В. Н. СОРОКИН

СИНТЕЗ РЕЧИ

МОСКВА «НАУКА»
ГЛАВНАЯ РЕДАКЦИЯ
ФИЗИКО-МАТЕМАТИЧЕСКОЙ ЛИТЕРАТУРЫ
1992

ББК 22.18
С65
УДК 519.76

Сорокин В. Н. **Синтез речи.**—М.: Наука. Гл. ред. физ.-мат. лит., 1992.—392 с.—ISBN 5-02-014665-X.

Синтез речи с использованием ЭВМ является составной частью современной информационной технологии. Методы синтеза речи находят широкое применение в информационно-справочных системах, в системах обучения с помощью ЭВМ и т. д. Читатель, обратившись к этой книге, сможет познакомиться с различными методами моделирования процессов речеобразования и восприятия речи.

Для научных инженерно-технических работников, специализирующихся в области разработки вычислительных и информационных систем, а также создания систем искусственного интеллекта.

Табл. 51. Ил. 176. Библиогр. 206 назв.

Рецензент доктор технических наук *М. А. Сапожков*

Научное издание

СОРОКИН Виктор Николаевич

СИНТЕЗ РЕЧИ

Заведующий редакцией *Е. Ю. Ходан*
Редактор *Т. И. Пташник*
Оформление художника *Б. М. Рябышева*
Художественный редактор *Г. М. Коровина*
Технический редактор *Л. В. Лихачева*
Корректор *Т. С. Вайсберг*

ИБ № 41232

Сдано в набор 03.08.91. Подписано к печати 03.03.92. Формат 60×90/16. Бумага тип. № 2. Гарнитура Таймс. Печать офсетная. Усл. печ. л. 24,50. Усл. кр.-отт. 24,50. Уч.-изд. л. 28,07. Тираж 920 экз. Заказ № 1011. С-047.

Издательско-производственное и книготорговое объединение «Наука»
Главная редакция физико-математической литературы
117071, Москва В-71, Ленинский проспект, 15

Ордена Октябрьской Революции и ордена Трудового Красного Знамени МПО «Первая Образцовая типография» Министерства печати и массовой информации РСФСР. 113054, Москва, Валовая, 28

Отпечатано в Новосибирской типографии № 4 ВО «Наука».
630077 Новосибирск 77, Станиславского, 25

С $\frac{1402070000-047}{053(02)-92}$ 129-92

ISBN 5-02-014665-X

ОГЛАВЛЕНИЕ

ПРЕДИСЛОВИЕ

В связи с развитием электронной технологии представления о сложности или простоте алгоритмов синтеза речи быстро меняются. В ближайшем будущем фактор ограниченной мощности процессоров перестанет определять возможности реализации синтезаторов. Тогда станет ясно, что так называемые «технически простые решения», игнорирующие или мало использующие знания о свойствах речеобразования и восприятия, не имеют никаких перспектив для использования в качестве средств передачи информации от машины человеку. Сдерживающим фактором в создании совершенных синтезаторов станет недостаток знаний о речевом сигнале. В этой книге фактически лишь намечены основные направления исследований. Предлагаемые решения в большинстве случаев служат только стартовыми позициями в формировании разделов нового научного направления — автоматического синтеза речи. По мере накопления опыта практического применения синтезаторов будут возникать новые проблемы, стимулирующие развитие этой области, лежащей на стыке наук о человеке и интеллектуальных машинах.

Предлагаемая книга в некотором смысле является развитием темы монографии автора «Теория речеобразования» (М.: Радио и Связь, 1985), поскольку разработка синтезаторов речи неотделима от математического моделирования процессов речеобразования. Задача книги состоит в описании вычислительных схем, необходимых для создания синтезаторов речи, к которым предъявляются различные требования. Почти все описанные алгоритмы реализованы на языке Фортран-5 и в том или ином виде испытаны при синтезе речи на универсальной ЭВМ. В Приложении приводятся сведения о речевом тракте, необходимые для разработки новых схем артикуляторно-формантных и артикуляторно-волновых синтезаторов.

При написании § 9.3 использовались результаты экспериментов на микролучевом рентгеноскопе университета штата Висконсин, а также программы анализа и синтеза речи, созданные Денисом Клаттом в Массачусетсском технологическом институте. В разработке метода прогонки и метода стрельбы, описанных в § 7.4, принимал участие А. Г. Миллер. Часть экспериментов по анализу длительности сегментов речи была выполнена на интонографе МПИИЯ им. М. Тореза. Сонограммы слогов были записаны В. Н. Ложкиным, МГУ.

ГЛАВА 1

ВВЕДЕНИЕ

§ 1.1. Задачи и способы синтеза речи

В технически развитом обществе сложность и скорость производственных процессов и процессов управления экономикой пришла в противоречие со способностью человека воспринимать, обрабатывать и передавать информацию. Становится все труднее получать имеющиеся где-либо сведения. Например, считается, что если научное исследование стоит не более 100 тысяч долларов, то часто его дешевле выполнить заново, чем разыскать нужные результаты в литературе. Поэтому доступ к информационным ресурсам становится столь же важным, как и доступ к сырьевым ресурсам, а увеличение скорости обработки информации необходимо для обеспечения устойчивости экономических систем и повышения производительности труда.

Трудности, созданные быстрым развитием техники, однако, могут быть решены средствами, которые эта же техника предлагает в виде электронных цифровых машин, способных не только управлять производственными процессами, но и обрабатывать информацию, необходимую для человека. В Японии, США, Франции и других странах созданы национальные программы по разработке ЭВМ пятого поколения, обладающих элементами искусственного интеллекта. Эти машины смогут читать тексты, распознавать зрительные образы, говорить и понимать человеческую речь.

Возможности речевого общения с искусственным интеллектом придается большое значение, поскольку в ряде случаев связь человека с машиной должна осуществляться на расстоянии, например, по телефонному каналу. Установлено, что речь является для человека наиболее удобным и естественным способом обмена информацией. Это означает, что при таком способе человек делает меньше ошибок, меньше устает, быстрее реагирует, а скорость обмена информацией выше, чем при других способах — визуальном, тактильном, тонально-звуковом. Повышение роли человека как элемента производственных и экономических процессов привело к созданию и интенсивному

развитию новой научно-технической области — речевой технологии или речевой информатики.

Речевая технология занимается разработкой систем автоматического распознавания и синтеза речи, распознавания (идентификации) и подтверждения (верификации) диктора, диагностики заболеваний по речевому сигналу, передачи и засекречивания речи по телефонным каналам. В основе речевой технологии лежат фундаментальные исследования свойств речевого сигнала, процессов речеобразования и восприятия, связи между речью и мышлением. Техническую базу речевой технологии создают ЭВМ, производительность которых непрерывно повышается, а габариты уменьшаются.

Синтез речи является важным элементом речевой технологии, который обеспечивает получение информации в речевой форме от различных источников. Применение синтеза речи целесообразно в тех случаях, когда зрительный канал не работает (в темноте), либо по каким-то причинам не может быть использован (для пилота), либо создает трудности в обработке информации. Если же связь осуществляется на расстоянии по телефону, как в информационно-справочных системах, автоматизированных системах управления, при общении с банком знаний или искусственным интеллектом, то синтез речи является основным способом получения информации для человека.

Применение синтеза речи. Синтез речи по тексту или коду сообщения может быть использован в технике связи, в информационно-справочных системах, для помощи слепым и немым, при управлении человеком со стороны автомата, для выдачи информации о технологических процессах, в военной и космической технике, в робототехнике, в акустическом *диалоге человека с* ЭВМ.

В США, Японии и других странах быстро увеличиваются объемы выпуска разнообразных синтезаторов и накапливается *опыт их практического применения. Имеющиеся сведения*, с одной стороны, подтверждают экономическую и техническую эффективность синтезаторов речи, а, с другой стороны, выявляют необходимость в дальнейшем улучшении их качества с тем, чтобы удовлетворить запросы всех потребителей. Приводимые ниже примеры существующих и потенциальных приложений синтеза речи свидетельствуют о большой экономической и социальной значимости этой области речевой технологии.

Имеется ряд категорий больных и инвалидов, лишенных возможности нормальной коммуникации: слепые, немые, люди с расстройствами речедвигательной системы. В США, например, насчитывается около 400 тысяч слепых и слабовидящих людей и более 1,5 миллиона людей, лишенных способности говорить. Значительная часть таких людей практически исключена из экономики, их труд малопроизводителен, сильно

ограничена возможность творческой работы. Синтезаторы речи могут изменить эти условия.

На протяжении нескольких десятилетий предпринимаются усилия по созданию читающих машин для слепых. Имеются опытные образцы, но широкому использованию читающих машин препятствует сложность и дороговизна оптико-механического блока, предназначенного для чтения текста с различными шрифтами. Тем не менее в США уже довольно давно выпускаются читающие устройства такого рода. Переход полиграфии на набор с помощью ЭВМ, при котором сначала создается копия книги в цифровом коде на магнитном носителе, устраняет необходимость в оптико-механическом блоке. Речь теперь может синтезироваться непосредственно по тексту, представленному в цифровом виде. Другая возможность состоит в чтении текста диктором, записи речи в сжатом виде, а затем синтезе по параметрам. Этот способ уже реализуется для выпуска речевых курсов обучения.

Для общения немых с говорящими людьми предназначаются синтезаторы речи, в которых сообщение набирается на алфавитной клавиатуре. В Финляндии создан экспериментальный образец такого переносного устройства. Его габариты невелики, а масса — около 1,5 кг — определяется, в основном, массой батарей [202].

В результате некоторых болезней, приводящих к расстройству управления движениями, люди лишаются возможности нормальной артикуляции, речь становится неразборчивой. Обслуживание таких людей часто требует специального персонала и стоит очень дорого, поэтому создание для них возможности сообщать о своих потребностях весьма актуально. Во Франции создано устройство, принцип работы которого состоит в том, что больной фиксирует свой взгляд на одной из ячеек прямоугольного табло, находящегося перед ним. На каждой ячейке написано слово, и устройство, следящее за положением глаз больного, фиксирует координаты слова на табло, а синтезатор речи генерирует фразу, состоящую из слов, последовательно выбираемых больным. Известный физик Стивен Хокинс, который вследствие болезни практически потерял способность к речи, для общения с людьми пользуется синтезатором.

В Калифорнийском университете было успешно испытано устройство предварительного диагноза шизофрении с помощью интеллектуальной системы, задающей пациенту вопросы в речевой форме, на которые он отвечал через клавиатуру алфавитного дисплея. Оказалось, что такой безличный опрос способствует уменьшению напряжения пациента и облегчает постановку предварительного диагноза, с которым пациент отправляется к специалисту.

Оценивая перспективы применения синтезаторов речи в сфере помощи инвалидам и больным, наряду с довольно высокой

стоимостью этих устройств, следует принимать во внимание и другие факторы. Поскольку число нуждающихся в синтезаторах исчисляется сотнями тысяч людей, то при таком массовом производстве их цена должна понизиться, тем более, что микропроцессоры и электронная память — один из немногих видов продукции, стоимость которых падает. Но наиболее важным является создание возможности более полной жизни для большого числа людей, что скажется на моральном климате общества. С экономической же точки зрения в долговременном плане более производительный труд этих людей наверняка окупит все затраты. К тому же, следует вспомнить, что шариковые авторучки и долгоиграющие пластинки, прочно вошедшие в нашу жизнь, были созданы в целях помощи слепым.

Другая гуманитарная область применения синтезаторов речи относится к обучению языку. На начальных этапах школьного обучения ребенку необходимо установить соответствие между произношением и написанием слов, и в этом ему может помочь синтезатор. В конце 70-х годов в США было выпущено устройство «Speak and Spell» («Говори и пиши»), которое по заданной программе произносило слова и оценивало правильность их буквенного представления на алфавитной клавиатуре. Это устройство приобрело огромную популярность, а стоимость его была вполне приемлемой — около 30 долларов. Затем появились различные варианты обучающих устройств. В перспективе вполне возможен перевод некоторой части процесса обучения в диалоговой режим с участием синтезаторов речи.

Правильное произношение при обучении иностранному языку также может быть достигнуто с помощью синтезаторов речи. Для нашей многонациональной страны весьма актуален вопрос обучения русскому языку в национальных школах, где зачастую сами преподаватели не владеют правильным произношением. Возможно также применение синтезаторов для устранения недостатков речи детей и больных, перенесших инсульт.

Использование синтезаторов наиболее эффективно в диалоговых системах с автоматическим распознаванием речи. В ближайшем будущем ЭВМ пятого и последующих поколений окажутся в состоянии вести речевой диалог в заданной предметной области, выступая как советчики. Перевод на диалоговый режим с использованием синтеза и распознавания речи систем автоматизированного управления (АСУ) и информационно-справочных систем (в том числе адресных и телефонных) обещает большой экономический эффект вследствие круглосуточной готовности и исключения человека из операций поиска информации.

В новых поколениях низколетающих вертолетов и самолетов, при операциях на орбите информационная нагрузка

на пилотов столь велика, что становится совершенно необходимой выдача информации о характеристиках полета и обстановке на борту с помощью синтезаторов речи. На бомбардировщиках ВВС США Б-52 установлена система речевого оповещения об аварийной ситуации. Синтезаторы речи используются в США и в тренажерах пилотов истребителей.

Имеется ряд технологических процессов, в которых использование зрения для получения информации невозможно или нежелательно: в темноте, при работе с микроскопом, во время хирургических операций. В США успешно прошла испытания система выдачи команд монтажнику сложной электронной аппаратуры с помощью синтезатора. Это позволило повысить производительность труда и сократить число ошибок путем исключения отвлечения внимания для чтения инструкций на алфавитно-цифровом печатающем устройстве, связанном с ЭВМ, контролирующей процесс монтажа.

Многие технические устройства, в том числе и бытовые, стали настолько сложны, что для их эксплуатации необходимо изучать иногда весьма трудно понимаемые инструкции. Это особенно затруднительно в ситуациях, требующих быстрого вмешательства в работу устройства. Перевод инструкций в речевую форму облегчает пользование такими устройствами. Например, в Японии выпускается швейная машинка с программным управлением, выдающая инструкции в речевой форме. Выпускаются также электронные калькуляторы, сообщающие результат действий через синтезатор. Речевой вывод в контрольных и измерительных приборах является не только удобным качеством, но и может предотвратить серьезные последствия, сообщая, например, об изменениях критически важных параметров технологических процессов, загазованности шахт, показателях пациента во время хирургической операции или в реанимационном отделении.

В токийском метро с конца 70-годов объявления об остановках совершаются синтезатором речи. Аналогичные информаторы устанавливаются в лифтах, в гостиницах, аэропортах и магазинах для объявлений и аварийного оповещения.

Синтезаторы речи могут стать средством творчества, если их использовать при создании радиоспектаклей и озвучивании сказочных персонажей.

В ближайшем будущем появятся роботы, которым потребуется общаться не только с человеком, но и друг с другом. Эти роботы могут быть электрически или программно несовместимы или находиться на расстоянии. Поэтому и здесь применение синтезаторов речи может оказаться вполне уместным. Наконец, синтезаторы речи являются мощными научными инструментами для исследования процессов речеобразования и свойств речевого сигнала, для изучения певческого голоса и патологии речи.

Из вышеприведенного краткого обзора следует, что область применения синтезаторов речи очень широка, а экономические

эффекты и социальные последствия их использования могут быть весьма велики. Вообще синтез речи может потребоваться во всех случаях, когда получателем информации является человек.

Классификация требований к синтезаторам. Анализ возможных применений синтетической речи позволяет сформулировать требования, предъявляемые к синтезу, а значит, определить соответствие того или иного способа синтеза задаче, в которой планируется его применение. Ниже приводятся основные факторы, определяющие выбор типа синтезатора.

1. Фиксированный или неограниченный набор сообщений. Фиксированный набор характеризуется объемом сообщений и степенью сменяемости — постоянный, редко и часто сменяемый.

2. Условия восприятия синтетической речи: наличие и тип шумов в канале связи или окружающей среде (в том числе реверберация, посторонние разговоры), искажения в канале связи (в том числе ограничения полосы частот, как в телефонном канале).

3. Степень умственной нагрузки, т. е. необходимость выполнения работы одновременно с прослушиванием синтезатора.

4. Натуральность, т. е. требуемая степень имитации человеческого голоса — от нарочито машиноподобного до естественной речи.

5. Необходимость в индивидуальной различимости, в частности, на мужской и женский голоса.

6. Необходимость в эмоциональной выразительности.

7. Длина непрерывных сообщений.

8. Требуемая разборчивость.

Способы синтеза. Можно определить синтез речи как восстановление формы речевого сигнала по его параметрам. В таком определении преобразование звукового давления в электрическое напряжение и наоборот в микрофоне и телефоне, а также запись и воспроизведение с магнитофонной ленты не являются синтезом. Дискретизация и квантование речевого сигнала при импульсно-кодовой модуляции также не относятся к синтезу речи, но генерация речевого сигнала в вокодерных системах может считаться синтезом. Более строгое определение синтеза состоит в требовании формирования речевого сигнала только по тексту. Мы будем пользоваться более широким определением синтеза. Все способы синтеза речи можно подразделить на три группы: параметрический, компиляционный и синтез по правилам. Каждый из этих способов удовлетворяет определенному набору требований из вышеперечисленных.

Параметрический синтез речи является конечной операцией в вокодерных системах, где речевой сигнал представляется набором небольшого числа непрерывно изменяющихся параметров. Параметрический синтез целесообразно

применять в тех случаях, когда набор сообщений ограничен и изменяется не слишком часто. Достоинством такого способа является возможность записать речь для любого языка и любого диктора. Качество параметрического синтеза может быть очень высоким (в зависимости от степени сжатия информации в параметрическом представлении). Однако параметрический синтез не может применяться для произвольных, заранее не заданных сообщений.

Компиляционный синтез заключается в сборке речевых сообщений из элементов, выделенных из естественной речи. Обычно в качестве таких элементов используются полуслоги — сегменты, содержащие половину согласного и половину примыкающего к нему гласного. При этом можно синтезировать речь по заранее не заданному тексту, но трудно управлять интонационными характеристиками. Качество такого синтеза не соответствует качеству естественной речи, поскольку на границах сшивки дифонов часто возникают искажения. Компиляция речи из заранее записанных словоформ также не решает проблемы высококачественного синтеза произвольных сообщений, поскольку акустические и просодические (длительность и интонация) характеристики слов изменяются в зависимости от типа фразы и места слова во фразе. Это положение не меняется даже при использовании дисков с лазерной записью, на которых может храниться огромное число словоформ.

Полный синтез речи по правилам обеспечивает управление всеми параметрами речевого сигнала и, таким образом, может генерировать речь по заранее неизвестному тексту. Синтез по правилам осуществляется, в основном, тремя способами: формантным, артикуляторно-формантным и артикуляторно-волновым. Разборчивость и натуральность таких синтезаторов может быть доведена до величин, сравнимых с характеристиками естественной речи.

§ 1.2. Кинематика и динамика артикуляторных органов

Механика и акустика речеобразования достаточно полно описаны в монографии [59], так что в этом разделе будут лишь кратко перечислены основные свойства речеобразования.

Речеобразующая система показана на рис. 1.1. Она включает в себя легкие, бронхи и трахею, составляющие подсвязочную область; голосовые складки, помещенные в защитную конструкцию, образованную щитовидным хрящом; к корню языка прикрепленный надгортанник, который, опускаясь, закрывает доступ в дыхательный тракт; язык, небная занавеска, губы и нижняя челюсть, являющиеся основными артикуляторными органами — их перемещения и изменения формы создают звукоразличительные признаки.

Сокращение грудных мышц заставляет легкие сжиматься, в результате чего из легких с определенной скоростью вытекает

воздушный поток. Протекая через голосовую щель, этот поток заставляет колебаться голосовые складки, создавая источник *голосового возбуждения, а в узких щелях ротовой полости* может возникнуть турбулентность, создающая шумовой источник возбуждения акустических колебаний. Когда объем легких уменьшается до некоторой критической величины, начинается вдох, так что в речи наступает пауза. Объем легких находится в пределах 3000—6000 см^3, причем уменьшение его на 20—30% вызывает необходимость вдоха.

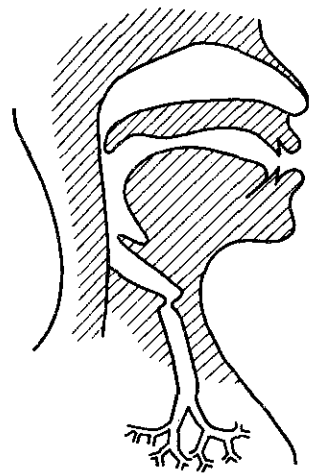


Рис. 1.1. Схема речевого тракта

Голосовые складки представляют собой упругие тела, и их взаимодействие с воздушным потоком сопровождается упругими колебаниями во всех трех измерениях (рис. 1.2). Вся гортань, в которой размещены голосовые складки, может смещаться вверх или вниз в процессе речеобразования, причем это смещение иногда достигает нескольких сантиметров. Перед началом фонации в исходном положении складки сведены, и амплитуда изменений площади голосовой щели при колебаниях складок составляет обычно 0,1—0,2 см^2. При генерации глухих согласных складки разведены таким образом, что площадь голосовой щели составляет примерно 0,2—0,4 см^2. Постоянная времени переходных процессов при сведении и разведении складок — 80—200 мс.

Небная занавеска при дыхании опущена. При артикуляции всех звуков, кроме (М, Н), она поднимается вверх и прижимается к задней стенке тракта, перекрывая проход в носовую полость (рис. 1.3). Поскольку упругие деформации небной занавески не играют заметной роли в формировании акустических характеристик речевого тракта, то ее движение можно описать уравнением для сосредоточенной системы

$$J\varphi'' + r\varphi' + c\varphi = F(t),$$

где φ — угол поворота небной занавески, J — момент инерции, r — *коэффициент вязкого трения*, c — *упругость*, F — *момент*, развиваемый мышцами. Амплитудно-частотные характеристики передаточной функции системы управления небной занавески показаны на рис. 1.4. Отметим разницу в движениях подъема и опускания, которая появляется в результате действия релаксационных процессов в мышечных тканях небной занавески, вследствие чего упругость c оказывается зависящей от времени. Коэффициент затухания r оценивается как 1—1,07, а постоянная времени $T = 1/c \approx 70 - 90$ мс.

Нижняя челюсть поворачивается относительно горизонтальной оси, причем сама ось может смещаться вперед или назад на несколько миллиметров, что используется при артикуляции переднеязычных звуков. Динамика нижней челюсти описывается

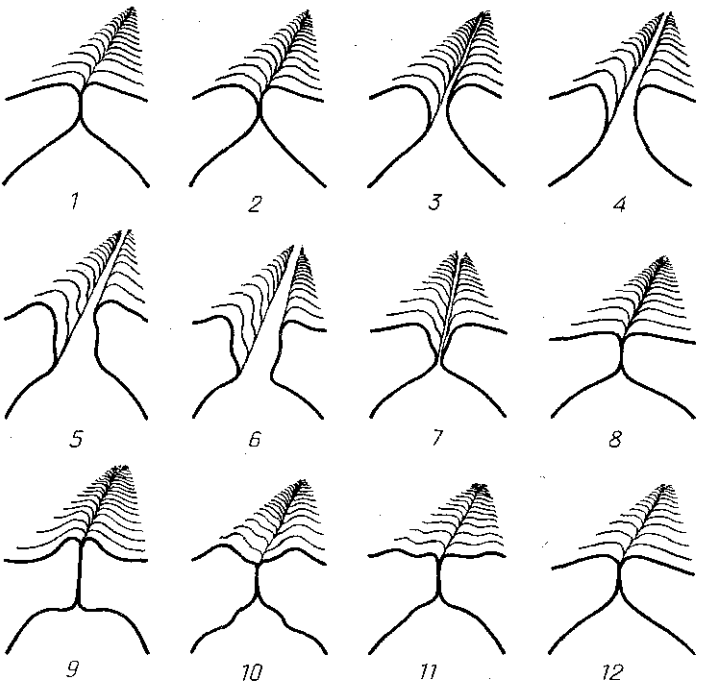


Рис. 1.2. Фазы колебаний голосовых складок

обыкновенным дифференциальным уравнением второго порядка с переменной жесткостью и вязкостью. Присоединенная упругость для нижней челюсти равна примерно 300 г/(см · с^2) коэффициент затухания — около 1,11, а характерная частота около 6 Гц. Измеренная амплитудно-частотная характеристика нижней челюсти показана на рис. 1.5.

Вертикальные движения каждой из губ могут быть описаны уравнениями второго порядка:

$$y''_{\text{г}} + 2g_{\text{г}}y'_{\text{г}} + \omega_{\text{г}}^2 y_{\text{г}} = F_{\text{г}}(t),$$

а форма губ описывается как

$$\Phi(y_{\text{г}}, z_{\text{г}}) = a_{\text{г}} y_{\text{г}} \sin \pi z_{\text{г}}/l_{\text{г}},$$

где $l_{\text{г}}$ — длина губ, $z_{\text{г}}$ — координата вдоль губ. Параметры $a_{\text{г}}$, $g_{\text{г}}$, $\omega_{\text{г}}$ различны для верхней и нижней губы, поскольку подвижность губ в вертикальном направлении неодинакова:

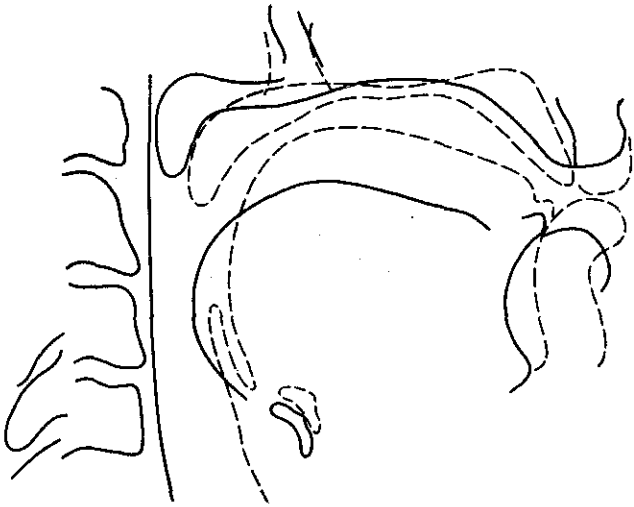


Рис. 1.3. Контурная рентгенограмма речевого тракта в слоге. МА: ---- середина сегмента назального гласного /М/, —— середина сегмента неназализованного гласного /А/

у нижней губы она гораздо больше. Уравнение относительно $y_г$ может использоваться и для расчета выпячивания губ при огублении гласных. В этом случае вертикальные и горизонтальные смещения определяются как $\alpha y_г$ и $\beta x_г$, где $\alpha+\beta=1$ ($x_г$ — смещение губ в горизонтальном направлении). Опускание нижней губы происходит гораздо быстрее, чем ее подъем — постоянная времени переходных процессов при опускании находится в диапазоне 70—90 мс (в зависимости от положения звука в звукосочетании), а при подъеме — в диапазоне 180—300 мс (при нормальном темпе артикуляции). Амплитудно-частотная характеристика системы управления движениями губ показана на рис. 1.6.

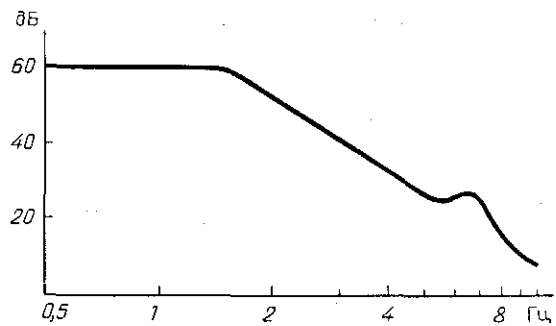


Рис. 1.4. Логарифмическая амплитудно-частотная характеристика системы управления небной занавески

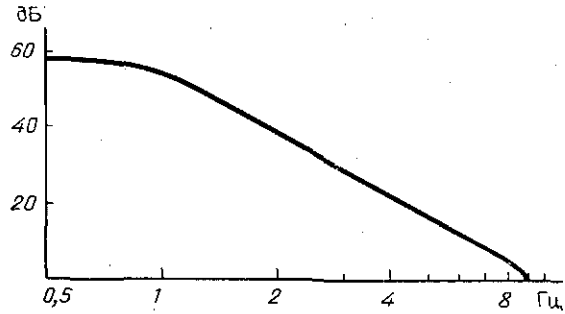


Рис. 1.5. Логарифмическая амплитудно-частотная характеристика системы управления нижней челюсти

Корень языка может смещаться в вертикальном направлении (к небу) и в горизонтальном направлении (к задней стенке или подбородку). При этом сам язык движется почти как твердое тело, и уравнение движения корня есть дифференциальное уравнение второго порядка с коэффициентом потерь $g_к$ и частотой $\omega_к$.

Движения кончика языка также могут быть описаны лишь одной собственной функцией его упругих деформаций:

$$\Phi_{кя}(\xi)=\psi_{1кя}(\xi)\,T_{1кя}(t),$$

где

$$\psi_{1кя}(\xi)=0{,}707\sqrt{\frac{2}{l_{кя}}}(\mathrm{ch}\,p_1\xi-\cos p_1\xi)-0{,}518\sqrt{\frac{2}{l_{кя}}}(\mathrm{sh}\,p_1\xi-\sin p_1\xi),$$

ξ — координата вдоль языка, $p_1=0{,}597\pi/l_{кя}$, $l_{кя}$ — «длина» кончика языка, обычно принимаемая равной 0,3—0,5 длины всего языка, $T_{1кя}(t)$ — решение дифференциального уравнения второго порядка:

$$T''_{1кя}+2g_{кя}T'_{1кя}+\lambda^2 ET_{1кя}=F_{кя}(t),$$

где $\lambda=p_1^2\sqrt{J_z/\rho}$, где J_z — момент инерции поперечного сечения

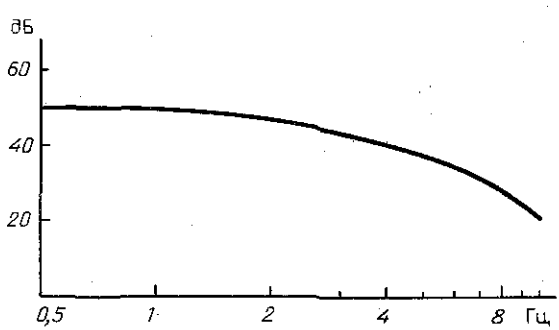


Рис. 1.6. Логарифмическая амплитудно-частотная характеристика системы управления нижней губы

кончика языка, E — модуль упругости тканей, ρ — плотность тканей ($\approx 1{,}1$ г/см^3). Амплитудно-частотная характеристика передаточной функции кончика языка показана на рис. 1.7.

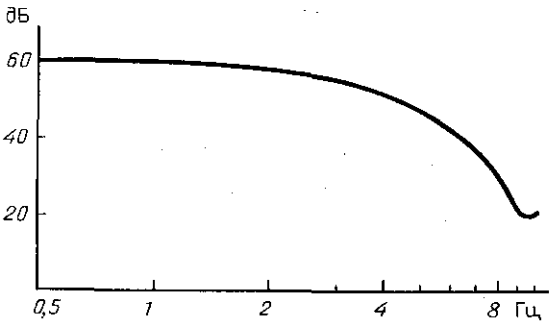


Рис. 1.7. Логарифмическая амплитудно-частотная характеристика системы управления кончика языка

Упругие деформации языка в целом описываются суперпозицией четырех собственных функций $\psi_{iя}(\varphi)$ и соответствующих им временных мод $T_{iя}(t)$:

$$\Phi(\varphi,\ t) = R_0 + \sum_{i=1}^{4} \psi_{iя}(\varphi)\, T_{iя}(t),$$

где R_0 — радиус поверхности языка в нейтральном положении ($R_0 \approx 3$—4 см), φ — угол в полярной системе координат с началом отсчета от корня языка,

$$\psi_{iя}(\varphi) = \operatorname{sh} p_i\varphi + \frac{p_i^2}{q_i^2} \sin q_i\varphi\, \frac{\operatorname{sh} p_i\varphi + p_i \operatorname{ch} p_i\varphi}{\sin \pi q_i + q_i \cos \pi q_i},$$

где $q_i^2 = 1 + p_i^2$, $p_1 = 0{,}8544$, $p_2 = 2{,}0347$, $p_3 = 3{,}1006$, $p_4 = 4{,}1353$. Временные моды получаются как решения уравнений

$$T''_{iя} + 2g_{iя} T'_{iя} + \omega_{iя}^2 T_{iя} = F_{iя}(t),$$

где

$$\omega_{iя}^2 = \frac{c_я}{\rho} + \frac{EJ_z}{\rho}(p_i q_i)^2,$$

$c_я$ — упругость подстилающего слоя ($c_я \approx 25$ г/(см $\cdot$ с^2)), E — модуль упругости ($E \approx 10^6$ Па), J_z — момент инерции ($J_z \approx 0{,}4$ см^4), $g_{iя} = 80$ с$^{-1}$.

Все вышеприведенные сведения о механических характеристиках артикуляторных органов нужны для построения системы управления артикуляцией. Более подробные данные приведены в монографии [59], к которой и следует обращаться в случае необходимости.

§ 1.3. Акустика речевого тракта

Речевой тракт представляет собой акустическую систему с распределенными параметрами, в которой распространяются, в основном, плоские волны. Если площадь поперечного сечения превышает 6 см^2, то в диапазоне частот от 4 кГц и выше в речевом тракте возникают и радиальные колебания. В этом случае уравнение речевого тракта, записанное в цилиндрической системе координат, есть

$$\frac{\partial^2\Phi}{\partial r^2}+\frac{1}{r^2}\frac{\partial\Phi}{\partial r}+\frac{1}{r^2}\frac{\partial^2\Phi}{\partial\varphi^2}+\frac{\partial^2\Phi}{\partial\xi^2}=\frac{1}{c_0^2}\frac{\partial^2\Phi}{\partial t^2},$$

где r — радиус, φ — азимут, ξ — криволинейная координата вдоль оси речевого тракта, c_0 — скорость звука, Φ — потенциал скорости акустических колебаний.

Форма речевого тракта мало влияет на резонансные частоты выше 4 кГц, поэтому разборчивость синтетической речи вряд ли будет зависеть от точности моделирования эффекта радиальных колебаний, но натуральность звучания, по-видимому, связана и с этим явлением. Обычно пользуются одномерным волновым уравнением, которое предполагает распространение только плоских волн. Для не слишком больших площадей поперечного сечения такое уравнение достаточно хорошо описывает акустику речеобразования в полосе до 5 кГц. В этом случае исходная система уравнений для речевого тракта записывается в следующем виде:

$$\begin{gathered}-S\frac{\partial P}{\partial\xi}=\rho_0\frac{\partial U}{\partial t}+r_1U,\\ -\rho_0c_0^2\frac{\partial U}{\partial\xi}=S\frac{\partial P}{\partial t}+r_2P+\rho_0c_0^2\frac{\partial S}{\partial t},\\ S=yL+S_0,\\ m\frac{\partial^2y}{\partial t^2}+b\frac{\partial y}{\partial t}+ky=P,\end{gathered}\qquad(1.1)$$

где ξ — координата вдоль оси речевого тракта, S — площадь поперечного сечения тракта, P — акустическое давление, U — объемная скорость акустических колебаний, r_1 и r_2 — коэффициенты потерь на вязкое трение и теплопроводность, y — смещение податливых стенок тракта, L — периметр сечения, S_0 — площадь сечения, создаваемая движениями артикуляторов, m, b и k — механические параметры стенок в расчете на единицу периметра.

Если в системе (1.1) пренебречь потерями и податливостью стенок, то можно получить так называемое уравнение Вебстера относительно давления P, объемной скорости $U=VS$

или потенциала скорости $\Phi = -\operatorname{grad} V$, где V — скорость акустических колебаний:

$$\frac{\partial}{\partial \xi}\left[S(\xi, t)\frac{\partial P}{\partial \xi}\right] = \frac{1}{c_0^2}\frac{\partial}{\partial t}\left[S(\xi, t)\frac{\partial P}{\partial t}\right],$$

$$\frac{\partial}{\partial \xi}\left\{\frac{1}{S(\xi, t)}\frac{\partial [S(\xi, t)V]}{\partial \xi}\right\} = \frac{1}{c_0^2}\frac{\partial}{\partial t}\left\{\frac{1}{S(\xi, t)}\frac{\partial [S(\xi, t)V]}{\partial t}\right\},$$

$$\frac{1}{S(\xi, t)}\frac{\partial}{\partial \xi}\left[S(\xi, t)\frac{\partial \Phi}{\partial \xi}\right] = \frac{1}{c_0^2}\frac{\partial^2 \Phi}{\partial t^2}.$$

Если площадь поперечного сечения тракта медленно меняется во времени, т. е. можно положить $\partial S/\partial t \approx 0$, то к этим уравнениям применима схема разделения переменных. Например, для давления из $P(\xi, t) = Z(\xi)\,T(t)$ получаем систему

$$(SZ')' + \lambda^2 Z = 0,$$
$$T'' + \lambda^2 c_0^2 T = 0,$$

где λ — собственные числа, которым соответствуют резонансные частоты тракта: $F_i = \lambda_i c_0/2\pi$.

От уравнения с распределенными потерями Q можно перейти к уравнению с сосредоточенными потерями $\delta_i(t)$, используя формулу

$$\delta_i = \frac{\int_0^l Q S \psi_i^2\, d\xi}{2\rho_0 \int_0^l S\psi_i^2\, d\xi},$$

где l — длина речевого тракта, ψ_i — собственные функции акустических колебаний. Тогда при малых потерях решение уравнения Вебстера может быть записано как

$$P(\xi, t) = \sum_{i=1}^{\infty} \Psi_i(\xi)\, T_i(t)\, e^{-\delta_i t}.$$

Радиальные колебания, создаваемые податливостью стенок, имеют резонансную частоту $F_{\text{рад}}$, которая зависит от объема речевого тракта:

$$F_{\text{рад}} = \frac{c_0}{2\pi\sqrt{L_{\text{ст}} V_{\text{пр}}/\rho_0}},$$

где $L_{\text{ст}} \approx 0{,}01$—$0{,}015$ г/см^4. Этот резонанс находится в диапазоне 150—350 Гц и повышает резонансные частоты F_i речевого тракта:

$$\bar{F}_i = \sqrt{F_i^2 + F_{\text{рад}}^2},$$

где $\bar{F}_i$ — резонансная частота акустической системы с подат-

ливыми стенками. При полной смычке частота первого резонанса в речевом тракте F_1 не стремится к нулю, как в системе с абсолютно жесткими стенками, а приближается к $F_{рад}$. Податливость приводит к излучению колебаний на частоте $F_{рад}$ во время звонкой смычки через стенки тракта. Излучение через стенки носа играет большую роль в создании эффекта назализации при распространении акустических колебаний через носовую полость.

Совместное действие различных видов потерь на вязкость, теплопроводность, излучение и колебания стенок, создают минимум потерь в среднем диапазоне частот (1—3 кГц), тогда как в низкочастотной области потери возрастают из-за податливости стенок, а в высокочастотной области — из-за излучения в свободное пространство.

§ 1.4. Нестационарные и параметрические явления

Вследствие того, что речевой тракт является системой с распределенными параметрами, заметную роль играют неустановившиеся процессы. При любых изменениях граничных условий, например, в голосовой щели, переходные процессы не могут длиться меньше удвоенного времени распространения сигнала от голосовой щели до губ, т. е. при длине голосового тракта в 17,5 см *длительность переходных процессов не меньше* 1 мс. За это время состояние голосовой щели может существенно измениться, и переходный процесс затягивается на весь интервал времени, на котором голосовая щель открыта, и захватывает часть интервала, на котором голосовая щель закрыта. В результате этого акустические процессы не успевают установиться, особенно на низких частотах, и резонансы не успевают полностью сформироваться. Часто это выглядит как понижение амплитуды колебаний, которое обычно трактуется как эффект возрастания потерь. Конечно, при открытой голосовой щели потери увеличиваются, но не только они определяют форму и амплитуду акустических колебаний. Кроме того, возникает заметный сдвиг по фазе между возбуждением от голосового источника и акустическими колебаниями в речевом тракте.

На *протяжении многих лет вновь и вновь исследуется* роль изменений формы речевого тракта во времени. В большинстве случаев приходят к выводу, что скорость изменения формы речевого тракта мала по сравнению со скоростью акустических колебаний, поэтому членом $\partial S/\partial t$ в уравнении Вебстера можно пренебречь, и допустимо решать это уравнение методом замороженных коэффициентов, считая все акустические процессы установившимися. Этот вывод, в общем, справедлив, но имеются два очень важных исключения. Действительно, если минимальная площадь поперечного сечения тракта не слишком мала (скажем, больше 1 см^2), то влияние изменения

площади на скорость изменения резонансных частот пренебрежимо мало. Однако дополнительная объемная скорость воздуха, втекающего или вытекающего из речевого тракта, при изменении его объема может привести к существенному изменению показателей затухания резонансных колебаний, и при определенных условиях ($\partial S/\partial t < 0$, $|\partial S/\partial t|$ больше некоторой величины) коэффициент затухания вместо отрицательного может стать положительным, что соответствует нарастанию вместо затухания колебаний [126]. Такие участки в самом деле иногда наблюдаются в речевых сигналах.

Второе явление связано с соотношением скорости изменения площади поперечного сечения тракта и скорости изменения резонансных частот. Как было показано в [59], при малых *площадях* (меньше 1 $см^2$) коэффициент зависимости скорости изменения частоты первого резонанса от скорости изменения площади тракта в сужении может увеличиться на порядок. При этом за один период колебаний на некоторой резонансной частоте сама эта частота меняется весьма существенно. Ясно, что при этом, как и во всех предыдущих случаях, квазистационарное решение волнового уравнения теряет силу.

Податливость стенок речевого тракта фактически создает нелинейные явления, поскольку величина смещения стенок зависит от акустического давления. Выше приводилась такая трактовка влияния податливости стенок, которая соответствовала описанию в терминах *радиального резонанса*, т. е. *допускала* существование установившегося режима акустических колебаний. Такое упрощение помогает пониманию физики явления и позволяет использовать приближенные решения волнового уравнения. Однако сомнительно, что такие упрощения при синтезе речи остаются незамеченными слуховым восприятием человека. В частности, одно из существенных отличий в механизмах речеобразования у женщин и мужчин заключается в неодинаковости механических характеристик (упругости и вязких потерь) тканей стенок речевого тракта. На низких частотах это различие проявляется в том, что ширина полосы первой форманты у женщин примерно в полтора раза больше, чем у мужчин. Моделирование этого свойства в синтезаторах существенно приближает синтетическую речь к качеству женского голоса, но это всего лишь квазистационарное решение, тогда как в действительности процессы, связанные с податливостью стенок, распределены и во времени, и вдоль речевого тракта.

Колебания голосовых складок могут рассматриваться как наиболее яркое проявление податливости стенок тракта. Вопреки распространенному мнению, сопротивление голосовой щели не является бесконечно большим даже по отношению к сопротивлению речевого тракта без заметных сужений. При артикуляции фрикативных звуков площадь сужения в тракте имеет те же величины, что и площадь голосовой щели, так что их сопротивления очень близки. Таким образом, голосовая

щель не замыкает речевой тракт, а расположена где-то посередине, если учесть длину трахеи и бронхов. Истинное начало речевого тракта лежит на выходе из легких, или даже в самих легких. Включая голосовую щель в речевой тракт, мы сразу осознаем последствия быстрого изменения площади голосовой щели во время фонации вплоть до полного перекрытия. Отсюда следует, что частоты и затухания резонансных колебаний подвергаются изменениям, синхронным с колебаниями голосовых складок. При смыкании складок резонансные частоты соответствуют длине тракта от голосовой щели до губ, при открытой голосовой щели — вдвое большей длине (от легких до губ), вследствие чего число резонансов в некоторой частотной полосе, например, в 5 кГц, меняется почти вдвое.

В силу конечной протяженности голосовой щели и малой площади ее максимального раскрытия на интервале открытой голосовой щели резонансные частоты могут не только понижаться, но и повышаться. Моделирование этого эффекта и анализ реальных речевых сигналов указывают на возможность девиации частоты первого резонанса до 20—30%. Поскольку формантные частоты и затухания, а также условия для возникновения дополнительных резонансов во время фонации меняются достаточно быстро, то весь комплекс связанных с этим явлений порождает существенно нестационарные процессы.

Особый класс составляют явления, связанные с аэродинамикой воздушного тока и его взаимодействием с акустическими процессами. Анализ этих явлений приводит к нелинейным уравнениям гидродинамических течений, т. е. уравнениям Навье — Стокса. Поскольку решение этих уравнений чрезвычайно трудоемко, можно попытаться разделить аэродинамическую и акустическую системы и искать приближенные способы описания их связи, как это делается, например, в гл. 5. Избавиться же от рассмотрения этих взаимосвязей невозможно, поскольку условия поддержания автоколебаний голосовых складок и характеристики этих колебаний в равной мере зависят как от нелинейности уравнений аэродинамики, так и от параметричности уравнений акустических колебаний. Грубо говоря, колебания голосового источника часто не могут возникнуть без наличия акустической нагрузки в виде речевого тракта.

Акустика речевого тракта, таким образом, оказывается нестационарной и нелинейной. Этим ограничивается область применения акустической теории Фанта [64] и объясняются трудности как в анализе речи, так и в разработке синтезаторов, генерирующих речь, близкую по всем характеристикам к естественной речи. В последующих разделах этой книги развиваются теоретические положения и описываются результаты моделирования основных свойств речеобразования, которые необходимо использовать в высококачественных синтезаторах речи.

ГЛАВА 2

МЕТОДЫ ЦИФРОВОЙ ОБРАБОТКИ СИГНАЛОВ В СИНТЕЗАТОРАХ

Методы цифровой обработки сигналов развивались в значительной степени для обеспечения анализа речи и передачи ее по каналам связи. В синтезе речи возникают специфические задачи, связанные с моделированием на электронных вычислительных машинах. Эти задачи требуют модификации известных методов цифровой обработки сигналов и создания специальных алгоритмов.

§ 2.1. Интегрирование

В процессе синтеза речи приходится вычислять интегралы разного вида. Например, при определении объема V речевого тракта по заданной площади поперечного сечения $S(x)$ нужно найти определенный интеграл $V=\int_0^l S(x)\,dx$, где l — длина речевого тракта. При решении дифференциальных уравнений первого порядка необходимо вычислять интегралы с переменным верхним пределом: $y(t)=\int_0^t x(\tau)\,d\tau$. Поскольку в синтезе речи подвергающиеся интегрированию функции редко бывают заданы аналитически, то интегрирование выполняется численными методами. Весьма эффективным способом организации вычислительных процедур оказывается рекурсивная форма. В дальнейшем мы ограничимся только вычислением интегралов от функций, заданных своими отсчетами через равномерные промежутки, хотя излагаемые методы вполне применимы и к функциям с неравномерными отсчетами.

Разные способы вычисления интегралов требуют разного количества операций и обладают разной погрешностью, поэтому выбор способа интегрирования зависит от требуемой точности. Рассмотрим три способа приближенного интегрирования, использующих интерполяцию интегрируемой функции полиномами нулевой, первой и второй степени. Эти способы

называются, соответственно, формулами прямоугольников, трапеций и парабол. Последний способ чаще встречается под названием формулы Симпсона.

Формула прямоугольников получается путем вычисления площади прямоугольника, ширина которого равна шагу h по аргументу интегрируемой функции, а высота равна отсчету функции f_i на i-м шаге:

$$y=\int_a^b f(t)\,dt\approx h\sum_{i=0}^{n} f_i.$$

Рекурсивная форма вычисления этого интеграла есть

$$y_i=hf_i+y_{i-1},$$

где y_{i-1} — значение интеграла на $(i-1)$-м шаге. Это простейшая формула вычисления интеграла, но она обладает наибольшей погрешностью, для снижения которой необходимо значительно уменьшить величину шага h, что не всегда возможно.

Формула трапеций получается при замене $f(t)$ на интервале $(ih, (i+1)h)$, хордой, соединяющей отсчеты f_i и f_{i+1}:

$$y=\int_a^b f(t)\,dt\approx\frac{h}{2}(f_0+2f_1+2f_2+\ \dots\ +2f_{n-1}+f_n).$$

Для вывода рекурсивной формы используем значение интеграла на $(i-1)$-м шаге:

$$y_i=\int_{(i-1)h}^{ih} f(t)\,dt+y_{i-1}.$$

Отсюда получаем рекурсивную форму для вычисления текущего значения интеграла:

$$y_i=\frac{h}{2}(f_{i-1}+f_i)+y_{i-1}.$$

Формула Симпсона выводится как результат параболической интерполяции функции $f(t)$:

$$y=\int_a^b f(t)\,dt\approx\frac{h}{3}(y_0+4y_1+2y_2+\ \dots\ +2y_{n-2}+4y_{n-1}+y_n).$$

Соответствующая рекурсивная форма есть

$$y_i=\frac{h}{3}(f_i+4f_{i-1}+f_{i-2})+y_{i-2},$$

где y_{i-2} — значение интеграла на $(i-2)$-м шаге.

Погрешность формулы трапеций есть

$$r_{\text{тр}}=-\frac{h^3}{12}f''(\xi),$$

где $f[(i-1)h]<f(\xi)<f(ih)$, а погрешность формулы Симпсона

есть

$$r_{\text{сим}} = -\frac{h^5}{90} f^{\text{IV}}(\xi).$$

Отсюда видно, что и формула трапеций и формула Симпсона дают значение интеграла с избытком, но формула Симпсона значительно точнее, а ее сложность не намного выше — требуется лишь на одно сложение и одно умножение больше, чем в формуле трапеций, и нужно запомнить на одно промежуточное значение интеграла больше. Впрочем, оценка погрешности зависит еще и от свойств интегрируемой функции, и могут встретиться случаи, когда $r_{\text{сим}}$ окажется значительно больше $r_{\text{тр}}$ из-за того, что $f^{\text{IV}} \gg f''$, так что меньшая погрешность будет обеспечена более простой формулой трапеций.

Существуют и другие, более точные способы численного интегрирования, но они значительно сложнее. Учитывая сравнительно высокую погрешность задания исходных величин при синтезе речи, по-видимому, в большинстве случаев достаточно использовать формулу Симпсона, а в ряде случаев приемлема и формула трапеций.

§ 2.2. Дифференцирование

Операции дифференцирования часто встречаются в алгоритмах синтеза речи. Здесь вопросы точности также играют большую роль, причем положение осложняется тем, что в большинстве случаев нужно вычислять производную от некоторой функции $f(t)$ в точке t^*, где известны значения этой функции только для $t \leqslant t^*$. При этом вычисление производной может осуществляться только с помощью экстраполяции функции $f(t)$ вперед, т. е. для $t > t^*$, и для достижения необходимой точности приходится пользоваться сложными формулами.

Рассмотрим сначала наиболее распространенные способы численного дифференцирования функции $f(t)$, заданной своими отсчетами через равномерно отстоящие промежутки аргумента t, равные h. Из определения производной

$$f'(t) = \lim_{h \to 0} \frac{f(t+h) - f(t)}{h}$$

в конечно-разностном виде получают две формулы для вычисления производной: так называемую левую производную

$$f'_{\text{л}i} = (f_i - f_{i-1})/h$$

и правую производную

$$f'_{\text{п}i} = (f_{i+1} - f_i)/h.$$

И та, и другая формулы подразумевают линейную интерпо-

ляцию функции $f(t)$ на конечных интервалах $((i-1)h, ih)$ или $(ih, (i+1)h)$. Существует еще один способ вычисления производной, использующий линейную интерполяцию функции $f(t)$, но не с участием точки t, где определяется производная, а между точками $t-h$ и $t+h$. Этот способ точнее, чем предыдущие два способа. Действительно, если представить функцию $f(t)$ в точках $f(t-h)$ и $f(t+h)$ разложением в ряд Тейлора, то

$$f(t-h)=f(t)-hf'(t)+\frac{h^2}{2}f''(t)-\frac{h^3}{6}f'''(t)+O(h^4),$$

$$f(t+h)=f(t)+hf'(t)+\frac{h^2}{2}f''(t)+\frac{h^3}{6}f'''(t)+O(h^4),$$

где $O(h^4)$ — остаточный член с погрешностью порядка h^4. Тогда, например, правая производная в точке t есть

$$\frac{f(t+h)-f(t)}{h}=f'(t)+\frac{h}{2}f''(t)+O(h^2),$$

а центральная производная по соседним точкам

$$\frac{f(t+h)-f(t-h)}{2h}=f'(t)+\frac{h^2}{6}f'''(t)+O(h^4).$$

Отсюда видно, что как правая, так и левая производные аппроксимируются лишь с первым порядком точности, а производная по симметричным соседним точкам — на порядок точнее, хотя и здесь нужно оговориться о влиянии производных f'' и f'''. Это является результатом того, что в последнем случае используется информация о значениях функции по обе стороны от точки t, т. е. фактически осуществляется интерполяция, тогда как в первых двух случаях осуществляется экстраполяция — вперед или назад.

В процессах синтеза речи обычно невозможно или очень трудно использовать значения функции в точках, находящихся правее той точки, в которой вычисляется производная. Использование же простейшего способа — левой производной с линейной интерполяцией часто не обеспечивает требуемой точности. Выход состоит в использовании интерполяционных многочленов более высокого порядка. Например, параболическая интерполяция на две точки назад дает следующие оценки производных:

$$f'(t-2h)=\frac{1}{2h}\left[-3f(t-2h)+4f(t-h)-f(t)\right]+\frac{h^2}{3}f'''(\xi),$$

$$f'(t-h)=\frac{1}{2h}\left[-f(t-2h)+f(t)\right]-\frac{h^2}{6}f'''(\xi),$$

$$f'(t)=\frac{1}{2h}\left[f(t-2h)-4f(t-h)+3f(t)\right]+\frac{h^2}{3}f'''(\xi).$$

Как видно, во всех трех точках погрешность порядка h^2, но

в средней точке $(t-h)$ эта погрешность все же вдвое меньше и, к тому же, вычисления проще — фактически оценка погрешности этого способа полностью соответствует оценке с линейной интерполяцией по симметричным соседним точкам. По изложенным выше причинам обычно требуется определить производную в точке t, и приведенная формула обеспечивает повышение точности за счет усложнения процедуры. Например, пользуясь формулой Рунге, можно повысить порядок точности на единицу:

$$\tilde{f}'_i = f'_i(h) + \left[f'_i(h) - f'_i(kh)\right]/(k^p - 1),$$

где $f'_i(h)$ — значение производной в точке i, полученное с использованием шага h, $f'_i(kh)$ — значение производной в той же точке, но для шага kh, p — порядок точности, т. е. степень при шаге h в оценке погрешности. Так, если взять $k=2$, то в случае вычисления левой производной по двум точкам

$$\tilde{f}'_i = f'_i(h) + \frac{f'_i(h) - f'_i(2h)}{2-1} = 2f'_i(h) - f'_i(2h).$$

Представляя производные через отсчеты функции f, получим

$$\tilde{f}'_i = \frac{1}{2h}(3f_i - 4f_{i-1} + f_{i-2}),$$

что в точности соответствует оценке производной по трем точкам с помощью параболической интерполяции. Естественно, что и погрешность получается того же, второго, порядка.

Для понижения погрешности еще на порядок нужно использовать уже четыре отсчета функции $f(t)$:

$$f'(t) = \frac{1}{6h}\left[-2f(t-3h) + 9f(t-2h) - 18f(t-h) + 11f(t)\right] + \frac{h^3}{4}f^{IV}(\xi).$$

Причина усложнения вычислительной процедуры ясна — для обеспечения той же точности при экстраполяции функции требуется информация о большем количестве точек, чем при интерполяции.

Описанные способы вычисления производной не являются упражнениями на тему вычислительной математики, так как при синтезе речи приходится решать системы дифференциальных уравнений, и сходимость решения зависит от точности представления производных. Поскольку одновременно необходимо экономить вычислительные ресурсы, то нужно располагать алгоритмами разной точности и сложности.

В рассмотренных выше процедурах дифференцирования подразумевалось, что дифференцируемая функция $f(t)$ не содержит случайного шума. Для таких функций полученные оценки погрешности являются точными. На практике же вычисления всегда сопровождаются погрешностями округления чисел в ЭВМ и погрешностями вычислительных алгоритмов.

Для таких функций операция дифференцирования приводит к возрастанию случайного шума, так как в частотной области дифференцирование соответствует подъему высоких частот с наклоном 6 дб/окт. В результате этого реальная погрешность может значительно превышать вышеприведенные оценки, и либо вычислительная схема теряет устойчивость, либо в синтетическом речевом сигнале начинают прослушиваться шумы. Таким образом, возникает задача фильтрации шумов операции дифференцирования. Корректное решение этой задачи требует знания вероятностных распределений шумов, что в большинстве случаев невозможно. Поэтому пользуются различными полуэвристическими процедурами, эффективность которых оценивают по качеству звучания синтезированного речевого сигнала.

Известно, что идеальное интегрирование подавляет высокочастотные шумы, осуществляя преобразование в частотной области с наклоном 6 дб/окт. Этим свойством пользуются для организации простейших схем фильтрации. Одна из таких схем называется «центральная разность с усреднением». В ней значение функции f' в точке i заменяется средним значением с учетом соседней точки, например

$$\bar{f}'_i=(f'_i+f'_{i-1})/2.$$

Применяя эту схему к операции вычисления левой производной, из

$$f'_i=(f_i-f_{i-1})/h$$

получим рекурсивную форму

$$f'_i=\frac{2}{h}(f_i-f_{i-1})-f'_{i-1}.$$

Аналогично, для параболической интерполяции

$$f'_i=\frac{1}{h}(f_{i-2}-4f_{i-1}+3f_i)-f'_{i-1}.$$

Во многих случаях этот способ сглаживания оказывается достаточно эффективным. Если же в результате подобного дифференцирования все же появляются нежелательные шумы, то следует применить фильтрацию с использованием большего числа точек. Построению фильтров посвящена большая литература (см., например [35, 40, 47]), поэтому здесь эти вопросы не рассматриваются.

Один из эффективных способов сглаживания производной называется реальным дифференцированием. В нем используются для сглаживания простейшие фильтры с различными постоянными времени, т. е. взвешенным усреднением по разному числу отсчетов функции. Реальное дифференцирование задается следующим уравнением:

$$Ty'+y=f'.$$

Для левой производной в конечных разностях получаем

сглаженную производную в рекурсивной форме:

$$y_i = \frac{f_i - f_{i-1} + T y_{i-1}}{T + h}.$$

Таким образом, при $T = 0$ имеем чистую левую производную — «идеальное» дифференцирование, а при $T > 0$ выполняется сглаживание. Аналогично, можно получить реальную производную и для параболической интерполяции.

§ 2.3. Интерполяция

Вычисляя значения какой-либо функции $f(x)$ в последовательности точек x_i, часто требуется найти $f(x)$ для других значений аргумента x. Иногда это задача интерполяции, как, например, при вычислении функции площади поперечного сечения $S(x)$, заданной в какой-либо момент времени на интервале $0 \leqslant x \leqslant l$. В других случаях это задача экстраполяции, когда необходимо предсказать функцию $f(t)$ по некоторому числу отсчетов, предшествующих точке t. В частном случае это задача представления функции $f(t)$ с другим числом отсчетов (или с другим размером шага дискретизации). Наконец, в алгоритмах формирования акустической волны функцию площади поперечного сечения речевого тракта необходимо аппроксимировать либо полиномами нулевого порядка (цилиндрическая аппроксимация), либо полиномами второго порядка (коническая аппроксимация). Наряду с хорошо известными методами аппроксимации в синтезаторах речи приходится применять и специальные методы, учитывающие особенности протекающих процессов.

Начнем с простейшего случая прореживания отсчетов, когда функцию $f(x)$, вычисленную с шагом h, нужно представить отсчетами с шагом kh, $k = 2, 3, \ldots$ Очевидно, при этом достаточно оставить лишь каждый k-й отсчет. Такая задача возникает, например, при вычислении формы речевого сигнала методом бегущих волн, когда частоту отсчетов приходится уменьшать вдвое (при дискретизации сигнала, подаваемого на прослушивание, с частотой 20 кГц), или даже вчетверо (при частоте дискретизации в 10 кГц). Поскольку после прореживания функция $f(x)$ представлена меньшей частотой отсчетов, а в ее спектре содержатся и более высокочастотные компоненты, то для избежания искажений вследствие наложения спектров после прореживания необходимо выполнить фильтрацию с помощью фильтра низких частот с частотой среза $F_{ср} = 1/(2kh)$.

Более типичной является ситуация, в которой требуется увеличение частоты отсчетов. Например, площадь поперечного сечения речевого тракта меняется довольно медленно во времени, и ее достаточно вычислять с частотой не выше

1 кГц. Однако в схеме бегущих волн значения этой площади должны обновляться с частотой в несколько десятков кГц (порядка 70—80 кГц). Другой пример — работа источников возбуждения также может рассчитываться с меньшими частотами отсчетов, чем тактовая частота схемы бегущих волн. Здесь существенным обстоятельством является невозможность интерполяции, поскольку обычно нельзя задерживать расчет акустических процессов до появления следующего отсчета функции площади поперечного сечения или сигнала источника возбуждения. Единственным способом определения значений некоторой функции с большей частотой служит экстраполяция. С этой целью разложим в ряд Тейлора экстраполируемую функцию в точке t:

$$f(t+h)=f(t)+hf'(t)+\frac{h^2}{2}f''(t)+\frac{h^3}{6}f'''(t)+\ldots,$$

где h — приращение аргумента. Пользуясь параболической интерполяцией, представим первую и вторую производные в точке t как

$$f'(t)=\frac{1}{2H}\left[f(t-2H)-4f(t-H)+3f(t)\right],$$

$$f''(t)=\frac{1}{H^2}\left[f(t-2H)-2f(t-H)+f(t)\right],$$

где H — исходный (большой) шаг отсчетов функции $f(t)$. Ограничиваясь первыми тремя членами в ряде Тейлора, получим экстраполяционную формулу

$$f(t+h)=\left[1+\frac{\eta}{2}(3+\eta)\right]f(t)-\eta(2+\eta)f(t-H)+\frac{\eta}{2}(1+\eta)f(t-2H),$$

где $\eta=h/H$. Поскольку $h<H$, то $\eta<1$, причем величина h не обязательно кратная к H. Если погрешность такой экстраполяции окажется слишком велика, то можно использовать интерполяционные формулы для производных на четыре точки, например, по [1].

В ряде случаев при экстраполяции необходимо обеспечить непрерывность первой производной. Например, первая производная по времени от площади поперечного сечения (точнее, от объема речевого тракта) используется для вычисления дополнительного потока воздуха в речевом тракте при изменении его объема. В таких случаях нельзя взять меньше трех членов ряда Тейлора для экстраполяции.

Наиболее распространенным способом описания функции площади поперечного сечения речевого тракта $S(x)$ является ступенчатая аппроксимация, эквивалентная представлению реального речевого тракта в виде последовательности отрезков цилиндрических труб разного сечения. Таким образом, ставится

задача наилучшей в некотором смысле аппроксимации $S(x)$ полиномами нулевой степени. Если в качестве критерия выбрать минимум среднеквадратичного отклонения искомой постоянной величины c_i от функции $S(x)$ на интервале (x_{i-1}, x_i), то

$$F(c_i)=\int_{x_{i-1}}^{x_i}[S(x)-c_i]^2\,dx.$$

Поскольку этот интеграл является непрерывной функцией от c_i, и на c_i не наложено ограничений, то минимум находится из условия

$$-\frac{1}{2}\frac{\partial F}{\partial c_i}=\int_{x_{i-1}}^{x_i}[S(x)-c_i]\,dx=0.$$

Отсюда имеем

$$c_i=\frac{1}{x_i-x_{i-1}}\int_{x_{i-1}}^{x_i}S(x)\,dx, \tag{2.1}$$

т. е. для среднеквадратичного критерия оптимальности значение аппроксимирующей постоянной c_i равно средней величине площади $S(x)$ на интервале (x_{i-1}, x_i). Если интервал (x_{i-1}, x_i) фиксирован, например, в виде отсчетов от $S(x)$ с равномерным шагом $h=x_i-x_{i-1}$, то такой метод гарантирует минимум среднеквадратичной погрешности. Однако этот минимум не обязательно находится в пределах допустимой погрешности аппроксимации $S(x)$, вытекающей из требований точности вычисления акустических процессов. Поэтому изменим задачу и сведем ее к поиску такой величины c_i и такого интервала (x_{i-1}, x_i), чтобы среднеквадратичная погрешность не превышала некоторого ε_i:

$$F(c_i, x_i)=\int_{x_{i-1}}^{x_i}[S(x)-c_i]^2\,dx\leqslant\varepsilon_i, \tag{2.2}$$

где x_{i-1} — фиксированное значение, а x_i — переменное. Для того чтобы найти $F_{\min}$ и сравнить его с требуемой погрешностью, подставим в (2.2) значение c_i из (2.1):

$$F_{\min}=\int_{x_{i-1}}^{x_i}S^2(x)\,dx-\frac{1}{x_i-x_{i-1}}\left[\int_{x_{i-1}}^{x_i}S(x)\,dx\right]^2\leqslant\varepsilon_i. \tag{2.3}$$

Здесь единственной неизвестной является переменная x_i, и алгоритм аппроксимации состоит в постепенном увеличении x_i до тех пор, пока условие (2.3) не перестанет выполняться, после чего вычисляется значение c_i для последнего x_i, при котором (2.3) еще выполнялось.

Поскольку известно, что для меньшей площади $S(x)$ требуется более точная аппроксимация, чем для большей площади, то можно ввести зависимость допустимой погрешности ε_i от c_i, например, в виде $\varepsilon_i = \alpha c_i$. Тогда условие (2.3) представляется как

$$(x_i - x_{i-1}) \frac{\int_{x_{i-1}}^{x_i} S^2(x)\,dx}{\int_{x_{i-1}}^{x_i} S(x)\,dx} - \int_{x_{i-1}}^{x_i} S(x)\,dx \leqslant \alpha.$$

Очевидно, что найденные точки отсчета x_i при этом окажутся на разных расстояниях, т. е. функция $S(x)$ будет аппроксимирована с неравномерными отсчетами. При этом на участках медленного изменения $S(x)$ отсчеты будут браться реже, а при быстром изменении $S(x)$ — чаще. В синхронных схемах бегущих волн типа Келли — Локбаума необходимо, чтобы длина цилиндрической секции была кратной некоторой величине Δx. Для таких схем интервал (x_{i-1}, x_i), полученный в результате минимизации ошибки, должен укорачиваться таким образом, чтобы $(x_i - x_{i-1})/\Delta x$ было равно целому числу, а следующий этап аппроксимации начинается с уточненной таким образом координаты x_i.

Возможно использование и других зависимостей $\varepsilon_i(c_i)$, например логарифмической. Это несколько усложняет вычисления, но не меняет существа алгоритма аппроксимации.

При кусочно-линейной аппроксимации функциями вида $c + bx$ для среднеквадратичного критерия ищем минимум функционала

$$F(b, c) = \int_{x_{i-1}}^{x_i} \left[f(x) - b_i(x - x_{i-1}) - c_i\right]^2 dx. \qquad (2.4)$$

Из системы

$$-\frac{1}{2}\frac{\partial F}{\partial b_i} = \int_{x_{i-1}}^{x_i} x\left[f(x) - b_i(x - x_{i-1}) - c_i\right] dx = 0,$$

$$-\frac{1}{2}\frac{\partial F}{\partial c_i} = \int_{x_{i-1}}^{x_i} \left[f(x) - b_i(x - x_{i-1}) - c_i\right] dx = 0,$$

найдем b_i и c_i, удовлетворяющие условию минимума среднеквадратичной погрешности. Подставив эти значения в (2.4), определим такое x_i, при котором

$$F_{\min}(b_i, c_i, x_i) \leqslant \varepsilon_i.$$

При кусочно-линейной аппроксимации возможны два варианта: с разрывом аппроксимирующей функции в точках x_i и без

разрыва. Во втором случае накладывается дополнительное условие

$$c_{i+1}=b_i(x_i-x_{i-1})+c_i,$$

так что задача аппроксимации снова становится одномерной и ограничивается лишь поиском наклона b_i линейной функции *на каждом из интервалов* (x_{i-1}, x_i).

При представлении речевого тракта в виде последовательности конических секций площадь поперечного сечения $S(x)$ аппроксимируется квадратичными полиномами специального вида:

$$\varphi_i(x)=c_i+b_i(x-x_{i-1})^2.$$

В этом случае минимизируется функционал

$$F=\int_{x_{i-1}}^{x_i}\left[S(x)-b_i(x-x_i)^2-c_i\right]^2 dx. \qquad (2.5)$$

Взяв частные производные по b_i и c_i, имеем систему уравнений

$$b_i\left(\frac{x_i^2+x_{i-1}^2}{3}-x_i x_{i-1}\right)+c_i=\frac{1}{x_i-x_{i-1}}\int_{x_{i-1}}^{x_i} S(x)\,dx,$$

$$\frac{3}{5}b_i(x_i-x_{i-1})^2+c_i=\frac{3}{(x_i-x_{i-1})^3}\int_{x_{i-1}}^{x_i} S(x)(x_i-x_{i-1})\,dx.$$

Решив эту систему и определив b_i и c_i, подставим их в (2.5) и найдем такое x_i, при котором

$$F_{\min}(b_i, c_i, x_i)\leqslant\varepsilon_i.$$

Так же как и в предыдущем случае, при аппроксимации без разрыва задача сводится к поиску лишь одного коэффициента раскрытия конуса b_i, так как $c_{i+1}=c_i+b_i(x_i-x_{i-1})^2$.

До сих пор мы рассматривали способы представления функции площади поперечного сечения речевого тракта $S(x)$, исходя из требований минимума ошибки аппроксимации. При таком строго математическом подходе за пределами внимания *оставался вопрос о целях аппроксимации. Между тем в синтезе* речи наиболее точная аппроксимация $S(x)$ полиномами того или иного порядка не является единственным способом *достижения целей синтеза. На самом деле задача заключается* в таком описании $S(x)$, при котором акустические характеристики речевого тракта были бы сохранены с наименьшей погрешностью. Можно предложить ряд способов такой аппроксимации $S(x)$, в которых погрешность представления $S(x)$

на отдельных участках достигает значительной величины, тогда как искажения акустических характеристик синтезированного речевого сигнала на слух не заметно. Опишем один из таких алгоритмов, который можно назвать алгоритмом аппроксимации по экстремумам $S(x)$.

Сначала вводится система нелинейных порогов $\Delta(S)$, позволяющая квантовать $S(x)$ с меньшим шагом при малых значениях площади и с большим шагом — при больших площадях. Зависимость $\Delta(S)$ может быть выбрана логарифмической.

Затем отыскиваются все локальные экстремумы (минимумы и максимумы), которые квантуются по порогу $\Delta(S)$. Если два соседних минимума или максимума после квантования оказываются одинаковыми, то один из них вычеркивается.

От каждого минимума движемся в обе стороны до ближайших максимумов и присваиваем $S(x)=S_{i\min}$, если $S(x)-S_{i\min}\leqslant\delta(S)$, где $\delta(S)$ — другая система порогов, $\delta(S)>\Delta(S)$.

Начиная с точки x_j, в которой выполняется условие $S(x_j)-S_{i\min}=\delta(S)$, меняем значение порога δ на большее, $\delta_j=\delta\,[S(x_j)]$. Присваиваем $S(x)=S(x_j)+\delta_j$, если $S(x)>S(x_j)$ и присваиваем $S(x)=S(x_j)-\delta_{j+1}$, если $S(x)<S(x_j)+\delta_j$, где $\delta_{j+1}=\delta\,[S(x_j)+\delta_j]$.

Процесс продолжается до тех пор, пока не будет достигнут соседний максимум, но прекращается, если потребуется присвоить значение, большее, чем ближайший максимум.

После окончания перебора по минимумам $S(x)$ рассматриваются окрестности всех максимумов и функция $S(x)$ заменяется на $S_{k\max}$, если $S_{k\max}-S(x)\leqslant\delta(S_{k\max})$.

В этом алгоритме сохраняются точные значения максимумов и минимумов площади поперечного сечения, а участки между ними могут квантоваться, например, на два или три уровня. При этом получается меньшее число цилиндрических секций, чем в алгоритме, минимизирующем среднеквадратичную погрешность. Такое сокращение весьма желательно, так как вычислительные затраты на синтез речевой волны сильно зависят от числа секций.

В принципе, можно поставить задачу минимизации числа цилиндрических секций при условии, что искажения резонансных частот речевого тракта при этом не будут превышать заранее заданных порогов. Эта задача решается путем использования одного из описанных в гл. 7 методов расчета собственных частот речевого тракта. Однако в задачах синтеза речи такой подход, по-видимому, представляет лишь теоретический интерес, поскольку выигрыш в количестве вычислений акустической волны, полученный в результате минимизации числа секций, скорее всего, окажется меньше объема дополнительных вычислений, требующихся для расчета собственных чисел и процесса оптимизации.

§ 2.4. Дифференциальное уравнение первого порядка

Обыкновенные дифференциальные уравнения первого порядка часто встречаются в процессах синтеза речи, причем во многих случаях они имеют переменные коэффициенты, так что их аналитические решения неизвестны. Поэтому необходимо использовать численные методы, но учитывать, что *они обладают разной точностью. Рассмотрим возникающие* проблемы на примере однородного уравнения

$$f'(x)+Rf(x)=0$$

с начальным условием $f(0)=1$. Взяв правую производную, получим конечно-разностное уравнение

$$\frac{f(x+h)-f(x)}{h}+Rf(x)=0,$$

которое в рекурсивной форме выглядит как

$$f(x+h)=(1-Rh)f(x).$$

Отсюда видно, что для выбранных начальных условий

$$f(nh)=(1-Rh)^n,$$

и при $h=1/n$

$$f(1)=\left(1-\frac{R}{n}\right)^n,$$

тогда как точное решение есть

$$f(1)=e^{-R}.$$

Хотя известно, что при достаточно малом h (т. е. при большом n) *эти решения мало отличаются [11], но в задачах синтеза* речи величина шага h обычно задается, исходя из других требований, и погрешность вычислений может оказаться довольно большой. С другой стороны, если пользоваться решением для уравнения с постоянными коэффициентами, то вычисление экспоненты требует довольно большого числа операций. Возникает вопрос, нельзя ли построить рекурсивную схему, используя знание точного решения.

Представим однородное уравнение как

$$f(x)=af(x-h),$$

и найдем коэффициент a путем сравнения его с точным решением

$$f(x)=e^{-Rx}.$$

Из $a=f(x)/f(x-h)$ получим $a=e^{-Rx}/e^{-R(x-h)}=e^{-Rh}$.

Частное решение неоднородного уравнения

$$f'(x)+Rf(x)=F(x) \tag{2.6}$$

найдем из

$$f_i+af_{i-1}=bF_i.$$

Принимая во внимание, что при дискретизации функции $F(x)$ ее значение от отсчета к отсчету меняется скачком, используем частное решение для ступенчатого возмущения F. Известно, что в общем случае решение (2.6) есть

$$f(x)=\exp\{-\int_{x_0}^{x} R(x)\,dx\}\left[f(0)-\int_{x_0}^{x} F(x)\exp\{\int_{x_0}^{x} R(x)\,dx\}\,dx\right],$$

а для ступенчатого возмущения, $F=\text{const}$ и $R=\text{const}$,

$$f(x)=\frac{F}{R}(1-e^{-Rx}).$$

Отсюда находим коэффициент

$$b=(1-e^{-Rh})/R,$$

так что рекурсивная схема представляется в виде

$$f_i=f_{i-1}+\left(\frac{F_i}{R}-f_{i-1}\right)(1-e^{-Rh}). \tag{2.7}$$

Если коэффициент R постоянен, то экспонента вычисляется только один раз. При переменном R нужно вычислять экспоненту на каждом шаге, однако здесь возможны некоторые упрощения. Обычно шаг h весьма мал. Например, при частоте отсчетов в 20 кГц, $h=5\cdot10^{-5}$ с. Поэтому и произведение Rh также может быть весьма мало, что позволяет от вычисления экспоненты e^{-Rh} перейти к вычислению одного — двух членов ее разложения в степенной ряд

$$e^{-Rh}=1-Rh+\frac{(Rh)^2}{2}+\ldots$$

Так, при использовании лишь двух членов разложения при $Rh\leqslant0{,}19$ ошибка вычисления экспоненты составляет менее 1%. При этом рекурсивная схема принимает особенно простую форму:

$$f_i=f_{i-1}+(F_i-R_if_{i-1})h. \tag{2.8}$$

Специальный вид уравнения первого порядка, содержащий нелинейный (квадратичный) член, т. е. уравнение Риккати, также может быть решен с помощью описанного приема. Это уравнение появляется при описании аэродинамических процессов в речевом тракте, и его решение описывается в § 5.3.

§ 2.5. Дифференциальное уравнение второго порядка

Дифференциальные уравнения второго порядка лежат в основе описания большинства процессов синтеза речи. Например, сюда относятся механические движения артикуляторных органов и голосовых складок. Акустические колебания в формантном синтезаторе также описываются этими уравнениями. Рассмотрим уравнение вида

$$f''+2gf'+\omega^2 f=F(x). \tag{2.9}$$

Пользуясь центральной производной

$$f'(x)=\frac{f(x+h)-f(x-h)}{2h},$$

получим конечно-разностную форму

$$f(x+h)=\frac{1}{1+gh}\left[F(x)h^2+f(x)(2-\omega^2 h^2)+f(x-h)(gh-1)\right],$$

которая применима и к уравнению с зависимыми от x коэффициентами g и ω. Здесь мы сталкиваемся с той же проблемой, что и при решении уравнений первого порядка. В частности, если $gh\ll 1$ и величина gh выходит за пределы разрядной сетки ЭВМ, устойчивое решение невозможно. Обеспечение устойчивости и повышение точности решения достигается использованием аналитического решения для уравнения с постоянными коэффициентами, как это было сделано в предыдущем разделе.

Обозначим $f_i=f(ih)$ и рассмотрим сначала общее решение (2.9), т. е. свободные колебания при $F(x)=0$. Перепишем это уравнение как

$$f_i=a_1 f_{i-1}+a_2 f_{i-2},$$

и найдем коэффициенты a_1 и a_2, сопоставляя их с коэффициентами в решении уравнения с постоянными параметрами g и ω. Если $\omega^2>g^2$ и решение (2.9) есть затухающие колебания, то

$$f(x)=e^{-gx}\cos(\lambda x+\varphi),$$
$$f(x-h)=e^{-g(x-h)}\cos[\lambda(x-h)+\varphi],$$
$$f(x-2h)=e^{-g(x-2h)}\cos[\lambda(x-2h)+\varphi],$$

где $\lambda^2=\omega^2-g^2$. Из системы

$$f_i=a_1 f_{i-1}+a_2 f_{i-2},$$
$$f_{i+1}=a_1 f_i+a_2 f_{i-1},$$

имеем

$$a_1 = \frac{f_{i+1} - a_2 f_{i-1}}{f_i}, \quad a_2 = \frac{f_{i-1} f_{i+1} - f_i^2}{f_{i-1}^2 - f_i f_{i-1}}.$$

Подставляя в эти выражения значения f_i, f_{i-1} и f_{i-2} и выполняя элементарные выкладки, которые здесь не приводятся ввиду их громоздкости, получим

$$a_1 = 2e^{-gh}\cos\lambda h,$$
$$a_2 = -e^{-2gh}.$$

Отметим, что используемое обычно в формантных синтезаторах значение

$$a_1 = 2e^{-gh}\cos\omega h$$

годится только в тех случаях, когда $\omega \gg g$.

Если $g^2 > \omega^2$ и решение (2.9) апериодическое, то коэффициент a_2 остается тем же, а

$$a_1 = (e^{\lambda h} + e^{-\lambda h})e^{-gh} = 2e^{-gh}\operatorname{ch}\lambda h.$$

Наконец, в вырожденном случае $\omega^2 = g^2$ решение для свободного движения есть

$$f(x) = e^{-gx}(c_1 x + c_2),$$

и, соответственно, коэффициенты a_1 и a_2 находятся наиболее просто:

$$a_1 = 2e^{-gh}, \quad a_2 = -e^{-2gh}.$$

При малых значениях λh допустимо вместо $\cos\lambda h$ использование первых двух его членов разложения в степенной ряд:

$$\cos\lambda h \approx 1 - (\lambda h)^2/2.$$

Для неоднородного уравнения рекурсивная схема имеет дополнительный член с коэффициентом b:

$$f_i = bF_i + a_1 f_{i-1} + a_2 f_{i-2}.$$

Пользуясь описанным в предыдущем разделе методом, найдем b как

$$b = (1 - a_1 - a_2)/\omega^2.$$

Полученная таким образом рекурсивная схема для уравнения второго порядка обеспечивает достаточную точность при синтезе речи.

ГЛАВА 3

ПАРАМЕТРИЧЕСКИЙ И КОМПИЛЯЦИОННЫЙ СИНТЕЗ

§ 3.1. Аппроксимация речевого сигнала

Как обсуждалось во введении, существует обширный круг задач, в которых объем речевых сообщений невелик, а сами эти сообщения меняются редко или вообще не меняются. В этом случае целесообразно заранее запомнить весь речевой материал, и по коду сообщения выводить тот или иной текст в звуковой форме. Простейший способ заключается в записи речевого сигнала в аналоговой форме на магнитную ленту и его воспроизведении по управляющему сигналу, как это делается, например, в системах «говорящие часы» или в объявлениях остановок в метро. Электромеханические устройства, однако, дороги в обслуживании и недолговечны, а выборка нужного сообщения из большого объема создает задержку во времени. Значительно более удобны цифровые системы, в которых речевой сигнал записывается в дискретной форме. При этом объем требуемой памяти пропорционален длительности множества сообщений. Несмотря на то, что стоимость и физические размеры запоминающих устройств падают, все же возникает задача экономного описания речевых сигналов через их параметры, при котором потребуется меньший объем памяти. Параметрическое представление, в свою очередь, связано с анализом и синтезом речевой волны, так что в действие вступает новый фактор — сложность этих процедур. Если множество сообщений фиксировано, то сложность анализа речевого сигнала не имеет никакого значения. Если же сообщение время от времени меняется, то стоимость и размеры параметрического синтезатора зависят не только от объема требуемой памяти, но и от сложности алгоритмов анализа и синтеза. Появление новых способов записи и хранения информации (например, лазерных дисков) может существенно ослабить влияние фактора стоимости памяти и, тем самым, снизить потребность в параметрическом представлении речевого сигнала, расширяя область применения систем вывода

речи без анализа параметров. Однако в любом случае остается обширная область применений, где необходим параметрический анализ и синтез речи.

Возможность записи речевого сигнала в более экономной форме основывается на его ограниченности и непрерывности. Это следует из первой теоремы Вейерштрасса, которая гласит, что если функция $f(x)$ непрерывна на конечном замкнутом интервале $[a, b]$, то всякому $\varepsilon > 0$ можно сопоставить многочлен $P_n(x)$ степени $n = n(\varepsilon)$, для которого на всем интервале $[a, b]$ имеет место неравенство $|f(x) - P_n(x)| \leqslant \varepsilon$. Это означает, что можно достигнуть равномерного приближения непрерывной функции $f(x)$ с заданной погрешностью ε, используя конечное число параметров приближающего многочлена $P_n(x)$. Запомнив n параметров этого многочлена, при синтезе $f(x)$ пользуются информацией о виде полинома $P_n(x)$, вследствие чего и создается экономия в требуемом объеме памяти. В первой теореме Вейерштрасса важным элементом является ограниченность интервала $[a, b]$ существования функции, в то время как речевой сигнал может длиться в течение практически бесконечного времени. Еще более важно то, что свойства речевого сигнала меняются во времени и бессмысленно пытаться найти полином, одинаково хорошо описывающий разные его участки.

Согласно второй теореме Вейерштрасса, для любой непрерывной функции $f(t)$ с периодом 2π существует такая тригонометрическая сумма

$$S_n(t) = a_0 + \sum_{k=1}^{n} (a_k \cos k t + b_k \sin k t),$$

которая для всех t удовлетворяет неравенству $|f(t) - S_n(t)| \leqslant \varepsilon$, причем $n = n(\varepsilon)$. Отдельные участки речевого сигнала близки к периодической функции, но периоды его компонент и их амплитуды также меняются во времени.

Трудности, связанные с особенностями речевого сигнала, привели к тому, что богатый аппарат теории аппроксимации функций практически не используется в параметрическом синтезе речи. Вместо этого заимствуются методы, первоначально предназначенные для сжатия требуемой частотной полосы в телефонных каналах. Кроме того, следует отметить, что подход к экономному описанию речевого сигнала как абстрактной функции сильно ограничен, поскольку для сохранения информационного содержания совсем не обязательно точно описывать форму речевого сигнала во времени или его спектр. Это способствовало развитию, может быть, менее математически строгих, но более содержательных подходов, учитывающих не только свойства самого речевого сигнала, но и свойства восприятия речи.

Обсуждая различные способы экономного описания речевого сигнала, почерпнутые из вокодерной телефонии, следует

учитывать одно серьезное отличие между задачами параметрического синтеза и экономной передачи речи. При разработке вокодеров нередко отвергаются такие описания, параметры которых непомехоустойчивы при передаче по телефонному каналу связи. Для синтеза речи этот фактор не играет никакой роли, поскольку канал передачи параметров отсутствует, и факторами, определяющими приемлемость метода, являются лишь степень сжатия информации и сложность алгоритмов анализа и синтеза.

Все методы экономного описания речевого сигнала можно условно разделить на три группы. Одна группа тяготеет к математической теории аппроксимации функций, стараясь по возможности точно описать форму сигнала во времени. К этой группе относится разновидность импульсно-кодовой и дельта-модуляции. В ней не используются в явном виде какие-либо модели речевого сигнала. Другая группа содержит эвристические алгоритмы, в которых форма восстанавливаемого речевого сигнала может подвергаться значительным искажениям, но сохраняется информация, необходимая для восприятия речи. К числу таких методов относится клиппирование — предельное ограничение сигнала, осуществляемое таким образом, что он представляется в виде последовательности прямоугольных импульсов одинаковой амплитуды и переменной длительности. Наконец, третья группа использует математические модели речевого сигнала и оценивает параметры этих моделей, как, например, в линейном предсказании.

Объем требуемой памяти для хранения речевых сообщений зависит от скорости передачи информации, обеспечиваемой параметрическим методом. Эта скорость находится в диапазоне от 100 Кбит/с для импульсно-кодовой модуляции (ИКМ) до 0,5 Кбит/с для векторного квантования, так что коэффициент сжатия может достигать 200 с соответствующим увеличением сложности алгоритмов анализа-синтеза. Для одного и того же метода параметрического описания можно получить разные коэффициенты сжатия, но его увеличение сопровождается ухудшением разборчивости и натуральности синтезированной речи. Поэтому, в конечном счете, выбор того или иного метода сжатия определяется условиями конкретной задачи. В последующих разделах мы обсудим характеристики некоторых параметрических методов.

§ 3.2. Импульсно-кодовая модуляция

Для того чтобы записать речевой сигнал в память вычислительной машины, его нужно сначала из непрерывной формы перевести в дискретную, т. е. взять отсчеты с некоторым интервалом по времени и квантовать эти отсчеты. Частота отсчетов зависит от способа интерполяции — чем сложнее этот

способ, тем реже могут браться отсчеты. Наибольшее распространение получил такой способ дискретизации, в котором восстановление исходной функции получается путем пропускания последовательности отсчетов через фильтр низких частот с частотой среза F_c. Переходная функция такого фильтра есть

$$h(t)=\sin(t/F_c)/t.$$

Согласно теореме Найквиста—Котельникова, если в спектре некоторой функции $f(t)$ не содержится частот выше F_c, то частота отсчетов $F_{отс}$ должна быть не меньше удвоенной частоты F_c, т. е. $F_{отс}\geqslant 2F_c$. *При этом гарантируется восстановление* исходной функции $f(t)$. Таким образом, перед дискретизацией функция $f(t)$ должна быть пропущена через низкочастотный фильтр с частотой среза F_c для ограничения спектра. Если же этого не сделать, то возникает явление наложения спектральных компонент, находящихся выше F_c, и восстанавливаемая функция $f(t)$ искажается.

В практической реализации такого способа дискретизации имеются противоречия, состоящие в том, что для эффективного подавления частот выше F_c нужно использовать фильтр с большой крутизной среза, а это создает помеху на частоте F_c. Поэтому применяют фильтры с более плавным склоном, используя быстрое падение спектра вокализованных участков речевого сигнала на частотах выше 5 кГц. Для сохранения в спектре дискретизированной функции частот до F_c рекомендуется использовать не минимальную частоту отсчетов $F_{отс}=2F_c$, а более высокую, например, $F_{отс}=(2{,}5-3)F_c$. Тогда полоса в 5 кГц обеспечивается частотой отсчетов не в 10 кГц, а в 12,5 кГц или 15 кГц. Высококачественное восстановление речевого сигнала достигается использованием частоты отсчетов в 20 кГц.

Каждый отсчет речевого сигнала квантуется и записывается в дискретной форме как степень двойки, причем максимальное число уровней квантования равно $N=2^n$. Величина n находится из условия перекрытия динамического диапазона речевого сигнала. Принимая этот диапазон близким к 90 дБ, получим $n=15$, поскольку увеличение показателя степени двойки на единицу соответствует возрастанию амплитуды сигнала вдвое, т. е. на 6 дБ.

Всякое квантование сопровождается шумом, так как непрерывное значение отсчета сигнала заменяется на ближайшее квантованное значение. Предполагая равномерное распределение шума квантования, в [48] получили следующую зависимость отношения сигнал/шум в децибелах от порядка квантования n: $SNR=6n-7{,}2$.

Таким образом, при $n=15$ отношение сигнал/шум равно 82,8 дБ, т. е. шумы квантования практически не слышны. Аналого-цифровые и цифро-аналоговые преобразователи на 15 и даже 16 бит существуют, хотя и довольно дороги.

В действительности порядок квантования n может быть несколько ниже, поскольку динамический диапазон речи каждого конкретного диктора меньше 90 дБ, особенно, если учесть ограничения на стиль произношения при записи текста, предназначенного для воспроизведения. Практически высокое качество синтезированной речи достигается при показателе степени $n=12$. Отсюда можно оценить скорость передачи информации в импульсно-кодовой модуляции с равномерными отсчетами и уровнями квантования:

$$I = F_{\text{отс}} \cdot n \text{ бит/с}.$$

Величина I указывает, какой объем памяти в двоичных единицах требуется для хранения одной секунды речевого сигнала. Принимая частоту отсчетов $F_{\text{отс}}=20$ кГц и порядок квантования $n=15$, получим $I=300$ Кбит/с или 37,5 кбайт/с. Это означает, что память объемом в 500 Кбайт может хранить около 13 с речи или около 10 слов, т. е. две — три фразы. Конечно, это очень маленький речевой запас, и даже уменьшение частоты отсчетов до $F_{\text{отс}}=10$ кГц и порядка квантования до $n=12$ ненамного увеличивает этот запас — всего лишь до 32 с.

Ценой некоторого усложнения процедуры квантования можно значительно уменьшить порядок квантования n. Достигается это путем неравномерного квантования или нелинейного преобразования с сохранением равномерного квантования сигнала до квантования и при восстановлении непрерывной функции. При равномерном квантовании шаг квантования постоянен и не зависит от уровня сигнала. Поэтому относительная ошибка мала для больших уровней сигнала и велика для малых уровней, из-за чего и приходится использовать большое число градаций. Естественно потребовать, чтобы относительная ошибка была постоянной для всех уровней сигнала. Очевидно, *что при этом уровни квантования должны быть распределены* логарифмически, или же, при сохранении равномерного квантования, сигнал предварительно должен быть подвергнут логарифмическому преобразованию, а перед синтезом — обратному преобразованию. При такой трансформации сигнала в абсолютных величинах достигается большая точность *представления малых уровней сигнала и меньшая — для больших* уровней, что соответствует и свойствам восприятия.

В силу неприменимости логарифмического преобразования к отрицательным величинам и величинам, близким к единице, обычно пользуются так называемым μ-законом сжатия:

$$y(t) = f_{\max} \frac{\log\left[1+\mu |f(t)|/f_{\max}\right]}{\log(1+\mu)} \operatorname{sign}\left[f(t)\right],$$

где $f_{\max}$ — максимальное значение сигнала, sign — знаковая

функция:

$$\operatorname{sign} f=\begin{cases} 1, & f>0, \\ 0, & f=0, \\ -1, & f<0, \end{cases}$$

а величина μ принимает значения от 30 до 500. Оценка относительной величины шума квантования показывает, что при $\mu=500$ неравномерное квантование на 7 разрядов эквивалентно равномерному квантованию с 11 разрядами. Таким образом, получается экономия требуемой памяти на 1 с речи более, чем в 1,5 раза.

Нелинейное квантование соответствует мгновенному сжатию речевого сигнала, но не принимает в расчет его средний уровень на данном отрезке времени. Можно улучшить отношение сигнал/шум или сэкономить еще один разряд в квантователе, если адаптировать величину шага квантования к среднему уровню. С этой целью вычисляется текущее среднее квадратическое отклонение сигнала, и шаг квантования устанавливается пропорционально ему, причем вводится новая величина — коэффициент усиления, обратно пропорциональный этой дисперсии, и средний уровень сигнала нормируется путем изменения коэффициента усиления. Выбирая соответствующий интервал усреднения, обычно равный 10—20 мс, добиваются адаптации к текущему уровню громкости сигнала. Нечто подобное выполняется в среднем ухе человека, где степень натяжения барабанной перепонки, т. е. ее чувствительность, меняется в соответствии с уровнем громкости звука. Выигрыш одного разряда в квантователе покупается ценой вычисления текущей дисперсии и коэффициента усиления, а также дополнительных операций по изменению уровня сигнала путем управления коэффициентом усиления.

Дальнейшее сжатие информации может быть получено, если учесть высокую степень корреляции, т. е. зависимости соседних отсчетов сигнала. Известно, что для любого отсчета речевого сигнала $f(t_i)$ в момент времени t_i значение отсчета этой функции в следующий момент времени t_{i+1} будет довольно близко к предыдущему отсчету. Это значит, что дисперсия разности соседних отсчетов меньше, чем дисперсия самого сигнала. Поэтому кодирование разности соседних отсчетов функции $f(t)$, а не самих отсчетов, при том же числе разрядов квантователя обеспечивает лучшее отношение сигнал/шум, т. е. лучшую точность описания сигнала. Еще большей точности можно добиться, предсказывая отсчет $f(t_{i+1})$ по m предыдущим отсчетам, и кодируя разность между предсказанным и истинным значениями. Однако сложность предсказания быстро растет с глубиной предсказания, т. е. с числом используемых отсчетов, тогда как наибольший прирост в точности дает простейший способ — кодирование разности $\Delta_{i,\,i+1}=f(t_{i+1})-f(t_i)$. Применяя

адаптивное квантование разности $\Delta_{i,i+1}$, можно получить выигрыш в отношении сигнал/шум, примерно равный 10—12 дБ, т. е. сэкономить еще 2 разряда квантователя. В итоге, с помощью адаптивно-разностной ИКМ можно сократить скорость передачи информации до 32 Кбит/с при отличном качестве звучания [38]. При этом оказывается, что при том же отношении сигнал/шум речевой сигнал в системе адаптивно-разностной ИКМ субъективно оценивается выше, чем в других системах ИКМ. Иными словами, количественная оценка системы анализа-синтеза речевого сигнала через отношение сигнал/шум не исчерпывает всех характеристик — важную роль играет также и способ аппроксимации сигнала.

Все рассмотренные выше способы сжатия касались кодирования отсчетов речевого сигнала при постоянной частоте отсчетов, что было связано со способом восстановления дискретизированного сигнала с помощью фильтра низкой частоты. Если же момент взятия следующего отсчета поставить в зависимость от изменения сигнала $f(t)$, например, так, как описывается в гл. 2, то возможно еще большее сжатие информации. Однако изменение расстояний между отсчетами выводит систему анализа-синтеза из рамок методов импульсно-кодовой модуляции, поскольку для аппроксимации сигнала уже используются другие функции.

§ 3.3. Дельта-модуляция

Дельта-модуляция является частным случаем разностной импульсно-кодовой модуляции, когда квантователь имеет только один разряд. Если частоту отсчетов сигнала значительно увеличить по сравнению с минимальной частотой по Найквисту — Котельникову, то соседние отсчеты окажутся настолько коррелированными, что для предсказания следующего отсчета достаточно указать лишь знак первой производной по времени. Кодирование и восстановление сигнала в дельта-модуляции осуществляется следующим образом. Определяется знак разности между отсчетами сигнала f в момент времени $t+\Delta t$ и оценкой сигнала $\bar{f}(t)$, т. е. $\delta=f(t+\Delta t)-\bar{f}(t)$. Если $\delta\geqslant 0$, то в память записывается 1, а при $\delta<0$ записывается 0. Значения функции $f(t+\Delta t)$ аппроксимируются как

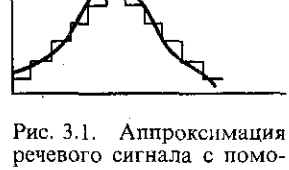


Рис. 3.1. Аппроксимация речевого сигнала с помощью дельта-модуляции

$$\bar{f}(t+\Delta t)=\bar{f}(t)+d,$$

где $d=\Delta$, если $\delta\geqslant 0$ и $d=-\Delta$, если $\delta<0$. Величина Δ определяет крутизну нарастания или спада сигнала (см. рис. 3.1). Очевидно, что между Δ и максимальной производной сигнала $f(t)$

существует прямая зависимость:

$$\frac{\Delta}{\Delta t} = \max \left| \frac{df}{dt} \right|.$$

Полагая $f(t) = A \sin \omega t$, можно оценить связь между порогом Δ и частотой ω как

$$\frac{\Delta}{\Delta t} = A\omega \cos \omega t|_{t=0} = 2\pi AF.$$

Приписывая F значение максимальной частоты в спектре сигнала, получим $\Delta = 4\pi A \Delta t / T_{\text{отс}}$, где $T_{\text{отс}}$ — интервал отсчета сигнала при дискретизации по Найквисту — Котельникову, A — относительная амплитуда спектральной компоненты сигнала на частоте F. По оценке [31] для телефонного канала на частоте 3500 Гц коэффициент $A \approx 0{,}076$.

При оптимальном выборе порога Δ максимальное отношение сигнал/шум квантования достигается при частоте отсчетов $F_{\Delta\,\text{отс}}$ примерно в 4—5 раз большей частоты отсчетов $F_{\text{отс}}$ по Найквисту — Котельникову [48]. Например, при частоте среза спектра $F_c = 3$ кГц и отношении сигнал/шум 35 дБ требуется скорость передачи 200 Кбит/с. Таким образом, классическая дельта-модуляция требует увеличения памяти для хранения речевого материала, и ее использование оправдывается лишь значительным упрощением процедур анализа-синтеза. Сжатие информации происходит путем подавления наиболее энергетических низкочастотных компонентов речевого сигнала операцией дифференцирования.

Увеличение порога Δ, с одной стороны, позволяет избежать ошибок при аппроксимации участков с быстрым изменением сигнала, но, с другой стороны, при отсутствии сигнала возникает так называемый шум дробления, связанный с попеременным переключением квантователя от $+\Delta$ на $-\Delta$. Следовательно, оптимальным было бы изменение порога Δ в соответствии с уровнем сигнала, т. е. адаптивная дельта-модуляция. При этом порог Δ изменяется пропорционально предыдущему значению, как $\Delta_i = M\Delta_{i-1}$, где

$$M = P > 1, \quad f(t_i) = f(t_{i-1}),$$

$$M = Q < 1, \quad f(t_i) \neq f(t_{i-1}),$$

и $PQ \leqslant 1$. Оптимум достигается при $1{,}25 < P < 2$ [48]. При скоростях передачи, меньших 40 Кбит/с, адаптивная дельта-модуляция обеспечивает лучшее отношение сигнал/шум, чем логарифмическая импульсно-кодовая модуляция. Практически, высокое качество синтезированного речевого сигнала сохраняется до скоростей не ниже 30 Кбит/с. Дальнейшее понижение

скорости ухудшает качество сигнала. Применяя различные схемы дельта-модуляции, можно несколько снизить скорость передачи при сохранении хорошего качества речи, но это покупается ценой усложнения анализа-синтеза [7].

§ 3.4. Клиппирование

Экспериментальным путем было установлено, что если пропустить речевой сигнал через усилитель с очень большим коэффициентом усиления и ограничить уровень, то такой предельно ограниченный (клиппированный) сигнал все же сохраняет определенную разборчивость. Поскольку положительный и отрицательный уровни сигнала при этом постоянны, то речевой сигнал выглядит как последовательность прямоугольных импульсов различной ширины. Ясно, что вся информация о речевом сигнале при этом заключается в интервалах между переходами через нуль, или мгновенной частоте. Если обозначить клиппированный сигнал как $\varphi(t)$, где

$$\varphi(t)=\begin{cases}1, & f(t)\geqslant 0,\\ 0, & f(t)<0,\end{cases}$$

то текущий спектр последовательности прямоугольных импульсов находится как

$$S(j\omega,\ t)=\int\limits_{t-T}^{t} f(\tau)\,e^{-j\omega\tau}d\tau=\frac{2}{\omega}\sum_k \varphi_k e^{-j\omega T_k}\sin\frac{\omega\Delta T_k}{2},$$

где ΔT_k — длительность k-го интервала, на котором функция φ_k сохраняет свой знак. Элементарные преобразования дают

$$S(j\omega, t)=\frac{1}{\omega}\sum_k \varphi_k\left[\sin\omega\theta_1+\sin\omega\theta_2-j\left(\cos\omega\theta_1-\cos\omega\theta_2\right)\right],$$

где $\theta_1=\Delta T_k/2-T_k$, $\theta_2=\Delta T_k/2+T_k$ [57]. При клиппировании речевого сигнала его спектр искажается из-за появления нечетных гармоник и, кроме того, изменяются амплитудные соотношения между частотными компонентами. Тем не менее, в спектре клиппированного сигнала сохраняется достаточно информации для различения звуков речи.

Вследствие того, что амплитуда низкочастотных компонент речевого сигнала обычно значительно больше амплитуды высокочастотных компонентов, при непосредственном клиппировании создаются нелинейные искажения, в результате которых высокочастотные компоненты спектра подавляются. Однократное или двукратное дифференцирование по времени поднимает уровень высокочастотных компонент, и разборчивость клиппированного сигнала заметно повышается. Естественно, что при восстановлении речевого сигнала нужно при-

менять обратные операции — однократное или двукратное интегрирование.

Клиппированный сигнал легко дискретизируется, исходя из тех же принципов, что и в импульсно-кодовой модуляции. Поскольку амплитуда клиппированного сигнала квантуется всего на два уровня (1 или 0), то скорость передачи информации равна просто частоте отсчетов. В [55] отмечается, что при скорости передачи в 25—30 Кбит/с клиппированная речь сохраняет высокую разборчивость, а при 8 Кбит/с — удовлетворительную. При этом, однако, натуральность клиппированной речи весьма низка. Натуральность повышается и почти не отличается от натуральности естественной речи, если перед клиппированием вычислить огибающую речевого сигнала, и при восстановлении наложить ее на клиппированный сигнал. Поскольку спектр огибающей очень узок — примерно 50—100 Гц, то для ее запоминания требуется очень мало информации — эта огибающая может быть закодирована одним из способов импульсно-кодовой или дельта-модуляции.

По простоте процедур анализа-синтеза и по скорости передачи информации клиппирование сравнимо с дельта-модуляцией.

§ 3.5. Линейное предсказание

Методы линейного предсказания позволяют уменьшить скорость передачи информации до 2,4 Кбит/с с сохранением приемлемой разборчивости и натуральности, хотя высокое качество требует скорости не ниже 9 Кбит/с. Это весьма существенное сжатие, позволяющее хранить речевой сигнал длительностью уже не в секунды, а в минуты, при технически вполне приемлемых объемах дискретной памяти. Поэтому, по крайней мере, по критерию требуемой памяти, методы линейного предсказания удовлетворяют требованиям большинства задач параметрического синтеза речи.

Сокращение скорости передачи информации в 10—20 раз по сравнению с прямым методом импульсно-кодовой модуляции, конечно, покупается ценой усложнения процедур анализа. В основе методов линейного предсказания лежит довольно простое предположение, что i-й отсчет речевого сигнала может быть представлен как сумма n предыдущих отсчетов, взвешенных с коэффициентами a_k, и сигнала возбуждения u_i:

$$f_i = \sum_{k=1}^{n} a_k f_{i-k} + G u_i, \tag{3.1}$$

где G — коэффициент усиления. Если удается определить такие n и оценки коэффициентов $\bar{a}_k$, что ошибка предсказания

$$\varepsilon = f_i - \sum_{k=1}^{n} \bar{a}_k f_{i-k}$$

будет минимальна в некотором смысле, то процедура синтеза

речи по (3.1) не представляет никаких трудностей. Этого нельзя сказать о процедуре анализа, которая, к тому же, опирается на некоторые предположения о свойствах речевого сигнала, т. е. на некоторую математическую модель. Методам линейного предсказания посвящена обширная литература, в том числе [35, 48], поэтому в данном разделе будут только кратко рассмотрены основные характеристики, существенные для параметрического синтеза речи.

Известно, что если некий сигнал образуется суммированием M комплексных компонент, то для вычисления параметров этих компонент достаточно взять $2M$ отсчетов сигнала. Интерпретируя для речевого сигнала комплексные компоненты как затухающие резонансные колебания, получим, что для оценки параметров пяти формант нужно взять 10 отсчетов. Поскольку такое представление все же не совсем точно, то эмпирическое правило состоит в выборе 4M отсчетов, или $F_{отс}+4$, где $F_{отс}$ — частота дискретизации в кГц. На практике оказывается, что увеличение числа коэффициентов a_k более 14 не улучшает качество синтезированной речи, поэтому в действующих системах M находится в пределах от 10 до 14. Одновременно установлено, что увеличение частоты дискретизации выше определенного предела не только не улучшает качество синтеза, но даже ухудшает его.

Коэффициенты a_k определяются путем минимизации среднеквадратичной ошибки

$$\varepsilon_i = \sum_m \Big(f_{i+m} - \sum_{k=1}^{n} a_k f_{i+m-k} \Big)^2. \qquad (3.2)$$

Хотя и не очевидно, что среднеквадратичный критерий адекватен свойствам речевого сигнала, его, как обычно, используют вследствие аналитической простоты, приводящей к системе линейных уравнений. Взяв частные производные от ε_i по a_k, т. е. $\partial\varepsilon_i/\partial a_k$, и приравняв их нулю, получаем

$$\sum_m f_{i+m-j} f_{i+m} = \sum_{k=1}^{n} a_k \sum_m f_{i+m-j} f_{i+m-k}, \quad 1 \leqslant j \leqslant n. \qquad (3.3)$$

Существует ряд методов решения (3.3) таких, как автокорреляционный, ковариационный, частных корреляций и т. д. Отличаясь в исходных предпосылках относительно свойств модели речевого сигнала, они обладают примерно одинаковыми характеристиками, хотя имеются и некоторые существенные отличия.

В автокорреляционном методе вычисления производятся в окне конечной длительности типа Хемминга (с плавно спадающей до нуля весовой функцией к краям окна) на интервале нескольких периодов основного тона. Ковариационный метод использует более короткий интервал 2—5 мс, но для надлежащей точности анализ нужно производить синхронно

с основным тоном. В методе частных корреляций предсказание осуществляется не только вперед, но и назад, и минимизируются ошибки для удвоенного набора коэффициентов — для прямого и обратного предсказания. Во всех методах в явном виде используется модель передаточной функции речевого тракта $H(z)$, содержащая лишь одни полюса, т. е.

$$H(z)=\frac{G}{1-\sum_{k=1}^{n} a_k z^{-k}}, \tag{3.4}$$

где z^{-1} — означает задержку сигнала на один отсчет.

При решении системы уравнений типа (3.3) возникает ряд трудностей, связанных с плохой обусловленностью матриц — отношения максимального и минимального собственных значений, а также конечной точностью представления чисел в ЭВМ. Установлено, что при анализе вокализованных звуков при увеличении частоты дискретизации точность вычислений должна быть повышенной. Применяя предыскажение речевого сигнала в виде дифференцирования, можно несколько понизить требования к точности.

При автокорреляционном методе вычисления коэффициентов линейного предсказания требуется число операций (умножений), пропорциональное n^2, а при ковариационном методе — пропорциональное n^3, однако сложность сопутствующих процедур такова, что на ковариационный метод в среднем требуется лишь в 1,4 раза больше операций, чем на автокорреляционный метод.

После того как найдены коэффициенты предсказания, определяется признак тон/шум, т. е. тип источника возбуждения — голосового или шумового. Для голосования источника необходимо определить частоту основного тона, которая обычно вычисляется по пику автокорреляционной функции сигнала ошибки ε. Кроме того, вычисляется общий коэффициент усиления G.

Для синтеза речевой волны можно воспользоваться прямой формулой (3.1), однако она требует высокой точности вычисления коэффициентов предсказания. Требования к точности понижаются, если вместо коэффициентов предсказания вычислять коэффициенты отражения на границах цилиндрических секций, аппроксимирующих форму речевого тракта, поскольку между ними на основе (3.4) установлена однозначная связь — передаточная функция последовательности цилиндрических секций есть

$$V(z)=\frac{1}{A(z)}\prod_{k=1}^{N}(1+r_k)z^{-N/2},$$

где r_k — коэффициенты отражения, $r_k=(S_{k+1}-S_k)/(S_{k+1}+S_k)$, S_k — площадь сечения секции, $A(z)$ — знаменатель (3.4),

вычисляемый с помощью рекурсивной процедуры:

$$A^{0}(z)=1,$$
$$A^{(i)}(z)=A^{(i-1)}(z)-c_i z^{-i} A^{(i-1)}(z^{-1}),$$
$$A(z)=A^{n}(z).$$

Коэффициенты отражения r_k равны коэффициентам c_i, взятым с обратным знаком, $r_k=-c_i$. Необходимо отметить, однако, что попытки восстановления формы речевого тракта на основе коэффициентов отражения r_k часто приводят к нелепым результатам, из чего следует ограниченная точность самой модели (3.4).

При имитации голосового источника синтезатор возбуждается последовательностью коротких импульсов с частотой основного тона. Такой способ возбуждения придает синтезированному сигналу характерное жужжание, которое можно слегка уменьшить, используя не прямоугольные, а треугольные импульсы. Наилучшее же качество достигается, если для возбуждения используется сигнал ошибки, но при этом скорость передачи информации повышается до 7—9 Кбит/с. Вместо возбуждения сигналом ошибки можно использовать возбуждение, создаваемое той или иной моделью голосового источника. В зависимости от модели, т. е. от числа управляемых параметров, в сигнале ошибки должны определяться либо только частота основного тона, либо и такие параметры, как длительность интервалов открытой и закрытой голосовой щели, амплитуды положительного и отрицательного импульсов возбуждения, уровень и длительность турбулентных шумов и их расположение относительно интервала с открытой голосовой щелью. Чем более адекватная модель голосового источника используется, тем большее сокращение требуемой памяти достигается при сохранении хорошего качества синтезированного речевого сигнала. В гл. 5 будут обсуждаться модели голосового источника, которые могут быть использованы не только в формантных и артикуляторных синтезаторах, но и в параметрических синтезаторах с линейным предсказанием.

Качество речевых сигналов у синтезаторов с линейным предсказанием весьма далеко от качества естественной речи. Например, при 10 коэффициентах предсказания и скорости 2,4 Кбит/с, разборчивость синтетической речи на 40% ниже разборчивости естественной речи как при низком, так и при высоком уровне шумов в канале связи [137]. Это является следствием, в первую очередь, чрезмерно упрощенной модели речевого тракта в виде системы, передаточная функция которой содержит только полюса. Системы линейного предсказания, в которых кроме полюсов, учитываются также и нули, обладают более высоким качеством. Еще лучше звучит синтетическая речь, если, кроме полюсов и нулей в модели

речевого тракта используется так называемое многоимпульсное возбуждение [71]. Развиваются и нелинейные методы рекуррентного оценивания параметров речевого сигнала [46].

§ 3.6. Векторное квантование

Для того чтобы сохранить высокое качество речи при скоростях 8—16 Кбит/с или добиться максимального снижения скорости (до 500 бит/с) при достаточной разборчивости, используется векторное квантование. Методы линейного предсказания применяют скалярное квантование, когда каждый отсчет сигнала или параметр квантуется независимо от других. В векторном квантовании кодированию подвергается сразу некоторое множество отсчетов или параметров. Основная идея векторного квантования близка к методам формальной теории распознавания образов, где сигналы представляются в виде точек или векторов в многомерном пространстве, и тем или иным способом осуществляется разбиение этого пространства на области, соответствующие разным классам (образам). В соответствии с этим подходом, в векторном квантовании пространство речевых сигналов разбивается на некоторое множество областей, и все векторы, попадающие в ту или иную область, заменяются на вектор — эталон этой области. Задачи разбиения пространства сигналов на области, выбора эталона и классификации векторов решаются весьма сложными математическими приемами, описание которых выходит за рамки данной книги. Хорошее введение в проблемы векторного квантования дается в [34]. Отметим только сходство и различие с методами линейного предсказания.

Как в методе линейного предсказания, в векторном квантовании используется линейная зависимость между соседними отсчетами речевого сигнала и форма функции плотности вероятности распределения значений сигнала. Дополнительное сжатие информации достигается путем кодирования сигнала ошибки линейного предсказания и совместным кодированием сразу многих отсчетов. Кодирование сигнала ошибки эквивалентно использованию нелинейной зависимости между отсчетами сигнала, а совместное кодирование отсчетов принципиально опирается на многомерное представление. Нелинейная зависимость важна для квантования спектральных параметров, а многомерность — для квантования сигнала во временной области.

Высококачественный синтез речи может быть достигнут при скорости передачи в 8 Кбит/с, а качество, характерное для служебной связи — вплоть до скоростей в 1 Кбит/с. Такое сжатие, однако, сопровождается необходимостью использования сложных алгоритмов анализа речевых сигналов, вследствие чего использование векторного квантования в синтезаторах оправдано лишь там, где требования на минимизацию объема памяти наиболее жесткие.

§ 3.7. Компиляционный синтез

Параметрический синтез не дает возможности изменять сообщения произвольным образом, задавая заранее неизвестный текст. Попытки синтеза фраз из заранее записанных слов наталкиваются прежде всего на проблему памяти, поскольку количество словоформ исчисляется сотнями тысяч. Но даже если этот фактор не очень важен, например, вследствие использования дисков с лазерной записью, то в этом подходе остается принципиально непреодолимая трудность изменения частоты основного тона, длительности и акустических характеристик сегментов слов в зависимости от типа фразы и положения слова во фразе. Очевидной альтернативой является сборка произвольных речевых сообщений путем последовательного соединения достаточно простых элементов, выделенных из естественной речи. Такой способ синтеза называется компиляционным.

Исходная предпосылка в компиляционном синтезе состоит в том, что речевое сообщение унаследует натуральность и разборчивость элементов, из которых оно собирается. Здесь, однако, имеется ряд трудностей. Прежде всего необходимо определить оптимальный размер элементов с тем, чтобы, с одной стороны, учесть взаимное влияние звуков (коартикуляцию), а, с другой стороны, чтобы число этих элементов было не слишком велико. Затем должны быть найдены правила сшивания этих элементов в слитную последовательность без разрывов во временной и спектральной областях. И, наконец, нужно предусмотреть возможность изменения просодических характеристик — интенсивности, длительности и частоты основного тона. Качество компиляционного синтеза зависит от того, насколько удачно решаются эти задачи.

Минимальный размер элементов для компиляционного синтеза определяется протяженностью во времени коартикуляционных связей. Эти связи максимальны для соседних звуков, но взаимное влияние часто проявляется и на триадах — трех соседствующих звуках, а некоторые признаки, например, огубление, могут распространяться до четырех — пяти звуков. В соответствии с тем, какому виду коартикуляции отдается предпочтение, и выбираются элементы компиляционного синтеза.

Исходя из лингвистических соображений, в качестве элемента компиляции должен быть выбран слог, хотя четкое определение слога отсутствует. Подсчитывая слоги, содержащие по три и четыре звука, т. е. СГС, ССГС и СГСС (С — согласный звук, Г — гласный звук), получим число элементов, близкое к 5000, хотя вероятности появления разных слогов сильно различаются и это число может быть значительно уменьшено. В английском языке считается, что число различных слогов превышает 10 тысяч. Это очень большие словари, для

составления которых требуется много времени. К тому же, если имеется необходимость в синтезе различных голосов, то объем работы увеличивается пропорционально числу дикторов. Поэтому обычно используют другие, более короткие элементы компиляции.

Наиболее популярными элементами являются дифоны, предложенные в [172]. Дифон — это сегмент речевого потока, заключенный между серединами соседних звуков. Полагая, что в языке имеется около 40 звуков, легко подсчитать число всевозможных комбинаций пар звуков — оно равно 1600. Учитывая, что в любом конкретном языке встречаются не все возможные пары звуков, фактически число дифонов меньше 1600. Для русского языка словарь дифонов оценивается в 1000—1200 элементов. С другой стороны, число дифонов должно быть увеличено с тем, чтобы охватить безударные позиции и случаи наиболее сильной коартикуляции, распространяющейся более, чем на два звука.

Недостатки системы дифонов состоят в появлении разрывов на границе двух гласных, особенно в дифтонгах, где непрерывный переход от одного гласного к другому принципиально важен. Кроме того, трудно имитируется коартикуляция гласных с находящимся между ними согласным. Созданы системы компиляционного синтеза с приемлемой разборчивостью и натуральностью, например [142, 169], однако их характеристики заметно ниже, чем у естественной речи. Так, для дифонной системы [77] найдено, что слоговая разборчивость всего лишь около 66%, тогда как для естественной речи — примерно 93%. Вместе с тем, тщательный пересмотр системы дифонов и правил сшивания может повысить слоговую разборчивость до 80% [137].

Если дифоны записываются в раздельном и нейтральном произнесении, то вероятно появление разрывов на границах вследствие отсутствия коартикуляционных связей более высокого порядка. Если же дифоны вырезаются из слов, то они носят отпечаток контекста.

К дифонам близка другая система элементов, называемых полуслогами [104]. Их преимущество заключается в использовании кластеров согласных в слогах, но они плохо описывают коартикуляцию между слогами.

В японском, русском и итальянском языках наиболее вероятно появление открытых слогов, т. е. сочетаний согласный — гласный. Так, в русском языке 164 слога СГ покрывают 77% текста [24]. Добавляя небольшое число изолированных гласных, в таких языках можно попытаться синтезировать слитную речь из весьма малого словаря элементов. В японском языке достаточно использовать 96 слогов и 5 гласных. В итальянском языке используется 110 слогов СГ и 44 сдвоенных согласных, начальных и конечных гласных [203]. При использовании СГ-элементов, очевидно, учитывается лишь

коартикуляция на переходном участке от согласного к гласному, но не учитывается влияние гласного, расположенного перед согласным звуком. Переходы от гласных к согласным вообще не представлены в этой системе, так же, как и переходы между соседними гласными.

В результате этого количество информации о месте артикуляции согласных звуков в компилированном речевом сообщении уменьшается. В частности, конечные согласные, особенно звонкие, представлены в речевом сигнале почти исключительно переходами от предшествующих гласных, и отсутствие этих переходов в СГ-системе делает такие звуки неразличимыми. Правда, в русском языке мало минимальных пар слов, различающихся лишь конечными согласными звуками (типа «лог — лоб — лом»), и можно надеяться, что общее снижение разборчивости компиляционного синтеза окажется не слишком велико. Более серьезное влияние на разборчивость должно оказывать ухудшение различимости первого согласного в кластерах согласных типа ГСС. Использование короткого сегмента гласного после первого согласного в таких кластерах может повысить разборчивость.

Одна из основных проблем в компиляционном синтезе — правила сшивания элементов в слитном потоке речи. Эти правила зависят от того, в какой форме записаны сами элементы. Если они хранятся в виде отсчетов речевого сигнала, представленного во временной форме, то необходимо обеспечить отсутствие разрывов на границе элементов как во временной функции, так и в текущем спектре. Напрашивающееся решение в виде суммирования сшиваемых элементов с линейно возрастающими и падающими весами на некотором отрезке времени оказывается не самым лучшим, поскольку при этом могут возникать звуки неопределенного фонетического качества.

Управление интенсивностью и длительностью сегментов при сшивании не представляет особых трудностей, и имитация этих просодических признаков для безударных и редуцированных слогов осуществляется довольно просто. Однако управление частотой основного тона при таком способе хранения элементов практически невозможно, и нужно примириться с тем, что синтезированные фразы будут лишены акцентов (логического ударения) и признаков вопроса, восклицания и т. д. К тому же, хранение элементов компиляции в виде отсчетов речевого сигнала требует большого объема памяти — от 0,5 Мбайт для СГ-системы до нескольких Мбайт в дифонных системах.

Необходимость в управлении частотой основного тона, улучшении правил сшивания и сокращения объема памяти для хранения элементов приводит к использованию одного из способов сжатия речевого сигнала. Обычно с этой целью применяется метод линейного предсказания. Поскольку в линей-

ном предсказании параметры передаточной функции речевого тракта и голосового источника разделены, то легко осуществляется управление частотой основного тона. При этом интерполяция контура частоты основного тона F_0, заданного в отдельных точках, должна быть линейной в логарифмическом масштабе [48]. Не представляет трудностей и управление коэффициентом усиления.

При сшивании сегментов во временной области возникают определенные трудности. Пусть, например, i-й и j-й сегменты описываются векторами коэффициентов линейного предсказания A_i и A_j. Тогда процедуру сшивания, обеспечивающую непрерывный переход от одного сегмента к другому во временной области, можно записать как

$$A_{ij}(t)=(1-\alpha t)A_i+\alpha t A_j,$$

где t — текущее время, $T=1/\alpha$ — интервал времени, на котором формируется переход от i-го сегмента к j-му. В момент времени $t=0$, $A_{ij}=A_i$, т. е. в начале перехода, коэффициенты линейного предсказания соответствуют i-му сегменту, а при $t=T$, $A_{ij}=A_j$, т. е. в конце перехода, коэффициенты соответствуют j-му сегменту, причем переход осуществляется непрерывно, и, казалось бы, задача сшивания решена. Однако может случиться так, что в какой-то момент времени модуль какого-либо коэффициента a_k станет больше единицы, и тогда нарушится условие устойчивости фильтра, синтезирующего речевой сигнал по коэффициентам линейного предсказания. Для того чтобы избежать такую ситуацию, нужно от представления элементов компиляции через коэффициенты линейного предсказания перейти к представлению через коэффициенты отражения или связанные с ними значения площадей цилиндрических секций, аппроксимирующих форму речевого тракта (см. § 3.3).

Интервал сшивания T — интервал перекрытия характеристик соседних сегментов — измеряется несколькими десятками миллисекунд. Поэтому и при сшивании в пространстве коэффициентов отражения могут возникать нежелательные звуки, так же как и при сшивании во временной области. Избежать эти искажения довольно трудно, поскольку существует большое число вариантов сшивания, и их все нужно прослушать на предмет обнаружения посторонних звуков. Альтернативой появления лишних звуков является допущение разрывов в текущем спектре на интервале сшивания сегментов.

Серьезной проблемой является ухудшение качества синтеза при изменении частоты основного тона. Поскольку все характеристики голосового источника, кроме частоты основного тона, входят в коэффициенты линейного предсказания, то изменение периода следования импульсов голосового источника (а, значит, и их формы) подразумевает изменение этих коэффициентов. Иными словами, можно рассчитывать на

достаточно высокое качество синтеза (и то с точностью до моделей речевого сигнала) лишь для тех частот основного тона, для которых производилось вычисление коэффициентов линейного предсказания. Так, даже для автокорреляционного метода, где интервал анализа равен 10—20 мс, замена F_0 на другое значение при синтезе приводит к изменению частоты первой форманты более, чем на 8% и еще большим погрешностям в значениях ширины формант [137]. В ковариационном методе интервал анализа еще короче (2—5 мс) и, по-видимому, можно ожидать еще большей зависимости качества синтеза от частоты основного тона.

Таким образом, компиляционному синтезу речи присущи некоторые принципиальные недостатки, в результате которых качество синтеза далеко не достигает качества естественной речи, несмотря на использование элементов, выделенных из естественной речи. Вместе с тем, компиляционные синтезаторы предоставляют возможность синтеза произвольного, заранее не заданного текста, и обладают способностями к управлению просодическими характеристиками: интенсивностью, длительностью сегментов и частотой основного тона, что позволяет решать задачи, не доступные параметрическим синтезаторам.

ГЛАВА 4

ПРОСОДИЯ РЕЧЕВОГО СИГНАЛА

Значительная часть информации в речевом сигнале передается посредством так называемых просодических параметров: частоты основного тона, интенсивности и длительности фонетических сегментов. С помощью этих параметров формируется тип фразы (утвердительный, вопросительный), указывается логически выделенное слово, ударные слоги противопоставляются безударным и т. д. Опыт применения синтезаторов показывает важность соблюдения просодических правил, свойственных конкретному языку.

§ 4.1. Временная структура речи

Временные параметры речевого сигнала содержат информацию о фонетическом качестве звуков речи и позиции сегмента в слове или слова во фразе. Кроме того, эти параметры характеризуют индивидуальную манеру артикуляции диктора и его эмоциональное и физическое состояние. Наиболее ярким примером информативности временных параметров в русском языке является противопоставление ударных и безударных гласных по длительности. В шведском, эстонском, венгерском и некоторых других языках существуют короткие и длинные гласные, играющие смыслоразличительную роль.

Свойства системы управления артикуляцией и механики процессов речеобразования проявляются в изменении длительности соседних сегментов, влиянии источника возбуждения и способа образования согласного на продолжительность гласного звука. Необычное слово длиннее, когда оно появляется в речи впервые [201], скорость речи замедляется у актеров, имитирующих страх или горе [203]. В зависимости от условий и темы разговора человек артикулирует с разной скоростью, которая зависит не только от механических характеристик артикуляторных органов, но и типа нервной системы и чувства времени.

Различия в скорости артикуляции у каждого диктора и еще большие различия в скорости между дикторами

свидетельствуют о необходимости регулирования темпа синтеза речи. Длительность различных звуков и ее коартикуляционные изменения являются важными источниками информации для распознавания речи. Известно, например, что в английском языке принятие решения о звонкой или глухой смычке производится, в основном, по длительности интервала между взрывом и началом фонации, а также по удлинению предшествующего гласного.

Временные параметры звуков играют большую роль в автоматическом синтезе речи. Например, в [84] было показано, что несоблюдение временных соотношений приводит не только к ухудшению натуральности синтетической речи, но и к падению разборчивости на 18%. В частности, на оценку натуральности влияют длительность гласных и интервал между началами ударных гласных во фразе [134].

Исследование временной организации артикуляторных программ дает возможность установить закономерности, общие для всех систем управления движениями человека. Нарушение в каком-либо звене системы управления артикуляцией проявляется и в изменении длительности сегментов речевого сигнала. Поэтому анализ временной структуры речи может также служить и диагностическим инструментом. Временные соотношения в речи человека, говорящего на иностранном языке, обычно остаются теми же, что и в его родном языке, составляя часть «акцента», так что знание временной структуры языка может быть использовано при обучении иностранному языку или распознавании происхождения человека.

Исследованию временной структуры речи посвящен ряд работ, из которых следует, что имеются существенные различия как между разными языками, так и между речью людей, говорящих на одном языке. В большинстве известных работ рассматриваются один—два фактора, влияющих на длительность сегментов речи, причем либо дается только качественное описание, либо данные представляются в численном виде без анализа механизмов изменения длительности. Исключение составляют работы [133, 134], где приводятся правила для синтеза длительности сегментов в различных контекстах, и работы [88, 144], в которых предпринимается попытка поиска механизмов, управляющих длительностью сегментов в речевом потоке. Исследованию некоторых аспектов временной структуры русского языка были посвящены работы [3, 19, 49, 66], однако много еще остается неизвестным.

Поскольку имеется более десяти факторов, влияющих на длительность сегментов [136], табличное описание всех возможных вариантов неосуществимо. Единственным способом синтеза длительности, сохраняющим параметры естественной речи, является обнаружение и количественное описание механизмов, порождающих изменение временных параметров. Изменения длительности, создаваемые системой управления специ-

ально для улучшения распознаваемости фонетических элементов, зависят от языка и от контекста. Механические характеристики артикуляторных органов проявляются в любом языке и зависят лишь от диктора. Энергетические критерии, связанные с мышечными усилиями, в том числе дыхательных мышц, также влияют на длительность сегментов. Некоторые особенности временных структур связаны с параметрами оперативной памяти, а другие — с индивидуальными тактиками систем управления артикуляцией у разных дикторов. Наконец, имеется довольно значительная случайная компонента, определяемая свободой выбора управляющих моторных команд, и связанная с кодовой избыточностью речевого потока [59].

Данный раздел, в основном, использует результаты работы [61], в которой была предпринята попытка изучения механизмов управления временной структурой речевого сигнала на основе различных экспериментальных методик. Для синтеза речи наиболее важным является формулировка правил, управляющих длительностью сегменов речевого сигнала для русского языка.

Экспериментальные методики. Одной из технических проблем при измерении длительностей в речевом сигнале является сегментация на участки, соответствующие изучаемым фонетическим элементам. Величина погрешности такой сегментации должна быть меньше воспринимаемой на слух разницы в длительности сегментов. Согласно [133], минимальная разница в длительности, замечаемая аудиторами, равна 25 мс или, в относительных единицах, около 20%. По другим данным, порог воспринимаемого различия в длительности составляет примерно 10 мс [49]. Таким образом, погрешность измерения длительности сегментов не должна превышать 10 мс.

На осциллограммах звукового давления некоторые фонетические элементы с большим трудом поддаются сегментации, в результате чего возникают большие ошибки в измерении длительности. В тех экспериментах, где используются осциллограммы, необходимо ограничивать набор изучаемых фонетических элементов только теми звуками, сегментация которых не представляет больших трудностей. Для остальных звуков должны использоваться другие формы представления речевого сигнала, например, видимая речь, огибающие в некоторых частотных полосах и др. В данном разделе описываются результаты измерений на речевом сигнале, представленном четырьмя различными способами.

Эксперимент 1. Слоги ГСГ, где в качестве гласных использовались звуки /А/ или /О/ с ударением на первом гласном, а в качестве согласных — /Б, Д, Г, М, Н, В/, в однократном произнесении были записаны от 30 дикторов на динамический микрофон, причем 10 дикторов записывались в заглушенной камере, а остальные — в обычной комнате. Затем речевой сигнал с помощью 48-канального спектрографа

был представлен в форме видимой речи и записан на кинопленку со скоростью 6 см/с. С кинопленки были сделаны отпечатки с увеличением масштаба примерно в два раза, и на них выполнялись измерения с точностью 5 мс.

Эксперимент 2. Трехсложные звукосочетания типа БАсгБА, где с — согласный /Б/ или /П/, а г — ударные гласные /А, О, У, И/, записывались в заглушенной камере одновременно с электромиограммами лицевых мышц для одного диктора. Всего было записано по пять реализаций каждого звукосочетания в нормальном и ускоренном темпе. Точность измерения на осциллограмме звукового давления составляла 10 мс.

Эксперимент 3. Слоги АБА, АПА, АДА, АТА, АГА, АКА с ударением на первом или втором гласном записывались в контексте «вот аба снова», произносимом как слитная фраза. Трое мужчин и три женщины повторяли весь список 3—4 раза в нормальном темпе. Запись велась в заглушенной камере на динамический микрофон, а затем речевой сигнал был представлен в виде осциллограммы звукового давления, огибающей и контура основного тона с помощью интонографа, разработанного в МГПИИЯ им. М. Тореза. Точность сегментации составляла 2 мс.

Эксперимент 4. В этом эксперименте, а также в экспериментах 5 и 6 участвовали 3 диктора (мужчины). Запись велась в обычной комнате на динамический микрофон. Речевой сигнал пропускался через набор фильтров высокой частоты с граничными частотами 100, 450, 900, 1 500 и 2 500 Гц, огибающие сигналы с выхода каждого фильтра вместе с осциллограммой звукового давления записывались на шлейфный осциллограф. Это устройство, дающее артикуляторно-ориентированное первичное описание речевого сигнала, описано в [25]. Такое представление речевого сигнала позволяло надежно сегментировать фонетические элементы. Точность измерений составляла 2 мс.

В слитных фразах «это СГС опять» и «вот ГСГ снова» исследовались все согласные звуки, а в качестве гласных использовались /А/ и /И/, причем в сочетании ГСГ ударение было на последнем гласном.

Эксперимент 5. В слитных фразах «вот $Г_1СГ_2$ снова» с ударением на первом гласном $Г_1$ в качестве согласных С использовались звуки /Т, Д, С, З/, а в качестве гласных — /А, О, У, Ы, Э, И, Е/ во всех возможных сочетаниях.

Эксперимент 6. В слитных фразах «это $С_1ГС_2$ опять» с гласными /А, И/ и всевозможными сочетаниями согласных /С, Ф, П, З, М, Б/ исследовалась роль степени открытости гласных.

Во всех экспериментах при измерениях в длительность гласного включался и переходный процесс от согласного звука. Длительность согласного определялась как длительность смычки или интервала с шумовым возбуждением. На видимой

речи для большинства записей было обнаружено, что в конце последнего гласного в течение нескольких десятков миллисекунд продолжаются колебания в низкочастотной области после того, как формантные колебания прекратились. Во временной области это проявляется в виде экспоненциально затухающих колебаний почти синусоидальной формы. Причиной этого является разведение голосовых складок из положения фонации в дыхательное положение, в результате чего в процессе их колебаний складки перестают соприкасаться, спектр импульсов голосового источника сужается, и формантные колебания не возбуждаются. При измерениях длительности последнего гласного эти затухающие колебания не учитывались.

Временная организация высказывания. В этом разделе описываются данные о высокой точности реализации длительности высказывания при относительно большой изменчивости длительности составляющих сегментов, а также эффекты укорочения сегментов при увеличении их числа в высказывании и удлинения конечных элементов высказывания. Предлагается гипотеза о механизмах этих явлений.

Длительность одних и тех же фраз типа «вот ГСГ снова» в эксперименте 3 при повторении каждым диктором подвергается весьма незначительным изменениям. Максимальное отклонение от среднего значения для каждого диктора находится в диапазоне 1%—11%, в то время как длительность сегментов колеблется в пределах от 4% до 34%. Стабильность длительности синтагм при повторении отмечалась в [49] (среднеквадратическое отклонение около 3%) при большой относительной ошибке длительности звуков (10—20%). В английском языке вариации длительности коротких фраз составляют примерно 15% при изменчивости длительности звуков около 13÷33% [69, 82]. Аналогичное явление наблюдается и в эксперименте 2, где длительность звукосочетаний при повторном произнесении изменялась на 4—10% как для нормального, так и для быстрого темпа. Длительность же гласных и согласных при этом изменялась в диапазоне 6—40%.

Необходимо отметить, что и в эксперименте 2, и в эксперименте 3 повторялся весь список, так что одни и те же звукосочетания не находились рядом во времени. Это исключает возможность кратковременной настройки на заданную длину высказывания.

Большая точность при повторной реализации высказывания — слова и фразы, говорит о том, что длительность высказывания задается в момент планирования моторной программы этого высказывания. Для обеспечения заданной длительности высказывания в процессе реализации моторной программы должны предусматриваться оценки текущей длительности в некоторых контрольных точках и компенсация отклонений от заданной временной программы. Контролируемыми

элементами могут быть длительности слогов, слов или расстояния между ударными слогами.

В эксперименте 1 оказалось, что средняя длительность сочетаний ГС постоянна почти для всех согласных и равна 365—370 мс (в среднем по всем дикторам), причем увеличение длительности гласного сопровождается уменьшением длительности звонкой смычки: АБ = 240 + 125 мс, АМ = 250 + 115 мс, АВ = 280 + 85 мс, АГ = 255 + 110 мс, АД = 255 + 115 мс. Как видно, хотя колебания длительности звонкой смычки невелики (кроме согласного /В/), они последовательно сопровождаются компенсационными изменениями длительности гласного, так что длительность конструкции ГС остается постоянной.

Если этот эффект не случаен, а соответствует некоторому механизму в реализации моторной программы, то он должен особенно проявляться при значительных изменениях длительности гласного, когда он находится в ударной позиции. Действительно, в эксперименте 3 длительность сочетаний АСоглА оказалась в среднем равной 350 мс, а длительность АСоглА с ударением на первом гласном — 295 мс, т. е. сочетания с первым ударным элементом сокращаются на 55 мс. Длительность ударного гласного А при этом практически постоянна и независима от позиции, тогда как длительность смычки заударного согласного сокращается примерно на 25 мс, независимо от звонкости или глухости, и на столько же укорачивается заударный гласный.

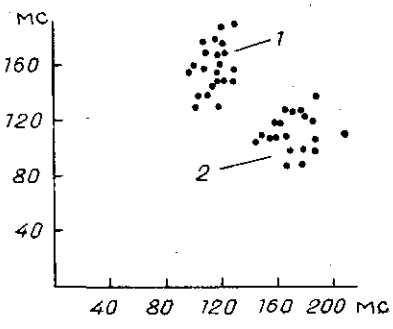


Рис. 4.1. Зависимость длительности смычки согласного /Б/ от длительности предыдущего гласного /А/: *1* — предударный /Б/, *2* — заударный /Б/

Аналогичная компенсация наблюдается и в эксперименте 2 в звукосочетаниях БАБАБА, где длительность звонкой смычки в сочетании АБ с ударной гласной почти обратно пропорциональна длительности гласного (рис. 4.1). Другие примеры согласованного изменения длительности гласного и согласного в сочетании ГС в зависимости от звонкости или глухости согласного получены в экспериментах 2—6. Если по временной оси отложить накопленную длительность звуков в сочетании $\Gamma_1 С \Gamma_2$ таким образом, что $T_{\Gamma_1 с} = T_{\Gamma_1} + T_с$, а $T_{\Gamma_1 с \Gamma_2} = T_{\Gamma_1} + T_с + T_{\Gamma_2}$, где T_{Γ_1}, $T_с$ и T_{Γ_2} — длительности звуков в одной и той же реализации сочетания $\Gamma_1 С \Gamma_2$, то гистограммы значений длительности для многих реализаций этого сочетания выявляют обратную зависимость длительности гласного Γ_1 и следующего за ним согласного С. В эксперименте 3, как видно из рис. 4.2, пики наиболее вероятных длительностей гласного Γ_1 и следующего за ним

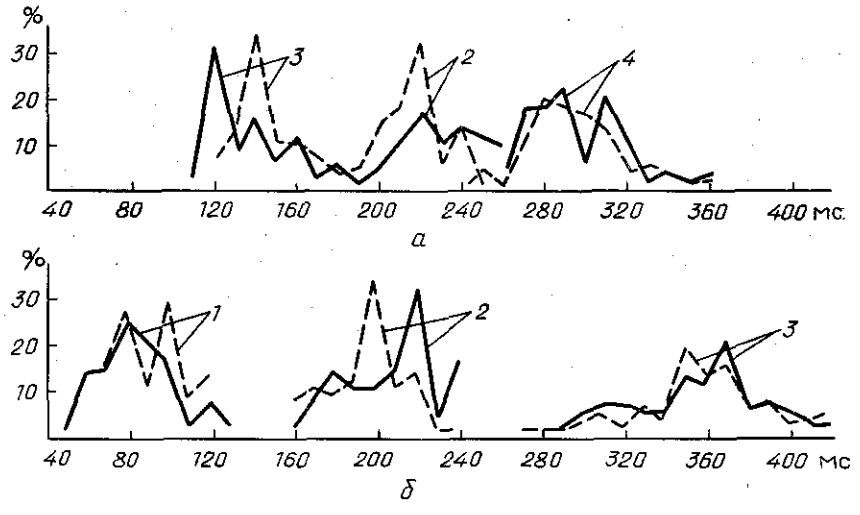


Рис. 4.2. Совместные распределения длительности звуков в слогах ГСГ (*а*) и ГСГ (*б*), ----слоги со звонкими взрывными, ———слоги с глухими, *1*—предударный гласный, *2*—согласный, *3*—ударный гласный, *4*—заударный гласный; нормальный темп

звонкого согласного сближаются, а пики длительностей гласного и глухого согласного раздвигаются примерно на 20 мс. В других экспериментах это различие доходит до 30 мс. Эта величина не зависит от того, является ли гласный $Г_1$ перед смычкой ударным или безударным. В противоположность этому явлению длительность гласного $Г_2$ следующего за согласным С не обнаруживает компенсационных изменений ни в ударной, ни в безударной позиции. Изменение темпа артикуляции мало влияет на величину компенсации, как это видно из рис. 4.3, полученного по результатам эксперимента 2.

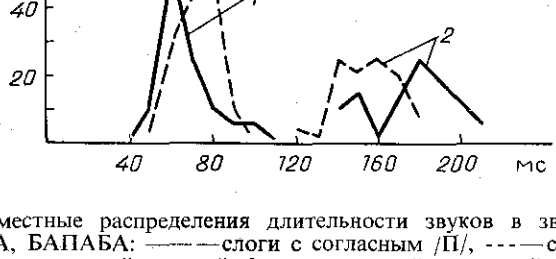


Рис. 4.3. Совместные распределения длительности звуков в звукосочетаниях типа БАБАБА, БАПАБА: ———слоги с согласным /П/, ----слоги с согласным /Б/, *1*—предударный гласный, *2*—предударный согласный; быстрый темп

О наличии отрицательной корреляции между длительностью гласных и согласных в сочетаниях ГС и СГ сообщается в [49]. В японском языке зависимость в сочетании СГ выражается как $T_{сг} = T_с + kT_г$, где $T_{сг}$ — длительность слога, $T_с$ — длительность согласного, $T_г$ — длительность гласного, $k \approx 2$ [115]. При этом так же, как и в [49], отмечалось, что полной компенсации изменения длительности одного элемента длительностью другого не существует. Обратная зависимость длительности гласного и согласного в сочетаниях ГС наблюдается в германских языках — английском, немецком, шведском и других [44]. В [150] найдено, что в английском языке длительность сочетания ГС больше длительности сочетания СГ. В шведском языке в слоге с короткой гласной используется только длинная согласная и наоборот [85].

Независимость длительности последующего гласного от звонкости или глухости предшествующего согласного при почти строгой обратной пропорциональности длительностей гласного и согласного в сочетании ГС наводит на мысль о том, что причиной этого явления служит сдвиг границы между гласным и согласным при постоянстве длительности сочетания СГС.

В сочетаниях СГ также имеется зависимость между длительностью согласного и длительностью гласного, причем здесь гласный влияет на согласный. В эксперименте 4 для сочетаний $Г_1СГ_2$ при одном и том же гласном $Г_1$ (/А/) согласный был длиннее с последующим коротким /И/ и короче с последующим /А/ (табл. 4.1).

Таблица 4.1. Зависимость длительности согласного от последующего гласного, мс

Согласный	И	Ы	У	А
С	124	117	114	107
Т	118	107	101	95
З	78	71	70	69
Д	80	84	78	73

Компенсационные явления могут распространяться и через звук, как это проявилось в эксперименте 3, с укорочением согласного и гласного, следующих за ударной гласной. Очевидно, что укорочение заударного гласного связано с необходимостью разделить требуемую компенсирующую длительность между согласным и гласным, поскольку длительность смычки согласного в нормальных условиях не может быть меньше некоторой величины (порядка 30—60 мс, по данным [49]).

Таким образом, эффект взаимной компенсации вариаций длительности соседних элементов в речи, по-видимому, является общим для многих языков и отражает действие некоторого

механизма, связанного с реализацией моторной программы в процессе речеобразования.

Аналогичное явление отмечается и в управлении движениями человека. Точность поддержания позиции, например, пальца или кисти руки оказывается значительно выше точности поддержания суставных углов руки. Это объясняется тем, что заданная позиция пальца может быть достигнута при относительно свободном выборе суставных углов всех элементов руки при условии, что значения каждого суставного угла планируются не отдельно, а одновременно со всеми другими, создавая, таким образом, возможность компенсации ошибок [13].

Фиксация длительности высказывания и контроль за выполнением временной программы на слогах и, возможно, словах подразумевает существование критерия, зависящего от длительности высказывания. Об этом свидетельствует и анализ явлений укорочения средней длительности сегментов при увеличении их числа в высказывании и удлинения последних сегментов в высказывании.

Средняя длительность изолированно произнесенных слогов ГСГ в эксперименте 1 составляет, в среднем, 540 мс (550 — для слогов с гласной /А/ и 530 мс — для слогов с гласной /О/). Средняя длительность конечного слога АБА в эксперименте 2 в звукосочетаниях типа БАБАБА равна 450 мс, а длительность этого же слога во фразе «вот аба снова» в эксперименте 3 равна 290 мс. Таким образом, наблюдается укорочение сочетания АБА в зависимости от длительности высказывания, в которое это сочетание включено.

Укорочение сегментов в зависимости от их числа в высказывании для английского и шведского языков описано в [101, 133, 134, 135, 144, 160], причем отмечается, что длительность данного сегмента находится в зависимости от числа сегментов, расположенных после него. В [144] эффект укорочения сегментов объясняется тем, что оперативная память имеет конечные размеры, хотя и может немного увеличиваться, т. е. она «эластична». Эта гипотеза представляется слишком механической.

В наших экспериментах относительная разница в длительности изолированного слога АБА и того же слога в звукосочетании БАБАБА составляет около 18%, хотя маловероятно, чтобы для звукосочетания БАБАБА объем оперативной памяти оказался недостаточным. Скорее, пропорционально числу сегментов растет «стоимость» записи моторной программы в оперативную память. Действительно, если измерить среднюю длительность сегментов в высказываниях АБА, БАБАБА, ВОТ АБА СНОВА, то окажется, что она обратно пропорциональна числу сегментов в высказывании, причем линейность соблюдается достаточно хорошо (рис. 4.4). При этом число сегментов считалось равным числу звуков плюс 1, поскольку необходимо

учитывать команды на переход от нейтрального состояния речевого аппарата к первому звуку и переход к нейтральному состоянию после последнего звука. Механически продолжая линейную зависимость с наклоном −4 мс/сегм на рис. 4.4, получим нулевую длительность сегментов при их числе в высказывании равном 34—36, так что такая экстраполяция недействительна для согласных, хотя, как известно, гласные звуки могут редуцироваться до полного выпадения из речевого потока.

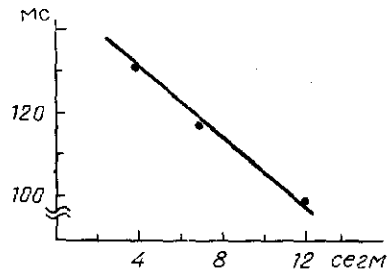


Рис. 4.4. Средняя длительность сегментов в зависимости от их числа в высказывании

При больших длительностях высказывания начинают влиять ограничения на продолжительность произнесения на одном выдохе, но, по некоторым данным, этот факт слабее ограничений на используемую оперативную память. В частности, конечное удлинение сегментов высказывания иногда наблюдается и без наличия паузы между двумя синтагмами. Это означает, что на одном выдохе произносится две синтагмы, каждая из которых планируется, исходя лишь из загрузки оперативной памяти.

Предположение о линейной зависимости стоимости оперативной памяти от длительности высказывания и, соответственно, стремлении системы управления сократить эту стоимость, подкрепляется различием длительности произнесения одного и того же высказывания разными дикторами. Время произнесения слогов ГСГ в эксперименте 1 разными дикторами колеблется от 375 мс до 735 мс. Ясно, что на скорость артикуляции влияют механические характеристики артикуляторных органов, а они различны у разных дикторов. Однако более интересен отмечаемый психолингвистами факт зависимости скорости речи от субъективного чувства времени. Рассмотрим подробнее возможный механизм этой зависимости.

На определенном этапе планирования высказывание представляется во вневременной форме, и лишь затем осуществляется его развертка во времени, что следует из анализа оговорок [103]. Это означает, что моторные команды для каждого комплекса артикуляторных движений извлекаются из долговременной памяти и помещаются в оперативную память. Не вдаваясь в обсуждение того, в какой форме моторные команды записываются в оперативной памяти, поскольку для этого нет данных, отметим лишь, что эти команды могут быть представлены своими отсчетами с последующей интерполяцией для достижения непрерывности. Если допустить, что частота отсчетов команд во времени является характерным свойством каждого человека, то окажется, что при большей

частоте (соответствующей более быстрому субъективному времени) для записи моторной команды одной и той же длительности в оперативную память потребуется больше места, чем при меньшей частоте отсчетов. В результате действия критерия экономии оперативной памяти, человек с более быстрым субъективным временем произнесет некоторое предложение за более короткий промежуток времени, чем человек с более медленным восприятием времени.

Известно, что люди говорят различным «стилем» в зависимости от условий в окружающей среде, темы, межличностных отношений и т. д. Среди прочих параметров различные стили характеризуются и разной длительностью высказываний. Например, по данным [49], фраза, произнесенная в разговорном стиле, в 1,5 раза короче фразы, произнесенной в так называемом полном стиле, при котором реализуется наиболее четкая артикуляция и достигается наибольшая разборчивость. Это свидетельствует о действии критерия экономии оперативной памяти, т. е. о стремлении к сокращению длительности высказывания в обычных условиях. При ускоренном темпе речи длительность высказывания еще более сокращается по сравнению с разговорным стилем. Следовательно, в зависимости от конечной цели речевого общения, система управления артикуляцией может изменять скорость развертки моторных команд во времени.

Наряду с укорочением средней длительности сегментов при увеличении их числа в высказывании, наблюдается и изменение длительности сегментов в зависимости от их положения в высказывании. Начальные сегменты — наиболее короткие, а конечные сегменты — длиннее начальных. В работах [101, 133, 144, 160] указывается, что длительность сегментов зависит, главным образом, от числа оставшихся сегментов до конца высказывания. Удлинение конечных сегментов во фразе отмечается во многих языках, в том числе и в русском [19], хотя в еврейском языке такого удлинения не наблюдается [75].

В эксперименте 2 со звукосочетаниями типа БАБАБА и различными гласными в ударной позиции средняя по 20 реализациям длительность звонкой смычки у первого /Б/ составляла 100 мс, а длительность смычки у второго /Б/ — около 165 мс для нормального темпа и соответственно 50 мс и 80 мс — для быстрого темпа. Последний безударный гласный /А/ имел примерно ту же длительность, что и ударный гласный звук — около 170 мс в нормальном темпе, и около 130 мс — в ускоренном темпе. В то же время длительность первого безударного гласного /А/ составляла около 120 мс для нормального темпа, и около 70 мс — для ускоренного. В эксперименте последний безударный гласный /А/ в слове «снова» длиннее безударного /А/ в предшествующем слоге в 2 раза для диктора А. К., в 1,4 раза — для диктора В. С. и в 1,5 раза — для диктора И. О. Даже ударный гласный

/О/ в слове «снова» подвергается удлинению по сравнению с этим же ударным гласным в предшествующем слоге: в 1,4 раза для диктора И. О и в 1,14 раза для диктора В. С.

В английском языке последний гласный удлиняется примерно в 2 раза [133], фрикативные переднеязычные /С/ и /З/ — примерно на 130%, а /В/ и /Ф/ — на 85% [89]. Слова в конце фраз в английском языке удлиняются на 108% [89].

Относительно причин удлинения конечных сегментов в [88] было высказано три предположения: 1) в это время планируется следующая фраза, 2) скорость выдачи команд из буфера обратно пропорциональна числу находящихся в нем сегментов, 3) это признак конца фразы для слушателя. В дополнение к этому, в [132] обращается внимание на то, что, возможно, удлинение конечного сегмента необходимо для формирования контура основного тона, сигнализирующего либо о завершении, либо о продолжении фразы. Таким образом, по существу предлагается два механизма удлинения: один связан с преднамеренным удлинением конечных сегментов для обеспечения маркировки типа фразы и ее границы, тогда как другой механизм отражает чисто внутренние процессы планирования и реализации высказывания системой управления.

Вполне вероятно, что эти механизмы действуют одновременно, и изменения длительности сегментов, связанные с ограничениями оперативной памяти, используются для передачи информации о типе и границах фраз. Как следует из имеющихся данных для относительно коротких высказываний, удлинение конечных сегментов является лишь частным случаем более общего закона удлинения сегментов по мере приближения к концу высказывания. В звукосочетаниях типа БАПАБА и даже слогах АБА, АДА и др. последний гласный звук имеет среднюю длительность около 170 мс, что превышает длительность ударного гласного в слогах ГСГ, произнесенных в слитных фразах в экспериментах 3 и 4. Высказанное в [88] предположение о переменной скорости выдачи моторных команд из оперативной памяти согласуется с критерием экономии стоимости памяти. В самом деле, длительность начальных сегментов высказывания будет близкой к средней длительности, полученной после сжатия всего высказывания. Однако по мере реализации моторных команд объем занятой оперативной памяти уменьшается, и сжатие может быть уже не столь большим, т. е. сегменты удлиняются. Тогда и удлинение конечного сегмента можно было бы считать не исключительным фактом, а просто эффектом освобождения памяти от высказывания. Возможно, что замедление скорости вывода команд (или увеличение интервалов между отсчетами команд) связано уже с другим критерием, устанавливающим стоимость частоты отсчетов в генераторе развертки.

Таким образом, последний сегмент высказывания оказывается в условиях как бы изолированного произнесения, поскольку

за ним ничего не следует. Поэтому можно было бы ожидать, что последний сегмент некоторого фонетического качества окажется одинаковой длительности у любого высказывания. Действительно, длительность последнего гласного /А/ в изолированных слогах ГСГ и в звукосочетаниях СГСГСГ примерно одинакова для диктора В. С. Экстраполируя прямую на рис. 4.4 на высказывание, содержащее лишь один сегмент, получим длительность, примерно равную 150 мс, что достаточно близко к значению 170 мс для конечного /А/ в слогах ГСГ и звукосочетаниях ГСГСГС. Правда, в эксперименте 4 последний гласный в слове «снова» у диктора В. С. имеет среднюю длительность около 100 мс, т. е. значительно короче. Обсуждая это явление, необходимо принять во внимание то, что к концу высказывания в оперативной памяти может начаться планирование следующего высказывания. Кроме того, степень удлинения конечных сегментов должна зависеть и от относительной важности критерия стоимости скорости развертки команд и от изменения этой стоимости как функции от числа оставшихся в оперативной памяти сегментов. Поэтому и необязательно возвращение к фиксированной длительности последнего сегмента.

Суммируя изложенное в этом разделе, сформулируем правило планирования длительности высказывания и его элементов. Длительность высказывания составляется как сумма длительностей сегментов моторной программы, уменьшенной на некоторую величину, так что все сегменты сжимаются во времени. По мере реализации моторной программы и освобождения оперативной памяти длительность сегментов увеличивается, в результате чего последние сегменты высказывания оказываются (при прочих равных условиях) длиннее начальных. В высказываниях, содержащих большое число артикуляторных сегментов, также наблюдается увеличение средней длительности сегментов по мере удаления от начала высказывания. Однако скорость нарастания длительности меньше, чем у коротких высказываний, тогда как длительность конечных сегментов заметно больше (см. рис. 4.5). Это свидетельствует в пользу предположения о преднамеренном увеличении длительности конечных сегментов с целью маркировки конца фразы. Ясно, однако, что конкретный алгоритм формирования длительности сегментов зависит как от индивидуальных особенностей каждого диктора, так и от самих моторных программ, определяющих фонетическое качество звуков и их коартикуляционное взаимодействие. В последующих разделах будет описано влияние на длительность гласных звуков степени открытости и ударной позиции, вида источника возбуждения, смягчения и места артикуляции согласных звуков, а также роль темпа артикуляции и индивидуальности тактик системы управления артикуляцией.

Ударные и безударные гласные. Ударные гласные в русском языке отличаются от безударных главным образом длительностью

Рис. 4.5. Длительность сегментов во фразе «Вроде более менее все нормально». Измерения по сонограмме

и степенью точности артикуляции. Частота основного тона на ударных гласных может и падать, и повышаться. Интенсивность ударных гласных может быть любой. Поэтому из всех просодических параметров только длительность играет различительную роль между ударными и безударными гласными. Поскольку нет принципиальной разницы в артикуляции гласного в ударной и безударной позиции, то следует признать, что различие в их длительности не определяется какими-либо врожденными свойствами процессов управления, а создается системой управления преднамеренно в качестве информативного признака. Этим признаком, однако, является не абсолютное значение длительности гласного, а некоторая относительная величина.

Из эксперимента 1 с изолированными слогами типа ГСГ получены распределения длительности ударного и безударного гласного /А/ (рис. 4.6) и гласного /О/. Средняя длительность /А/ и /О/ в ударных и безударных позициях оказалась

Рис. 4.6. Распределение длительности безударного (*1*) и ударного (*2*) /А/ в изолированных слогах ГСГ

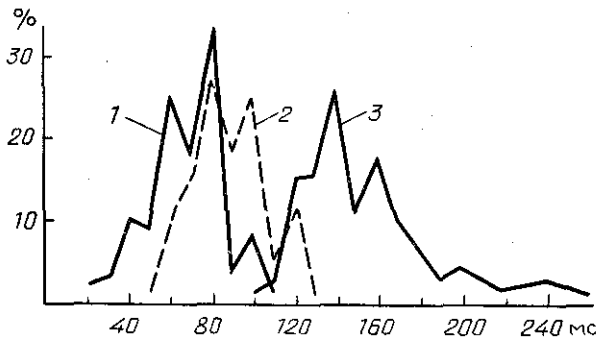


Рис. 4.7. Распределение длительности ударного гласного /А/: *1* — заударного, *2* — предударного, *3* — ударного

практически одинаковой: 245 мс для /А/ и 250 мс для /О/ в ударной позиции и, соответственно, 175 мс и 190 мс в безударной позиции. Распределение длительности ударного, предударного и заударного /А/ в эксперименте 2 показано на рис. 4.7. В этом случае средняя длительность ударного /А/ равна 140 мс, предударного — 90 мс, заударного — 70 мс. Таким образом, видим, что средняя длительность заударного гласного /А/ в эксперименте 1 близка к средней длительности ударного гласного в эксперименте 2, хотя в каждом из этих экспериментов распределения ударного и безударного звука разделены достаточно хорошо.

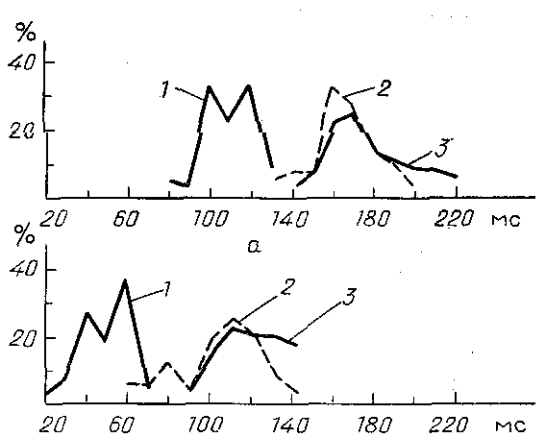


Рис. 4.8. Распределение длительности предударного гласного /А/ (*1*), ударных гласных /А, О, У, И/ (*2*) и заударного /А/ (*3*) в звукосочетаниях типа БАБОБА, *а* — нормальный темп, *б* — быстрый темп

Кроме того, в эксперименте 3 в звукосочетаниях БАБАБА средняя длительность предударного и ударного гласного была, соответственно, 110 мс и 170 мс в нормальном темпе артикуляции, и 80 мс и 110 мс в ускоренном темпе, т. е. длительность предударного гласного в нормальном темпе близка к длительности ударного гласного в быстром темпе. На рис. 4.8 показаны распределения длительности гласных в эксперименте 3 для нормального и ускоренного темпа. Хорошо видно удлинение конечного гласного, распределение длительности которого полностью перекрывается с распределением ударного гласного.

Следовательно, различение ударной или безударной позиции гласного должно производиться по каким-то относительным величинам, например, по отношению их длительностей. Относительные длительности ударного и безударного /А/ приведены в табл. 4.2.

Таблица 4.2. Отношение длительности ударного и безударного /А/, %

Эксперимент 1	Эксперимент 2		Эксперимент 3	
заударный	предударный	заударный	предударный	
			норма	быстро
139	166	214	154	138

Таким образом, ударный /А/ длиннее безударного примерно на 40—110%. Эти значения близки к данным [94] для разных языков: в английском это отношение равно 60%, в немецком — 44%, в испанском — 30%, в шведском [144] — 33—53%. В еврейском языке это отношение меньше — оно составляет 15—20% [75]. Отношения длительностей между ударными и безударными гласными более наглядно представлены на рис. 4.9, где по оси абсцисс отложена длительность ударного, а по оси ординат — длительность безударного гласного в слогах типа ГСГ. В левой нижней части этого рисунка находится кластер, соответствующий результатам эксперимента 2, а в правой верхней — результатам эксперимента 3. Как видно, существует примерно линейная (в среднем) зависимость между длительностью безударного и ударного гласного, однако случайные отклонения от этой зависимости довольно велики.

Помимо различий в длительности гласных, связанных с ударной или безударной позицией, имеются меньшие различия, определяемые степенью открытости гласных. Наиболее заметно эти различия проявляются при сравнении двух гласных, имеющих максимальную и минимальную открытость — /А/ и /И/. В эксперименте 5 со слогами $С_1ГС_2$, где в качестве

гласных использовались /А/ и /И/, средняя длительность ударного /А/ равнялась 135 мс, а ударного /И/ — 102 мс, т. е. разница между ударными /А/ и /И/ примерно такая же, как и разница между ударными и безударными гласными /А/.

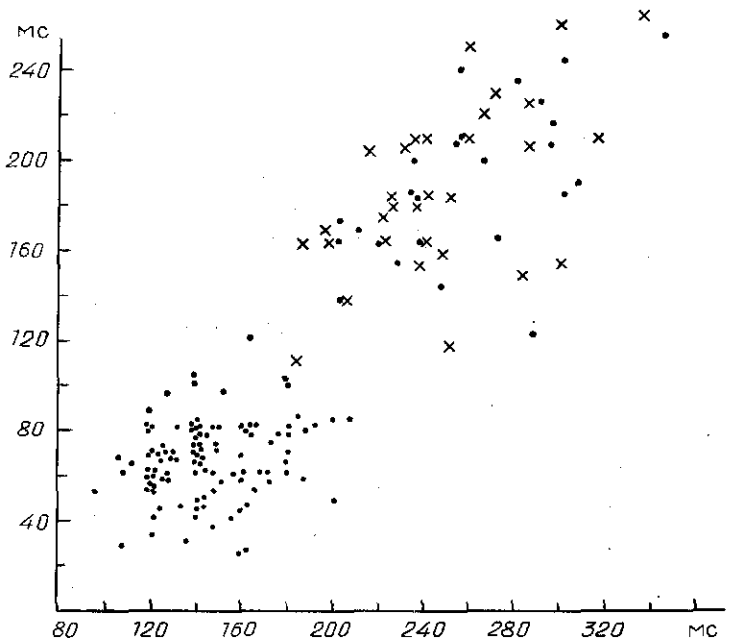


Рис. 4.9. Длительность безударного гласного как функция длительности ударного гласного в слогах ГСГ, ··· — гласный /А/, × × × — гласный /О/

Поэтому для правильного определения ударной или безударной позиции того или иного гласного нужно знать его «собственную» длительность в обеих позициях. Различие по степени открытости гласных отмечается и для других языков. Например, в шведском языке открытые гласные на 40 мс длиннее закрытых [144].

Собственные длительности гласных исследовались в эксперименте 4 со слогами $Г_1СГ_2$, где в качестве согласных использовались только переднеязычные /С, З, Т, Д/, и были реализованы все возможные сочетания гласных. Результаты показаны в табл. 4.3, где гласные упорядочены по степени подъема нижней челюсти, определяющей открытость гласного.

Как видно, длительность безударных гласных примерно постоянна и не зависит от степени открытости.

В [161] предлагается формула для расчета длительности гласного как функции либо от степени подъема h нижней

Таблица 4.3. Собственные длительности гласных, мс

Гласный	А	О	Э	У	Ы	И
Ударный	133	119	120	111	103	95
Безударный	71	77	82	76	75	76

челюсти:

$$T_г = -19h + 31{,}1k + 95{,}3,$$

либо по значениям частот первой и второй формант F_1 и F_2:

$$T_г = 0{,}95k + 0{,}08F_1 - 0{,}01F_2 - 10{,}61,$$

где k — коэффициент, зависящий от того, является ли гласный длинным или коротким, назальным или ротовым (имеется в виду французский язык). То, что во второй формуле вместе с частотой первой форманты F_1, обычно связываемой со степенью подъема нижней челюсти, присутствует и частота второй форманты F_2, говорит о влиянии и места наибольшего сужения в речевом тракте, т. е. зависимости длительности гласного не только от открытости, но и от того, является ли он передним или задним.

Линейная зависимость длительности гласного от степени опускания нижней челюсти является весьма грубой аппроксимацией. *Движения нижней челюсти описываются* дифференциальным уравнением второго порядка с переменными параметрами, причем скорость ее опускания в 1,2—1,6 раза больше скорости подъема (для одного из дикторов, принимавшего участие в экспериментах [59]). В результате этого скорость перехода от открытых гласных к закрытым меньше, а длительность перехода, соответственно, больше, чем у перехода от закрытых к открытым гласным. В частности, в эксперименте 4 средняя длительность слогов АСоглИ равна 297 мс, а средняя длительность слогов МСоглА равна 260 мс. Большая скорость переходов ИА по сравнению с АИ отмечается и в [121]. Кроме того, на время достижения позиции нижней челюсти, соответствующей артикуляции данного гласного, влияет и исходная позиция нижней челюсти. Поэтому длительность звука в контексте является следствием решения уравнения

$$h'' + 2gh' + \omega^2 h = F(t), \quad h(0) = h_0,$$

где h_0 — *начальное положение нижней челюсти*, $F_\infty/(\omega^2 + g^2)$ — конечное (целевое) положение, причем g, ω и $F(t)$ зависят от *подъема или опускания челюсти. Величины опускания нижней* челюсти для диктора В. С. примерно равны 0,88 см для открытых гласных /А, О, Э/, и 0,39 см для закрытых гласных /И, У, Ы/ [59].

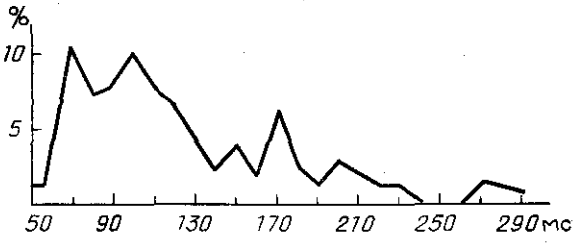


Рис. 4.10. Распределение длительности звонкой смычки заударных взрывных в изолированных слогах ГСГ

Звонкие и глухие согласные. Во многих языках длительность глухого согласного больше длительности звонкого согласного. Причины этого явления не вполне ясны. Артикуляция глухих согласных сопровождается разведением голосовых складок после начала смычки и их сведением в позицию фонации после взрыва смычки, тогда как у звонких согласных колебания голосовых складок на интервале смычки обычно не прекращаются. В начале глухого согласного (смычки или шумного интервала) в течение нескольких десятков миллисекунд обычно наблюдаются колебания голосовых складок с постепенным затуханием, а в конце глухого фрикативного звука имеется пауза, в течение которой сближаются голосовые складки. Возможно, что именно на эти интервалы и удлиняются глухие согласные для того, чтобы в системе слухового восприятия успели сформироваться соответствующие признаки звонкости/глухости.

Длительность согласных звуков, так же как и гласных, зависит от длительности высказывания и от позиции в высказывании. На рис. 4.10 показано распределение длительности звонкой смычки заударных согласных /Б, Д, Г, М, Н/ в слогах ГСГ с гласным /А/ (эксперимент 1). Средняя длительность каждого взрывного согласного близка к 120 мс, и только длительность /В/ заметно меньше — около 85 мс. На рис. 4.11 показаны распределения длительности звонкой и глухой смычек предударных согласных /Б/ и /П/ в звукосочетаниях типа БАБАБА в нормальном и ускоренном темпе. На рис. 4.12

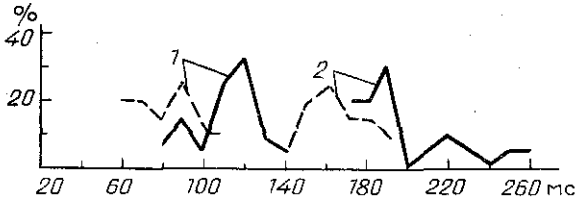


Рис. 4.11. Длительность звонкой (- - -) и глухой (——) смычек предударного согласного в быстром (*1*) и нормальном (*2*) темпе

показаны распределения длительности глухой и звонкой смычек в предударном и заударном положении (эксперимент 2), а на рис. 4.13 показаны распределения длительности фрикативных звуков (эксперименты 4, 5).

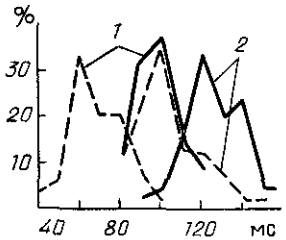


Рис. 4.12. Длительность заударной (*1*) и предударной (*2*) смычек. - - - — звонкие, ——— — глухие

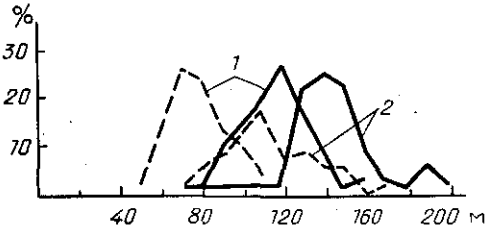


Рис. 4.13. Распределение длительности звонких (- - -) и глухих (——) фрикативных, *1* — заударные, *2* — предударные

Из сравнения этих распределений прежде всего видно, что имеется минимальное (40 мс) и максимальное (290 мс) значения длительности звонкой смычки, причем весь этот диапазон реализуется только в изолированных слогах ГСГ. О существовании минимального значения длительности смычки согласных указывалось в [49], а в [66] было показано, что при удлинении согласного /С/ более, чем на 200 мс, этот согласный начинает восприниматься как удвоенный.

Абсолютная разница между средней длительностью глухой и звонкой смычки меняется в зависимости от положения относительно ударного гласного и темпа речи, составляя 20—50 мс, или 8—26% в относительных единицах (разность средних значений, деленная на их сумму). Из рис. 4.12 видно, что в слитной речи по длительности смычки лучше всего различаются заударные звонкие и предударные глухие взрывные — разница между модами распределений составляет 60 мс или 33% от суммы мод. Примерно такое же соотношение наблюдается и между глухими и звонкими фрикативными (рис. 4.13). Но в других случаях эта разница не столь велика и длительность одного лишь согласного является слабым признаком звонкости/глухости, хотя вместе с другими признаками, например, признаком фонации, она может быть сильным признаком [25].

Наряду с собственной длительностью смычки, ее звонкость/глухость влияет и на длительность предшествующего гласного. В английском языке, например, это влияние весьма велико — перед звонким согласным гласный удлиняется на 50—100 мс [133]. В шведском языке гласные удлиняются перед звонкими на 20—30 мс [144], а в арабском языке вообще не наблюдается такого удлинения [157]. В английском

языке, где конечные звонкие согласные не оглушаются независимо от наличия или отсутствия фонации, конечные фрикативные воспринимаются как глухой или звонкий только на основании длительности предшествующего гласного [96].

В русском языке гласный /А/ длиннее перед звонкими взрывными примерно на 20 мс независимо от темпа и ударной или безударной позиции (рис. 4.3 и 4.4). На такую же, в среднем, величину отличаются гласные /А, О, Э, У, Ы, И/ и перед звонким /З/ по сравнению с глухим /С/ (эксперимент 4). Учитывая эффект компенсации в сочетаниях ГС, можно предложить отношение $\delta_T = (T_г - T_с)/(T_г + T_с)$ как величину, характеризующую одновременно и ударность/безударность и звонкость/глухость последующего согласного (табл. 4.4, составленная по результатам эксперимента 2).

Таблица 4.4. Относительная длительность гласного и согласного

Тип звука	Ударный	Безударный
Звонкий	0,27	0,0
Глухой	0,13	−0,27

Как видно, ударному гласному соответствуют положительные значения этого отношения, а безударному — отрицательные, хотя имеются различия, связанные со звонкостью/глухостью последующего согласного. Здесь знаменатель $T_г + T_с$ играет роль нормировки, но эта нормировка хороша лишь для небольшой вариации темпа артикуляции. При очень быстром произнесении соотношения между длительностями гласных и согласных изменяются, поэтому решение о звонкости/глухости лучше производить на плоскости (δ_T, $T_г + T_с$). В иностранной литературе в качестве просодического признака звонкости/глухости используют отношение длительностей гласного и следующего за ним согласного звука — чем больше это отношение, тем больше вероятность наличия голосового возбуждения [96]. Отношения $T_г/T_с$ для тех же данных, что и в таблице 4.4, представлены в таблице 4.5.

Таблица 4.5. Отношение длительности гласного и согласного

Тип звука	Ударный	Безударный
Звонкий	1,75	1,0
Глухой	1,33	0,57

По сравнению с δ_T отношение $T_г/T_с$ представляется менее чувствительным к изменению признака звонкости/глухости.

Фрикативные согласные. Фрикативные согласные /С, Ш, Х, Ф, З, Ж, В/ несколько длиннее взрывных согласных с тем же местом артикуляции. Это отмечалось и в [19]. Из сравнения распределений длительности согласных на рис. 4.12 и 4.13 видно, что фрикативные на 10—20 мс длиннее взрывных, причем предударные фрикативные примерно на 20 мс длиннее взрывных, а заударные — на 10 мс. При этом различие в длительности между глухими и звонкими фрикативными такое же, как между глухими и звонкими взрывными, как в предударной, так и в заударной позициях.

Гласные перед звонкими фрикативными удлиняются так же, как и перед звонкими взрывными, причем безударные (предударные) гласные в среднем увеличиваются примерно на 20 мс, а ударные — на 10 мс.

Заударные фрикативные на 30—50 мс короче предударных, в чем проявляется такой же эффект компенсации длительности в сочетании ГС, как и для взрывных. Поскольку фрикативные немного длиннее взрывных, можно было бы ожидать, что, согласно принципу компенсации, гласные перед фрикативными будут короче, чем перед взрывными. Однако в действительности имеется обратное явление — гласные перед фрикативными длиннее, чем перед взрывными. Величина этого удлинения колеблется от 7 мс до 24 мс в зависимости от ударной или безударной позиции гласного и степени его открытости. Наибольшие изменения длительности в ударной позиции наблюдаются для гласного /И/ — 20—24 мс, а в безударной позиции — для гласного /А/ перед звонкими фрикативными — 13 мс. Следовательно, эффект компенсации длительностей гласного и согласного в сочетаниях ГС не распространяется на изменение длительности согласного при изменении источника возбуждения. Имеющийся экспериментальный материал слишком ограничен для того, чтобы ответить на вопрос о существовании эффекта компенсации в сочетаниях ГС с фрикативными.

Несмотря на относительную малость увеличения длительности фрикативных по сравнению с длительностью взрывных, вместе с увеличенной длительностью предшествующего гласного суммарная длительность сочетаний ГС с фрикативными может оказаться уже заметно больше, чем длительность сочетаний ГС со взрывными, играя, таким образом, роль вторичного признака фрикативности. Действительно, средняя длительность сочетаний с ударным /И/ и последующими звонкими фрикативными на 26 мс больше длительности сочетаний со звонкими взрывными, что составляет 15% от длительности сочетания со взрывными. Эта величина сравнительно мала, но она превышает порог восприятия изменения длительности и, таким образом, может участвовать в процессах принятия решения о способе возбуждения. Вместе с тем, надо отметить, что для использования таких малых различий

в длительности важно иметь информацию об общем темпе артикуляции, месте слога в слове и положении слова во фразе.

Твердые и мягкие согласные. Артикуляция мягких согласных осуществляется с подъемом средней части языка, т. е. в дополнение к артикуляционным движениям, соответствующим тому или иному твердому согласному, выполняется еще одно движение. Поэтому можно ожидать увеличения длительности мягких согласных по сравнению с твердыми. Действительно, распределение длительностей предударных твердых и мягких согласных, включая звонкие и глухие, взрывные и фрикативные (по эксперименту 5), демонстрирует некоторое увеличение длительности мягких согласных — в среднем на 10 мс (рис. 4.14). Твердые и мягкие переднеязычные согласные /Т/ отличаются на 13 мс, /С/ — на 8 мс, /З/ — на 20 мс. В то же время губные твердые и мягкие почти не отличаются по длительности. Это можно объяснить тем, что при артикуляции мягких губных согласных подъем средней части языка может происходить одновременно с движениями губ, тогда как движения кончика и середины языка сильно взаимодействуют.

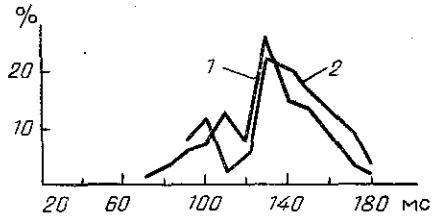


Рис. 4.14. Распределение длительности твердых (*1*) и мягких (*2*) предударных согласных

Время возникновения фонации. Глухие согласные — взрывные и фрикативные, образуются при разведенных голосовых складках. По окончании согласного складки сводятся в положение, при котором начинаются их автоколебания. Время, в течение которого происходит переходный процесс голосовых складок из одного положения в другое, характеризуется отсутствием сигнала, т. е. паузой. Длительность этой паузы зависит как от скорости движения голосовых складок, так и от фазы начала их движения относительно движений артикуляторных органов.

Время возникновения фонации (в английской литературе — *voice onset time*) для взрывных согласных неодинаково для разных языков. В английском языке интервал между взрывом смычки и первым импульсом голосового источника служит наиболее надежным различительным признаком звонкости/глухости [96, 178]. Длительность этого интервала для глухих взрывных составляет 30—90 мс, а для звонких — менее 30 мс. В русском языке взрыв звонких смычных проявляется весьма редко, поэтому здесь противопоставление «звонкий — глухой» скорее происходит не по времени возникновения фонации, а по самому факту наличия или отсутствия взрыва. В эксперименте 3 со взрывными /П, Т, К/ измерялось время возникновения фонации предударной и заударной позиций.

Эти данные представлены на рис. 4.15. Среднее время возникновения фонации для /П/ равно 14 мс, для /Т/—17 мс и не зависит от положения относительно ударного гласного. Для /К/ оно отличается в полтора раза и равно 28 мс для

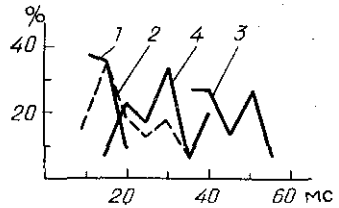


Рис. 4.15. Время возникновения фонации после взрыва: *1*—/П/, *2*—/Т/, *3*—предударный /К/, *4*—заударный /К/

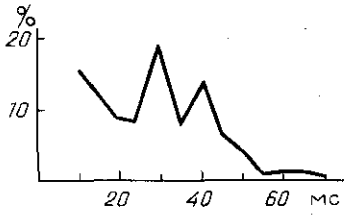


Рис. 4.16. Длительность паузы между /С/ и /Н/ в слове «снова»

заударной позиции и 43 мс—для предударной, причем часто наблюдаются повторные импульсы через 4—24 мс. Из рис. 4.14 видно, что место артикуляции влияет на распределение времени возникновения фонации. Поскольку измерения производились только в симметричном окружении с гласным А, осталась неизвестной роль смены гласных.

Имеются сведения о пропорциональной связи между временем возникновения фонации и длительностью последующего гласного [178]. Там же сообщается, что это время наибольшее в случае, если за данным взрывным следует группа глухих согласных, и это время убывает, если за ним следуют назальные, напряженные гласные и ненапряженные гласные (в порядке убывания).

Пауза между глухим фрикативным и последующим звонким звуком вызывается тем, что по мере уменьшения площади голосовой щели при сближении голосовых складок ее сопротивление возрастает и в определенный момент скорость воздушного потока падает настолько, что турбулентное шумообразование прекращается, в то время как фонация еще не началась. Распределение длительности паузы между фрикативным /С/ и назальным /Н/ в слове «снова» во фразе «вот ГСГ снова» (эксперимент 3) показано на рис. 4.16. Как видно, это распределение захватывает длительности вплоть до 70 мс со средним значением, близким к 30 мс. Из других экспериментов установлено, что между глухим фрикативным и последующим гласным всегда существует хотя бы короткая пауза. Таким образом, хотя эта пауза появляется в результате чисто аэродинамических процессов, она может играть перцептивную роль в идентификации глухих фрикативных.

Необходимо отметить, что измерения времени между концом смычки и началом фонации с помощью датчиков касания и ларингофона в [49] показывают бóльшие значения этого

времени как для /П/ (около 40 мс), так и для /Т/ (20—50 мс с последующими открытыми гласными /А/ и 83—111 мс с последующими /И/). Поскольку на величину этого времени влияет ротовое давление, то быстрое его падение после взрыва приводит к уменьшению времени начала фонации. Известно, что у многих дикторов при артикуляции гласного /А/ небная занавеска слегка опущена и приоткрывает проход в носовую полость. Поэтому в зависимости от индивидуальной степени назализации последующего гласного время возникновения фонации может быть больше или меньше. В частности, при артикуляции мягких согласных, характеризующихся наиболее высоким положением небной занавески, время возникновения фонации увеличивается вдвое по сравнению с твердыми (с 24—28 мс до 52—83 мс для /Т/ и /Т'/ [34]). Если за глухим фрикативным следует назальный звук, то небная занавеска опускается еще во время интервала турбулентности. При этом основной поток воздуха идет через носовую полость, скорость потока через ротовую полость падает и турбулентность прекращается еще до начала сведения голосовых складок в позицию для фонации.

Темп артикуляции. Смена темпа артикуляции приводит к разнообразным изменениям в длительности звуков речи. Уже давно известно, что длительность гласных и согласных при этом меняется с разными коэффициентами. Например, в [49] сообщается, что при увеличении длительности фразы происходит уменьшение доли времени, занятой согласными относительно доли времени, занятой гласными. В [108] приводятся следующие коэффициенты укорочения звуков в слогах $С_1ГС_2$ при ускоренном темпе: для первого согласного — 0,9—0,95, для второго согласного — 0,81—0,94, для гласного — 0,74—0,82. Из данных обеих работ следует, что при ускорении темпа гласные укорачиваются больше, чем согласные. Результаты наших экспериментов, однако, показывают, что при изменении темпа происходят весьма сложные явления, связанные иногда с полной реорганизацией активности мышц, управляющих артикуляцией [59].

Планирование высказывания в заданном темпе охватывает не только моторные команды, но и паузы между высказываниями. Например, в эксперименте 3 в нормальном темпе при средней длительности звукосочетаний 830 мс, средняя длительность пауз между ними была равна 670 мс, а в ускоренном темпе в 1,6 раза паузы сократились в 1,4 раза. При ускорении фраз длиной 1,47—2,02 с в среднем в 1,45 раза паузы длиной 0,62—0,87 с сократились в 1,38 раза. Очевидно, длительность паузы между высказываниями зависит от длительности самих высказываний, поскольку для записи моторных команд в оперативную память требуется время, пропорциональное их длительности. Иными словами, длительность паузы перед высказыванием определяется длительностью этого

высказывания, хотя ее могут изменять процессы дыхания, необходимость обдумывания следующей фразы, степень важности сообщения и другие факторы.

При ускорении темпа артикуляции степень укорочения сегментов зависит от их положения в высказывании, как это

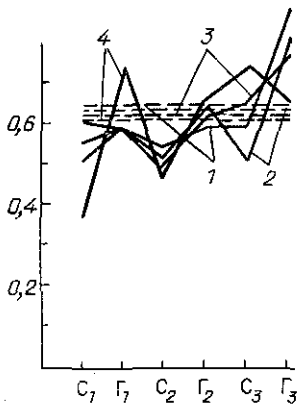


Рис. 4.17. Отношение длительности сегментов в быстром темпе к длительности сегментов в нормальном темпе: *1* — БАБИБА, *2* — БАБОБА, *3* — БАБАБА, *4* — БАБУБА; штриховая линия — средний темп

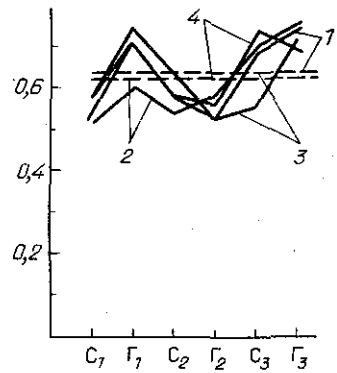


Рис. 4.18. Отношение длительности сегментов в быстром темпе к длительности сегментов в нормальном темпе: *1* — БАПИБА, *2* — БАПАБА, *3* — БАПУБА, *4* — БАПОБА; штриховая линия — средний темп

следует из рис. 4.17 и 4.18, где показано среднее отношение по 5 реализациям длительности высказывания в быстром и нормальном темпе и отношение длительности звуков в звукосочетаниях БАБА́БА, БАБО́БА, БАБУ́БА, БАБИ́БА, (рис. 4.17) и БАПА́БА, БАПО́БА, БАПУ́БА, БАПИ́БА (рис. 4.18). Во всех случаях первый согласный укорачивается больше, а последний согласный — меньше, чем среднее укорочение. Начальные слоги также укорачиваются больше, чем последние. Поэтому при ускорении темпа степень укорочения падает по мере реализации высказывания. В этом наблюдается аналогия с вариацией длительности сегментов при заданном темпе в зависимости от положения в высказывании, которая обсуждалась ранее в связи с эффектом укорочения сегментов при увеличении их числа. В результате относительное удлинение (точнее, неукорочение) последнего сегмента в высказывании увеличивается с ускорением темпа. Таким образом, похоже, что в основе укорочения сегментов при ускорении и при увеличении числа сегментов в высказывании лежит один и тот же механизм переменной скорости развертки моторных команд, а роль темпа сводится лишь к дополнительному сжатию всего высказывания.

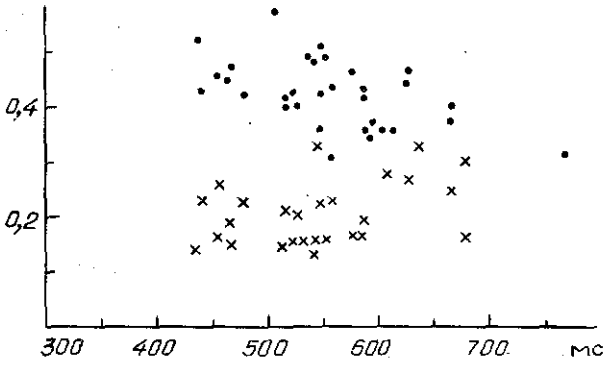


Рис. 4.19. Отношение длительности ударного /А/ (точки) и заударного /Б/ (крестики) к длительности слога АБА как функция длительности слога

Что касается относительного ускорения артикуляции гласных и согласных, то результаты измерений не подтверждают мнения о преимущественном укорочении гласных при ускорении темпа. Для первых трех звуков — БАП или БАБ, гласный А всегда укорачивается меньше, чем любой из согласных, т. е. относительная доля начальных согласных при ускорении темпа уменьшается. Наряду с этим, имеются отличия в поведении слогов АБА и АПА: глухой взрывной /П/ укорачивается хотя и в большей степени, чем среднее укорочение, но все же меньше, чем звонкий взрывной /Б/. Предударный гласный перед /Б/ обычно укорачивается больше, чем перед /П/, а ударный /А/ после /Б/ укорачивается меньше, чем после /П/. Механизм влияния звонкости / глухости согласного на укорочение комплекса ГСГ не ясен. Вариации длительности сочетаний ГС и СГ значительно меньше вариаций входящих в них гласных и согласных звуков, и обычно степень укорочения этих сочетаний близка к общему укорочению.

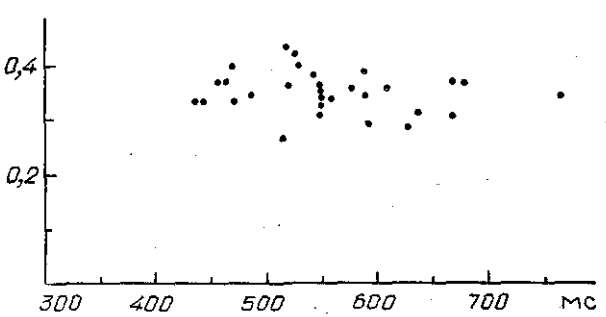


Рис. 4.20. Отношение длительности заударного /А/ к длительности слога АБА как функция длительности слога

При рассмотрении изменений длительности согласных при вариации темпа в [49] указывалось, что существуют минимальные и максимальные значения длительности смычки. Минимальное значение определяется наибольшей скоростью изменения направления движений артикуляторных органов. Максимальное значение связано с восприятием удвоенного согласного или конца слова. Длительность гласных не ограничена ни сверху, ни снизу. Поэтому в [49] принимается, что изменения темпа происходят, в основном, за счет изменения длительности гласных. В наших экспериментах длительность смычки первого согласного /Б/ в отдельных случаях укорачивалась до 30 мс и даже 20 мс, а длительность смычки второго /Б/ не опускалась ниже 60 мс, что, в общем, совпадает с данными [49]. Однако в наших экспериментах согласные все же укорачиваются больше, чем гласные при ускорении темпа. Возможно, это является следствием особой тактики управления артикуляцией у данного диктора, но возможно также, что здесь играют роль интересы системы восприятия — при укорочении гласного в большей степени затрудняется определение его фонетического качества, чем при укорочении смычки согласного.

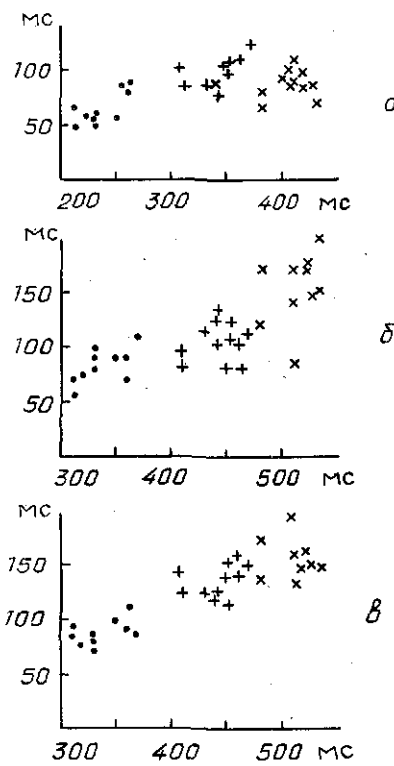


Рис. 4.21. Относительные длительности: *а* — предударного гласного, *б* — согласного, *в* — ударного гласного в слоге типа ГСГ как функция длительности слога (··· — диктор В. С., +++ — диктор И. О., ××× — диктор А. К.)

Особый интерес представляет различие темпа артикуляции у разных дикторов, хотя при этом надо помнить о том, что для каждого диктора его собственный темп является нормой. В эксперименте 1 скорость произнесения слогов АБА в массиве из 30 дикторов различалась почти в два раза — от 440 мс до 770 мс. Относительная длительность первого гласного и согласного в этом эксперименте показана на рис. 4.19, а последнего гласного — на рис. 4.20. Прежде всего обращает на себя внимание постоянство относительной длительности последнего гласного. Это означает, что в среднем длительность последнего гласного изменяется пропорционально длительности слога, причем разброс относительно невелик. Из рис. 4.18 следует,

что у дикторов, говорящих в разном темпе, имеются разные относительные длительности первого гласного и согласного. С увеличением длительности слога относительная длительность первого гласного падает, тогда как доля согласного остается примерно постоянной в диапазоне длительности слога 440—600 мс, а затем возрастает с увеличением длительности слога. Таким образом, для большинства дикторов длительность согласного и последнего гласного изменяется пропорционально общей длительности слога с одним и тем же коэффициентом, примерно равным 1, а длительность первого гласного изменяется с коэффициентом, меньшим 1.

В эксперименте 4 принимали участие три диктора с большой разницей в темпе артикуляции. Временные соотношения в слоге АСоглА для этих дикторов оказались такими же, как и в эксперименте 1 (рис. 4.21). Как видно, длительность ударного гласного изменяется пропорционально длительности слога, т. е. его относительная длительность остается постоянной. Длительность согласного мало меняется в диапазоне общей длительности 200—350 мс, а затем быстро растет. Хорошо виден эффект компенсации в сочетании ГС, который приводит к постоянному отношению длительности ГС к длительности всего слога у каждого диктора (см. табл. 4.6).

Таблица 4.6. Относительные длительности в слоге $Г_1СГ_2$

Диктор	$Г_1$	С	$Г_2$	$Г_1С$	Средняя длит. $Г_1СГ_2$, мс
В. С.	0,27	0,35	0,38	0,62	237
И. О.	0,31	0,3	0,4	0,61	340
А. К.	0,23	0,38	0,39	0,61	410

Отсюда следует, что при изменении темпа артикуляции как одним диктором, так и у разных дикторов длительность сегмента ГС изменяется пропорционально общей длительности высказывания, с учетом позиции этого сегмента в высказывании.

Порождающая модель для временной структуры. При синтезе речи по тексту необходимо иметь порождающую модель для управления длительностью сегментов речевого потока. В [133, 134] было сформулировано правило для изменения длительности гласных, исходя из принципа «несжимаемости». Предполагается, что в ударной позиции длительность i-го гласного не может быть меньше некоторой величины $T_{i\min}$, а влияние контекста учитывается коэффициентом k в формуле

$$T_i = k\left(T_i^{(0)} - T_{i\min}\right) + T_{i\min},$$

где $T_i^{(0)}$ — собственная длительность гласного, определяемая как длительность в ударной позиции в слоге ГСГС,

включенном в короткую фразу. В [144] эта формула была модифицирована с учетом сжатия сегментов в зависимости от их числа и положения в высказывании:

$$T_i = \alpha^a \beta^b (T_i^{(0)} - T_{i\min}) + T_{i\min},$$

где α и β — коэффициенты укорочения, a — число слогов после данного сегмента, b — число слогов перед данным сегментом. Эти формулы описывают лишь влияние соседних сегментов на длительность гласного, тогда как собственные длительности задаются таблично, а изменение длительности согласных вообще не описывается. Вместе с тем, понятие о собственной длительности звука представляется весьма важным, хотя и требует обоснования для выбора контекста с целью измерения $T_i^{(0)}$.

Исходя из предположения о минимальном укорочении (или даже отсутствии такого укорочения) последнего сегмента в высказывании, можно определить $T_i^{(0)}$ для гласных как длительность ударного гласного в изолированных слогах СГ, для фрикативных согласных — как длительность в изолированных слогах ГС, а для взрывных — как длительность в слогах ГСГ с ударением на последнем гласном. Эти слоги должны произноситься в нормальном темпе данного диктора.

В наших экспериментах было показано, что существует ряд факторов, влияющих на длительность гласных и согласных, причем эти факторы действуют независимо, т. е. здесь справедлив принцип суперпозиции. Независимость факторов отмечалась и в [178]. Поэтому попытаемся сконструировать правила, управляющие длительностью сегментов во фразе, хотя в настоящее время еще неизвестны некоторые необходимые для этого свойства временной структуры речи.

1. По фонетической последовательности считываются собственные длительности гласных и согласных и формируется вспомогательный двоичный код высказывания в виде последовательности /С/ и /Г/.

2. Если гласный — безударный, то его длительность укорачивается на 60—80%, при этом длительность последующего согласного остается неизменной.

3. Если гласный — ударный, то длительность последующего согласного укорачивается на 25 мс, и длительность следующего за согласным звуком также укорачиваается на 25 мс.

4. Если согласный — звонкий, то предшествующий гласный удлиняется на 20 мс.

5. Если согласный — щелевой, то предшествующий гласный удлиняется на 15 мс.

6. Если согласный — губной назальный, то предшествующий гласный удлиняется на 10 мс.

7. Если согласный — переднеязычный мягкий, то он удлиняется на 10 мс.

8. Вычисляется длительность $T_{\text{гс}}$ сочетания ГС, и длительность гласного $T_{\text{г}}$ подвергается небольшому (около 10%)

случайному изменению, а длительность согласного определяется как $T_с = T_{гс} - k_{гс} T_г$, где $k_{гс} < 1$. Этот коэффициент отражает неполную компенсацию изменения длительности элементов сочетания ГС.

9. Если число сегментов в высказывании меньше 15, то длительность сегмента, находящегося на j-й позиции в высказывании, вычисляется как

$$T_j = \left[T_j^{(0)} - k_п (N - j)\right] \left[1 - (1 - k_т) \frac{N - j + 1}{N}\right] \frac{T}{N},$$

где $k_п$ — коэффициент сжатия, требуемого по условиям экономии оперативной памяти, $k_т$ — коэффициент темпа ($k_т = 1$ означает отсутствие изменения темпа, $k_т < 1$ — ускорение), N — число сегментов в высказывании, T — длительность высказывания.

10. Если число сегментов в высказывании больше 15, то их средняя длительность примерно постоянна, и лишь один — два последних сегмента в 1,5—2 раза длиннее среднего уровня.

Величины изменения длительности сегментов в этих правилах соответствуют средним данным по некоторому множеству дикторов. Для придания индивидуальности временным параметрам синтетической речи эти величины должны соответствовать измерениям в речи конкретного диктора. Дикторским особенностям посвящена работа [26]. Для завершения модели временной структуры речи предстоит исследовать сочетания согласных, роль логического фразового ударения, индивидуальные особенности ускорения темпа, влияние типа фразы.

§ 4.2. Интонация

Изменение частоты основного тона является одной из наиболее важных просодических характеристик. С помощью основного тона передается информация о поле, физическом и эмоциональном состоянии диктора, степени семантической выделенности слов и фраз, маркируется тип фраз — вопросительный, восклицательный и т. д. Каждый язык обладает своей системой правил управления основным тоном, и аудиторы осознают любое нарушение этих правил. В синтезе речи адекватное управление частотой основного тона необходимо не только для обеспечения натуральности звучания, но и для правильного понимания смысла синтезированного речевого сообщения.

Частота колебаний голосовых складок F_0 зависит от их геометрических размеров, плотности тканей и напряжения мышц голосовых складок. Размеры складок и плотность тканей связаны, главным образом, с полом диктора, а также с размерами тела и физическим состоянием человека. Человек

с большим телом и физически крепкий обладает и большими голосовыми складками, частота колебаний которых низка, в импульсах голосового возбуждения проявляются высшие моды колебаний складок и голос звучит довольно резко с заметными нерегулярностями. Поэтому низкая частота основного тона обычно связывается с силой, господством, агрессивностью, уверенностью в себе, высоким общественным статусом. Высокая частота основного тона, свойственная детям, связывается с подчинением, неуверенностью, почтительностью. Предполагается, что механизмы декодирования высокого и низкого основного тона — врожденные, и проявляются также и в поведении животных [168]. Повышение F_0 — это мимикрия под младенца, демонстрация отсутствия угрозы, взывание к родительскому инстинкту.

Дикторов радио и телевидения специально обучают низкому основному тону для создания образа уверенности. Главный герой в радио и телеспектаклях обычно имеет более низкий основной тон. В определенных кругах США, например, не только мужчины, но и женщины стараются говорить вблизи нижней границы своего частотного диапазона основного тона. В восточных странах, наоборот, женщины говорят намеренно высоким, нежным, почти детским голосом. Эти различия, конечно, связаны с социальными нормами того или иного общества.

Возможно, что именно с врожденными механизмами оценки высоты частоты основного тона связано повышение F_0 в вопросительных предложениях, где как бы демонстрируется неуверенность, просьба о помощи, а также аномальное повышение F_0 в утвердительных предложениях с целью показать почтительность и уважение к собеседнику [168].

Зависимость частоты основного тона от напряжения мышц гортани приводит к тому, что все процессы в нервной системе человека так или иначе сказываются на интонации. Изменение частоты основного тона может быть непроизвольным, например, при эмоциональном напряжении, и зависит от вида и силы эмоций. Кроме того, на частоту и форму импульсов голосового источника влияет степень расширения или сужения кровеносных сосудов, расположенных вдоль голосовых складок. Известно, что определенные эмоциональные состояния могут привести к изменению диаметра кровеносных сосудов и, соответственно, к изменению характеристик голосового источника. Имеются и другие виды влияния состояния диктора на интонацию. Например, увеличение степени заинтересованности диктора в предмете разговора проявляется в повышении начальной точки $F_0^{(0)}$ контура основного тона, а также увеличении максимальных и минимальных значений F_0 [83]. Эмоциональное напряжение при докладе в большом зале повышает среднюю частоту основного тона с 129 Гц до 188 Гц, т. е. примерно на 60 Гц [36].

Уровень F_0 выше для эмоций гнева, удовольствия, радости, изумления, и ниже — для эмоций сожаления, вины, нежности, оскорбления [166]. В [27] отмечается повышение F_0 для положительных эмоций, и понижение — для отрицательных. Сердитая ссора характеризуется жестким метрическим ритмом с равномерно отстоящими основными ударениями, причем жесткая мелодическая линия F_0 прерывается внезапными подъемами [100]. Связь частоты основного тона с состоянием системы управления движениями человека проявляется в повышении F_0 и уменьшении диапазона изменений F_0 при ускорении артикуляции [205]. Отмечается ускорение темпа в вопросительных предложениях [56], что также может коррелировать с повышением F_0. Все эти явления в естественной речи в определенной степени непроизвольны, хотя имеется возможность имитировать в голосе те или иные эмоции. В синтетической речи также необходимо научиться управлять факторами, связанными с выразительностью и убедительностью речи.

На границах между гласными и согласными возникают нерегулярности следования импульсов голосового источника, вызывающиеся изменениями физических условий, в основном, перепадом давления на голосовой щели. Сразу после взрыва глухой смычки F_0 слегка повышается, а после звонкой смычки — слегка понижается [168], но и в том, и в другом случае на протяжении 75—100 мс после взрыва смычки F_0 понижена относительно среднего уровня [185]. В отличие от этих непроизвольных изменений, на конце фраз могут создаваться управляемые возмущения частоты и формы импульсов голосового источника, служащие специальными маркерами [183]. Изменение относительной активности мышц гортани может привести к изменению формы импульсов и относительной длительности интервалов с открытой и закрытой голосовой щелью. Это меняет тембр голоса, в частности, вследствие изменения отношений амплитуд формант, и может использоваться для передачи специфической информации, например, эмоций презрения, насмешки, иронии.

Еще один пример непроизвольного изменения частоты основного тона, уже связанного с артикуляцией — это так называемая «внутренняя частота» (*intrinsic pitch*) гласных. Установлено, что каждая гласная у данного диктора имеет характерную частоту F_0, повышающуюся по мере повышения положения средней поверхности языка в ротовой полости. В языке Хинди разница между низкими и высокими гласными может достигать 40 Гц [185]. В итальянском языке гласным */a, o, e, u, i/* соответствуют частоты $F_0 = 130$, 135, 137, 140 и 139 Гц, причем в некоторых контекстах разница между */a/* и */u/* может доходить до 20 Гц [173]. Различие по внутренней частоте F_0 уменьшается с понижением F_0 [56], а для коротких гласных в немецком языке разница в F_0 больше, чем для

длинных, хотя здесь велико влияние индивидуальности диктора [158].

Причины различия частоты основного тона у гласных не вполне ясны. Скорее всего, повышение F_0 возникает как результат совместного действия нескольких факторов: изменения сопротивления аэродинамическому потоку при сужении речевого тракта, понижения частоты первой форманты, влияния подъема языка на натяжение мышц гортани и иррадиации нервного возбуждения от центров, управляющих движениями языка, на центры, управляющие напряжением мышц гортани. Отмечается, что изменения внутренней частоты основного тона зависят от индивидуальных особенностей дикторов и потому отличаются большой вариативностью.

Частота основного тона служит ярким примером того, что в речевом сигнале один и тот же параметр одновременно несет информацию разного типа. Эта информация кодируется как абсолютными значениями частоты F_0, так и ее изменениями. Первое возможно потому, что известны границы диапазонов мужских и женских голосов. Так, для мужских голосов 70 Гц $\leqslant F_0 \leqslant$ 240 Гц со средним значением около 130 Гц, а для женских голосов 140 Гц $\leqslant F_0 \leqslant$ 450 Гц со средним значением около 250 Гц [36], тогда как средняя частота детских голосов — около 360 Гц. Все же, по-видимому, абсолютные значения играют небольшую роль в передаче просодической информации — более важны значения F_0 относительно некоторого уровня, характерного для определенного голоса или регистра.

Дифференциальный порог восприятия F_0, т. е. минимально различимая разница, по [65], составляет 0,3—0,5% от F_0, или меньше 1 Гц. В [131] указывается еще меньшая величина — 0,3 Гц на частоте 120 Гц. Следовательно, в диапазоне мужских голосов оказывается различимыми около 570 градаций основного тона, а весь диапазон F_0, включая женские голоса, соответствует примерно 1270 градациям, т. е. больше 2^{10}. На верхней границе диапазона F_0 скорость передачи информации составляет примерно 5 Кбит/с. Это довольно высокая скорость, обеспечивающая передачу большого разнообразия просодических типов. К тому же наблюдаются и высокие скорости изменения F_0 — до 6000 Гц/с [36], что соответствует 60 Гц на каждом периоде основного тона с частотой в 100 Гц.

Известно, что слуховое восприятие очень чувствительно к нарушениям правильных соотношений в мелодическом контуре. Например, ошибка всего лишь в одном периоде следования импульсов голосового источника в системах вокодерной телефонии сразу отмечается аудиторами как сильно ухудшающее качество передачи речи.

Речевое сообщение разделяется не только на синтаксические, но и на просодические фразы, причем границы этих фраз часто не совпадают. Иногда синтаксическая фраза (от паузы

до паузы) содержит две или более просодических фраз. В случае незавершенной интонации просодическая фраза объединяет две синтаксические. Просодическая фраза характеризуется определенным контуром F_0, который называется базовым. На фоне этого базового контура формируются локальные изменения частоты основного тона, которые передают информацию путем изменения максимальных и минимальных значений F_0 на некотором интервале времени, а также скорости и направления движения F_0. Предположительно, базовому и локальному контурам соответствуют два физических механизма управления частотой основного тона. Медленные изменения F_0 в базовом контуре приписываются влиянию подъема и опускания гортани, когда вследствие механических связей черпаловидные хрящи поворачиваются и изменяют натяжение голосовых складок [27, 152]. Сравнительно быстрые изменения F_0 отождествляются с напряжением мышц внутриголосовых складок. Разница в скорости изменения частоты основного тона в базовом контуре и локальных изменениях объясняется различием в инерционности движения гортани и напряжения внутрискладочных мышц.

В ряде языков на протяжении одной просодической фразы в базовом контуре F_0 сначала быстро нарастает, а затем медленно спадает к концу фразы. Предложено несколько способов описания формы базового контура, причем можно предположить, что эта форма зависит и от языка. Так, в [56] базовый контур аппроксимируется отрезками прямых линий (рис. 4.22), в [29] используется плавно поднимающаяся и спадающая кривая (рис. 4.23). Одна из версий описаний базового контура задает не только минимальные, но максимальные

Рис. 4.22. Линейная аппроксимация базового контура частоты основного тона

Рис. 4.23. Аппроксимация базового контура частоты основного тона с плавным подъемом

значения F_0 в виде прямых линий разного наклона, так что диапазон изменений уменьшается к концу фразы (рис. 4.24) [137]. Для шведского языка характерно не плавное, а ступенчатое понижение базовой линии F_0 (рис. 4.25) [83].

В [105, 107] базовый контур и локальные изменения F_0 описываются как реакции двух линейных инерционных

звеньев, причем все сигналы рассматриваются в логарифмическом масштабе частот:

$$\ln F_0(t) = \ln F_{0\min} + \sum_{i=1}^{I} A_i G_i(t - T_{0i}) + \sum_{j=1}^{J} B_j [Q_j(t - T_{1j}) - Q_j(t - T_{2j})],$$

где

$$G_i(t) = \begin{cases} \alpha_i^2 t e^{-\alpha_i t}, & t \geqslant 0, \\ 0, & t < 0, \end{cases}$$

$$Q_j(t) = \begin{cases} \min\left[1 - (1 + \beta_j t) e^{-\beta_j t}, \ \theta\right], & t \geqslant 0, \\ 0, & t < 0. \end{cases}$$

Здесь $F_{0\min}$ — минимальное значение частоты F_0 на базовом контуре, I — число фразовых команд, J — число локальных (акцентных) команд, A_i, B_j — амплитуды команд, T_{0i} — момент подачи фразовой команды, представляющей собой бесконечно короткий импульс с амплитудой A_i; T_{1j} и T_{2j} — начало и конец акцентной команды, представляющей собой прямоугольный импульс длительностью $T_{2j} - T_{1j}$, α_i — характеристическая частота (угловая) линейного звена для i-й фразовой команды, β_j — характеристическая частота для j-й акцентной команды, θ — параметр, определяющий максимальный уровень акцентной компоненты, обычно равный 0,9. Частота $\alpha = 2{,}8\ \text{с}^{-1}$, и она линейно уменьшается с ростом длительности фразы, а $\beta = 20\ \text{с}^{-1}$, и она постоянна. Амплитуды A_i и B_j линейно возрастают с увеличением длительности фразы, причем эта длительность измеряется числом слогов. Такое описание базового контура и локальных акцентов позволяет аппроксимировать реальные контуры с очень высокой точностью (рис. 4.26). На этом рисунке видно, что просодическая фраза характеризуется обновлением базового

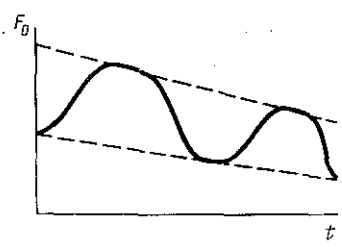


Рис. 4.24. Аппроксимация базового контура частоты основного тона типа «шляпы»

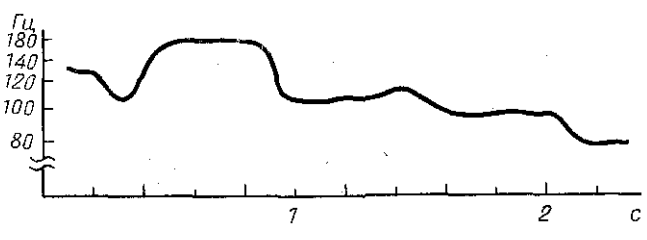


Рис. 4.25. Ступенчатый контур базовой линии в шведском языке

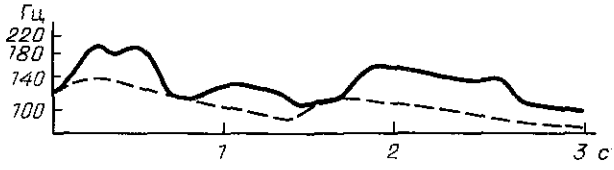


Рис. 4.26. Аппроксимация базового контура и локальных акцентов по [106]

контура, хотя в сложно-сочиненных предложениях у некоторых дикторов вместо обновления базового контура синтаксическая граница маркируется удлинением последнего сегмента и увеличением числа акцентов [106].

Несмотря на столь хорошую аппроксимацию контуров F_0, трудно согласиться с тем, что управление высотой гортани, соответствующей базовой линии F_0, осуществляется с помощью очень коротких импульсов. Импульсное управление подразумевает очень большую амплитуду, что энергетически невыгодно, и имеет своим следствием баллистическое (неуправляемое) движение на интервалах между импульсами. Из рис. 4.27 видно, что базовая линия может изменяться очень быстро — почти ступенчато, и не наблюдается затянутого инерционного движения. Поэтому можно предположить, что и управление базовой линией также происходит с помощью импульсов конечной длительности, как и управление быстрой компонентой.

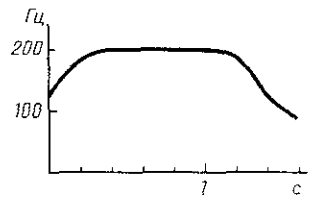


Рис. 4.27. Контур базовой линии фразы «Голованова жила в Иваново» с логическим ударением на последнем слове

Аппроксимация в логарифмическом масштабе представляется физиологически и перцептивно оправданной. Во многих случаях достигаются хорошие результаты при использовании линейной аппроксимации в логарифмическом масштабе [196]. Правдоподобно также и суммирование базовой и быстрой компонент в логарифмическом масштабе.

Быстрая компонента содержит локальные изменения частоты основного тона, обычно привязанные к ударным слогам слов, причем особо выделяется ударный слог слова, на который приходится фразовое ударение — акцент. Такая привязка локальных изменений F_0 к ударным слогам подразумевает зависимость планирования контура F_0 от временно́й структуры высказывания. Поэтому при синтезе речи сначала необходимо разделить текст на просодические фразы. Поскольку при синтезе произвольного текста прагматические и семантические знания очень ограничены, то для формирования просодии нужно в максимальной степени использовать лексические

и грамматические правила. Наименьшая текстовая единица — дыхательная группа, обычно равная грамматическому предложению, отделенному знаками препинания /./, /!/ или /?/. Внутри дыхательной группы выделяются пунктуационные группы, разделенные знаками /,/, /;/ или /:/. Каждая из этих групп может содержать одну или более просодических фраз.

В просодической фразе нужно разметить ударные слоги, чередующиеся с определенным ритмом, выделить среди них акцентные слоги, на которые приходится фразовое ударение. Просодическая фраза может содержать один или несколько акцентов, причем последний акцент называют ядром. В доядерной позиции наблюдается большая вариактивность контура F_0 [119], тогда как поведение F_0 в ядре во многом определяет тип фразы и вследствие этого ограничено в своих вариациях.

Такое планирование просодических контуров подразумевает анализ всего предложения, но существует мнение, что формирование F_0 на отдельных сегментах происходит рекурсивно, т. е. без заглядывания вперед по времени, хотя здесь велико влияние индивидуальных тактик управления контуром F_0 у разных дикторов. Так, в [73] обнаружено, что только у одного диктора из трех наблюдается зависимость между значением первого пика F_0 и длиной фразы. В [83] также отмечается это явление, и, кроме того, найдено, что от длины фразы не зависит и величина ступенчатого падения F_0, которое остается постоянным (в логарифмическом масштабе) для последовательности ступенек, что дает возможность диктору не планировать весь контур, а вычислять его рекурсивно от ступеньки к ступеньке.

Минимальное значение частоты $F_{0\min}$, достигаемое в конце просодической фразы, меньше всего зависит от типа просодической информации, и характеризует нижнюю границу диапазона для данного диктора (точнее нижнюю границу одного из регистров, в котором в настоящий момент говорит диктор). Минимальное значение $F_{0\min}$ асимптотически понижается с увеличением силы ударных групп во фразе, и если это число больше 4, то $F_{0\min}$ *практически остается постоянным* [197]. Это свойство контуров F_0 предлагается использовать для нормализации, т. е. для оценки текущего отклонения F_0 по сравнению с $F_{0\min}$. Замечено, однако, что если две просодические фразы не разделены паузой, то $F_{0\min}$ первой фразы может быть больше $F_{0\min}$ второй фразы, тогда как при наличии физической паузы минимальное значение в обеих фразах одинаково [83]. В методе, описанном в [105], это явление моделируется автоматически, причем выявляется зависимость значения $F_{0\min}$ от длительности первой просодической фразы.

Контур частоты основного тона F_0 передает два вида информации — о типе высказывания и степени семантической выделенности слов или словесных групп. В дополнение к этому, как обсуждалось выше, передается информация и об эмоцио-

нальном состоянии. С целью классификации контуров F_0 в [92] была предложена комбинация тонального описания и описания формы. Предполагается, что диапазон частот основного тона для каждого диктора от $F_{0\min}$ до $F_{0\max}$ разбит на четыре тональных уровня, и каждая просодическая фраза начинается и кончается на одном из этих уровней. Форма контура описывается как комбинация постоянного уровня и возрастающего и ниспадающего движения F_0. Всего в [92] различается 10 типов фраз, включая общий и специальный вопрос, восклицание, большую и малую завершенность и т. д. Эта классификация находится в основе многих работ по исследованию роли частоты основного тона в передаче информации.

Фразовое и словесное ударения кодируются комбинациями высоких и низких значений F_0 с разным сдвигом относительно ударного слога как различительным признаком. Обычно различаются две степени ударения и безударное состояние. Фразовое ударение (акцент) представляется самым высоким значение F_0, следующим за высоко-низкой точкой словесного ударения. В шведском и датском языках величина F_0 фразового акцента остается примерно одинаковой независимо от положения относительно начала просодической фразы, тогда как величина F_0 на словесном ударении зависит от общего контура F_0 [83, 197]. Чем длиннее фраза (в количестве ударных слогов), тем выше F_0 на первом фразовом акценте, и следующий за ним минимум F_0 также оказывается повышенным. Это наблюдается в шведском, датском, японском и английском языках [83]. В шведском языке смещение фразового ударения вправо по времени приводит к относительному постоянству общего контура F_0 от начала фразы до фразового ударения. На фоне этого контура поднимаются локальные пики словесных ударений, причем разность значений F_0 локального пика и базовой линии остается примерно постоянной независимо от его положения в верхней или нижней области диапазона F_0 данного диктора [83].

Наиболее информативным оказывается поведение F_0 в области ударных слогов. Так, по [197] при различении вопросительных, утвердительных и незаконченных фраз, аудиторы опираются на контур F_0 не в конечной части фразы, а на ударных сегментах. При этом большую роль играет степень синтаксической маркированности — чем больше синтаксической информации о вопросе или незавершенности, тем круче спад контура F_0 и тем больше он похож на контур утвердительной фразы. Одновременно с этим, положение максимума F_0 относительно ударного слога несет дополнительную информацию о речевом высказывании — если $F_{0\max}$ приходится на ударный слог, то это уточнение или противопоставление, если пик F_0 сдвигается на следующий слог, а на ударный слог приходится локальный минимум F_0, то это неуверенность или недоверие, причем характер

интонации тяготеет к категориальному [174]. Это дает основания в некоторых синтезаторах использовать скачкообразные переходы между целевыми значениями F_0, как это предлагается в [139], где контур F_0 рассчитывается как

$$F_0(t) = F_{0\min} f(N) f(T),$$

где $F_{0\min}$ — минимальное значение частоты основного тона для данного диктора, $f(N) = Nd^i$, $f(T) = W^T$, $d = 0{,}8$, $W = 1{,}5$, i — номер акцента, $T = 1$ для высокого регистра, $T = 0$ для среднего регистра и $T = -1$ для низкого регистра.

Система правил, управляющих частотой основного тона в каждом языке своя, поэтому, наряду с некоторыми общими свойствами, имеются сильные языковые различия. Исследованию характеристик русского языка посвящен ряд работ, в том числе [5, 39, 56]. В [5] было сконструировано 7 интонационных образов, близких по идее к классификации [92]. В других работах предлагаются несколько отличающиеся системы интонационных единиц. Разнообразие этих систем, очевидно, отображает большую вариативность, свойственную контурам F_0 и объективную трудность их классификации.

По [56] наиболее устойчивы характеристики повествования, общего и частного вопросов, а также восклицания. Повествование характеризуется понижением частоты F_0 на ударном слоге слова, находящегося в позиции фразового ударения, от уровня средней индивидуальной частоты $\bar{F}_0$ до уровня средней минимальной частоты $\bar{F}_{0\min}$. При этом интонационный центр (максимум F_0) обычно совпадает с последним знаменательным словом фразы. Общий вопрос кодируется резким повышением F_0 на ударном слоге наиболее важного слова до уровня средней максимальной частоты $\bar{F}_{0\max}$. На заударных словах F_0 падает вплоть до средней минимальной частоты $\bar{F}_{0\min}$. Вопрос с вопросительным словом имеет подъем F_0 на вопросительном слове, стоящем в начале предложения, до уровня выше средней частоты $\bar{F}_0$ с сохранением высокого уровня F_0 вплоть до последующего ударного слога, на котором происходит падение F_0 до уровня средней частоты $\bar{F}_{0\min}$. Интонация незавершенности приближается к интонации общего вопроса, но с меньшим интервалом повышения F_0, причем на заударных слогах возможны высокие значения F_0. Восклицание имеет восходяще-нисходящий контур F_0, причем для него характерна большая вариативность.

Выраженность пиков F_0 на словесных ударениях зависит от типа фразы — в общем вопросе эти пики ослаблены или вообще отсутствуют, при сильном выделении какого-либо слова в повествовательном предложении эти пики также сглаживаются [56]. Ситуация осложняется тем, что характеристики контуров F_0 в разговорном стиле сильно отличаются от литературного стиля речи. Так, в [167] сообщается, что

в русской спонтанной речи наблюдаются пилообразные контуры F_0 с максимумами, соответствующими ударениям в неконечной позиции перед конечным акцентом. Эти контуры не принадлежат ни к одному из контуров, наблюдающихся в литературном стиле речи.

Множество форм контуров основного тона отличается большим разнообразием, и значительная часть правил управления интонацией не только не формализована, но даже и неизвестна. Эта ситуация несколько облегчается тем, что при синтезе речи число возможных ситуаций, а, значит, и форм речевого сообщения, значительно меньше, чем при разговоре людей между собой. Основное назначение синтезатора (по крайней мере в ближайшем будущем) состоит в простой передаче информации, что может быть обеспечено лишь с помощью утвердительных и вопросительных предложений. Тонкости эмоционального плана на первых порах могут быть опущены. Но и при таком подходе необходимо обеспечить адекватное кодирование просодическими средствами семантически выделенных групп слов.

С целью выяснения некоторых характеристик контуров основного тона были проведены эксперименты с использованием речевого материала, включающего последовательность повествовательных предложений, прочитанных как информационный текст в стиле лекции, а также предложения, отличающиеся либо длиной, либо местом расположения семантически выделенного слова во фразе, включая вопросительные типы фраз. В этих экспериментах было установлено, что в русском языке базовая линия имеет форму, отличную от той, которая наблюдается в шведском, английском и японском языках. Эта линия начинается с крутого подъема, длительность которого занимает около 100 мс. Достигнутый уровень $F_{0б}$ сохраняется примерно постоянным вплоть до семантически выделенного слова во фразе, а затем происходит спад до уровня F_{0min} на интервале около 200 мс. Величина постоянного уровня $F_{0б}$ зависит от ряда факторов, в том числе от степени новизны информации, разговорного стиля и т. д. Если семантически выделенное слово находится в конце фразы, то и уровень базовой линии $F_{0б}$ держится постоянным вплоть до ударного слога этого слова. Если же выделенное слово находится в начале фразы, то вся оставшаяся часть фразы произносится при постоянном значении частоты основного тона базовой линии $F_{0б}=F_{0min}$, где F_{0min} — минимальное значение для текущего регистра данного диктора (рис. 4.28).

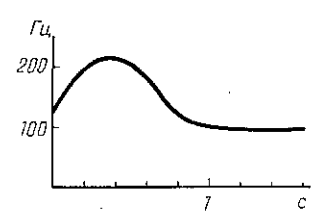


Рис. 4.28. Контур базовой линии фразы «Голованова жила в Иваново» с логическим ударением на первом слове

Короткие фразы, состоящие из одного—двух слов, обладают формой базовой линии, похожей на непрерывно спадающие линии, как, например, на рис. 4.23 или рис. 4.24. Это видно из рис. 4.28, если оставить лишь первое акцентное слово во фразе «*Голованова* жила в Иваново». Однако более длинные фразы демонстрируют существование плато на базовой линии. Если синтагма содержит две просодические фразы, то в конце первой фразы происходит спад базовой линии к $F_{0\min}^{(1)}$, затем вновь частота основного тона поднимается до постоянного уровня $F_{0б}^{(2)}$ и держится на нем до конца второй фразы. Величины $F_{0б}^{(1)}$, $F_{0\min}^{(1)}$ и $F_{0б}^{(2)}$ зависят от семантических и синтаксических факторов.

Не вполне ясно, меняется ли величина постоянного уровня $F_{0б}$ непрерывно, или существуют предпочтительные квантованные уровни. В последнем случае необходимо выяснить число таких уровней, которое не должно быть слишком велико, а также их расположение и зависимость от индивидуальных особенностей диктора. Есть основания полагать, что минимальное значение $F_{0\min}$ в конце фразы перед паузой имеет не произвольную величину, а находится в некотором отношении к абсолютному минимуму частотного диапазона основного тона данного диктора, т. е. $F_{0\min} = kF_{0\min\min}$, где $k < 1$. Таким образом, по крайней мере, конечное значение F_{0min} тяготеет к дискретным уровням, причем число этих уровней невелико—около 7. Нечто подобное можно ожидать и от постоянного уровня базовой линии $F_{0б}$.

Повествовательное законченное предложение обладает ниспадающим контуром F_0 на последнем ударном слоге. При этом, если базовая линия уже установилась на уровне $F_{0\min}$, то для обозначения спада F_0 перед ударными гласными или в его начале происходит быстрый подъем F_0. В повествовательном незаконченном предложении разница между максимальным и минимальным значением F_0 на интервале ударного слога невелика. В предложениях с общим вопросом ударный слог слова, к которому обращен вопрос, имеет в среднем подъем частоты основного тона, причем разница между минимальным и максимальным значениями F_0 должна быть не меньше определенной величины, иначе предложение не будет воспринято как вопросительное.

Диапазон частот основного тона для диктора обычно находится в пределах трех октав, причем для многих исследователей привлекательна идея квантования F_0 по шкале музыкальных тонов. Так, в [30] принимается, что вариации F_0 должны занимать, по крайней мере, октаву, и коммуникативный тип предложения определяется подъемом или спадом F_0 на определенное число музыкальных тонов. Например, в законченном повествовательном предложении в последнем слоге последнего слова F_0 падает на кварту или квинту (1,336 или 1,496 от исходного уровня), а в незаконченном

предложении F_0 падает лишь на секунду или терцию (1,122 или 1,259 от исходного уровня). В вопросительном предложении используется подъем F_0 в предпоследнем слоге последнего слова на кварту, квинту или сексту, а в последнем слоге — еще на терцию или квинту, в зависимости от степени выраженности вопроса. Акцентное слово во фразе выделяется путем повышения F_0 на секунду, терцию или кварту.

Допуская, что тонкие изменения эмоциональной скорости окраски речевого высказывания передаются путем изменения частоты основного тона на величины, кратные дифференциальному порогу, т. е. около 0,3 Гц, можно принять, что коммуникативный тип предложения и степень семантической значимости могут кодироваться изменениями F_0 на кванты, соответствующие десяткам и сотням дифференциальных порогов, т. е. тонам музыкальной шкалы.

§ 4.3. Просодический анализ текста

Как мы видели в предыдущих разделах, просодические характеристики играют важную роль в передаче информации о содержании речевого высказывания. Вместе с тем, в письменном тексте содержится очень мало указаний на просодические параметры — обычно это знаки препинания: запятые, точки с запятой, двоеточия, точки, а также вопросительный и восклицательный знаки. Этого явно мало для управления просодическими характеристиками при синтезе речи, поэтому нужно найти просодические законы, исходя из анализа самого текста.

Прежде всего разделим общую задачу синтеза речи по тексту на три группы задач. К первой группе относится задача чтения заранее неизвестного текста типа учебника или художественной литературы. В этой задаче просодические характеристики могут быть найдены только путем синтаксического и семантического анализа текста. К сожалению, методы семантического анализа пока еще не достигли уровня, необходимого для использования в синтезаторах, поэтому выразительность синтетической речи не слишком высока. Имеются, правда, сообщения о том, что слепые при использовании читающей машины предпочитают как можно более нейтральное произношение. Если это верно, то задача формирования просодии в данном случае облегчается. Не ясно, однако, чем вызвано такое желание, и не связано ли оно просто с плохими просодическими правилами, формирующими неадекватные смыслу текста интонацию, длительность и интенсивность сегментов.

Вторая группа задач относится к такому синтезу речи, в котором текст сопровождается некоторыми просодическими маркерами, назначаемыми человеком. Это возможно в гово-

рящих машинах для немых или в тех случаях, когда текст формируется человеком (объявления, инструкции). При этом могут быть обозначены темп, ударные слоги в словах, акцентные слова во фразах, размечена относительная информационная важность тех или иных групп слов. Необходимо отметить, однако, что и в этом случае невозможно полностью обозначить длительность каждого сегмента и детали интонационного контура, так что некоторый дополнительный автоматический анализ текста все-таки необходим. Кроме того, трудоемкость просодической разметки может свести на нет эффективность использования синтезатора.

Третья группа задач имеет место тогда, когда текст формируется неким источником информации, например, системой искусственного интеллекта в речевом диалоге с человеком. В этом случае источник сообщения осведомлен о типе фразы, акцентных группах, информационной нагрузке, размещении ударений в словах. Это знание позволяет расставить просодические маркеры, аналогично тому, как это сделал бы человек, но окончательное формирование просодии все же должно происходить автоматически, как и в предыдущем случае.

Итак, мы видим, что в зависимости от типа задачи синтеза, в большей или меньшей степени, но все же необходим такой анализ текста, который позволил бы сформировать адекватные просодические характеристики. Такой анализ нужен еще и потому, что восприятие устной речи сильно отличается от чтения текста, где всегда можно вернуться к непонятному месту, сделать паузу для осмысления прочитанного. Восприятие устной речи сопровождается периодическим понижением внимания, связанным, вероятно, с процессами понимания, вследствие чего отдельные участки слитной речи воспринимаются менее отчетливо. Поэтому и озвучивать текст нужно таким образом, чтобы в нем просодически выделялись наиболее информативные участки, тогда как менее важные фрагменты могут быть синтезированы с меньшей степенью разборчивости без потери смысла текста. Известно, например, что чересчур четко артикулируемая речь вызывает раздражение, по-видимому, в связи с непроизвольным повышением внимания и увеличением затрат мозговых ресурсов для понимания не только важных, но и малозначительных фрагментов речи. Аналогичный эффект вызывает такая манера речи, в которой часто используются длительные паузы, предполагающие особую важность последующего сообщения.

Еще один аспект автоматического анализа просодических характеристик текста связан с необходимостью придания синтезатору способности менять стиль произношения — от официального до дружеского. Кроме того, от одного и того же синтезатора может потребоваться имитация разных дикторских манер произношения. Выбор того или иного стиля

относится уже не к самому синтезатору, а к интеллектуальному блоку, формирующему речевой диалог с человеком.

Одной из достаточно простых просодических характеристик является *темп* — среднее число слогов в единицу времени, определенное на длинном интервале времени. Как мы видели в § 4.1, темп речи отображает тип нервной системы говорящего и условия разговора. В синтезаторе необходимо предусмотреть возможность как ручной, так и автоматической регулировки темпа речи, поскольку скорость восприятия информации у разных людей также различна. Наряду с этим, существуют и некоторые правила управления темпом. На периоде, содержащем несколько фраз, темп никогда не поддерживается строго постоянным. Наиболее заметно различаются начальная, средняя и конечные стадии периода, причем соотношения темпа в них могут быть самыми различными, в зависимости от содержания периода. Информативные участки произносятся в относительно медленном темпе. Малоинформативные участки, вроде вводных фраз, произносятся быстрее и тише. В разговорном стиле к концу фразы может произойти ускорение с деформацией конечных звуков [56]. При переспросе та же фраза произносится медленнее, причем каждое слово получает самостоятельное ударение.

Считается, что почти в каждой синтаксической фразе (от точки до точки) могут находиться две интонационные фразы, каждая с подъемом и падением частоты основного тона, одна из которых связана с группой подлежащего, а другая — с группой сказуемого. С целью определения границы между интонационными фразами в синтезаторе *DECTalk* используется словарь глаголов, которые в английском языке однозначно маркируют начало интонационной глагольной фразы [137]. Отмечается, однако, что лучше пропустить эту границу, чем поставить ее в неправильном месте — в первом случае просто создается впечатление слишком быстрой речи, тогда как во втором случае ложная граница дезориентирует слушателя при анализе фразы.

Такт относится к другой просодической характеристике, связанной с временными интервалами. Под тактом понимается число слогов, объединенных одним полным словесным ударением. В русском языке так же, как и во многих других языках, в разговорной речи наблюдается тенденция к ритмизации, т. е. к сохранению постоянства размера такта, причем чаще всего встречается интервал в 2—3 безударных слога между ударными. В результате ритмизации одно и то же слово может оказаться ударным или безударным в зависимости от требований ритма, а безударность может привести к сильной фонетической деформации и потере слогов. Распределение ударений в процентах по фразе в разговорном и полном стилях по данным [56] показано в табл. 4.7.

Таблица 4.7. Распределение ударений, %

Слог	1	2	3	4	5
Начало фразы	43	33	19	4,5	0,5
Конец фразы	36,5	47	15	1,5	
Номер от конца					
Середина фразы	34	32	19,5	6	1,3
Интервал между ударениями					
Полный стиль	28	31,5	17,5	8	2
Середина фразы					
Интервал между ударениями					

Соседство ударных слогов очень редко: в разговорном стиле — 7%, а в полном стиле — 12%.

Стремление к ритмизации в разговорном стиле приводит к редукции и выпадению отдельных звуков и даже групп звуков в безударной позиции, тогда как в полном стиле (например, при публичном выступлении) такое выпадение встречается редко. Исходя из требований ритма, в разговорном стиле словесное ударение может попасть на предлоги, служебные слова, частицы и местоимения, что не характерно для полного стиля. Свойство ритмизации является существенным фактором, облегчающим формирование просодических характеристик при синтезе речи по тексту.

Паузы определяются знаками препинания, причем наименьшая пауза соответствует запятой, а наибольшая — точке, вопросительному или восклицательному знакам. Однако как показали специальные исследования, в речи более 36% пауз не соответствует никаким пунктуационным знакам [166]. Как мы видели в § 4.1, длительность паузы зависит и от среднего темпа, укорачиваясь с ускорением темпа. В периоде, содержащем несколько фраз, должна присутствовать дыхательная пауза, которая длиннее обычной паузы, обозначающей конец предложения. Расчет момента появления дыхательной паузы зависит от предполагаемого объема легких (см. п. 5.4.4), причем обычно дыхательная пауза совпадает с паузой конца фразы. Для 80% дыхательных групп число ударных слогов равно 1—4, и во всяком случае не превышает 16 слогов [166]. Длительная пауза предшествует фрагменту речевого сообщения, несущему повышенную информационную нагрузку. Имеется связь между длиной паузы и интонацией: нисходящей интонеме соответствует долгая пауза, а восходящей — короткая [166].

Одной из главных просодических характеристик фразы является *акцент* — выделение наиболее семантически важного слова посредством наиболее сильного ударения. Так, простая фраза «мальчик пришел домой» может быть прочитана по крайней мере тремя разными способами, не считая эмоциональных нюансов. Акцент, поочередно смещаемый от пер-

вого к последнему слову, меняет смысл фразы. Неправильно поставленный акцент, или отсутствие акцента, например, во фразе «шасси не выпущено» может привести к игнорированию пилотом аварийного сообщения. В задачах первого рода, т. е. при чтении произвольного текста обычно неизвестно место акцентного слова. В этом случае ни одно из слов не получает акцентного ударения, а в полном стиле каждое слово несет ударение, несколько более слабое, чем акцентное. В разговорном стиле, учитывая ритмику, слово может оказаться и вовсе безударным. Таким образом, во фразе имеется как минимум три степени ударения: акцент, ударное слово и безударное слово. Эти степени характеризуются уровнем частоты основного тона, интенсивностью, длительностью и точностью артикуляции. В безударной позиции наиболее вероятна редукция или полное выпадение звуков. Если синтаксическая фраза содержит две интонационные фразы, то в каждой из них может быть по акцентному слову.

В разговорном стиле часто встречающиеся слова или словосочетания редко бывают ударными, они могут подвергаться сильной фонетической деформации, что не замечается слушателями. В полном стиле ударение никогда не падает на односложные предлоги, частицы, союзы и местоимения, тогда как двухсложные слова этих типов могут быть слабоударными [56].

Наконец, в каждом слове нужно определить ударный слог или даже два, если слово многосложное. По теории Чомского — Халле, можно определить положение ударения в слове, привлекая понятие сильных или слабых слогов. Однако, например, для английского языка даже улучшенные правила дают лишь 65% правильных решений [120]. В [137] обращается внимание на то, что 90% двухсложных существительных имеют ударение на первом слоге, тогда как лишь 15% двухсложных глаголов имеют ударение на первом слоге. В русском языке для двухсложных слов число ударений на первом слоге такое же, как и на втором слоге [42]. В трехсложных словах вероятность ударения на втором слоге вдвое выше вероятности ударения на первом или третьем слоге. В значительной мере проблема определения места ударения в слове решается путем запоминания базового словаря. Такой словарь, кстати, позволяет определить и часть речи, к которой принадлежит слово. По [137], 2000 слов покрывает 70% текстов, а в синтезаторе *DECTalk* используется 6000 слов, покрывающих около 90% текстов. Считая среднюю длину слова в 5 звуков, найдем, что на каждую тысячу слов требуется около 4 Кбайт, что вполне приемлемо. Сочетая словарь с правилами сдвига ударения при изменении падежа, числа, времени и т. д., можно добиться высокой точности определения места ударения в слове.

§ 4.4. Фонетический анализ текста

После того, как расставлены фразовые и словесные ударения, необходимо преобразовать буквенную запись текста в фонетическую, по которой и выполняется формирование команд управления параметрами синтезатора. В разных языках связь между буквенной и фонетической формой текста отличается очень сильно [120, 171]. Например, в белорусском или финском языках текст записывается почти так же, как и произносится, и поэтому нет никаких трудностей в управлении синтезатором. В английском языке связь между буквенной и фонетической формами очень сложна, поэтому наиболее успешным оказывается использование словаря с фонетической транскрипцией в сочетании с довольно большим числом правил. Иногда запоминаются слова (*DECTalk* — 6000 слов), иногда морфемы (*MITalk* — 12000 морфем, синтезатор Белловских лабораторий *CONVERSANT* — 43000 морфем). Число правил в английском языке доходит до 500 [137]. В синтезаторе *CONVERSANT* на хранение словаря и правил отводится 900 Кбайт. Правила необходимы в любом случае, так как могут появиться слова, отсутствующие в памяти.

Процедура преобразования букв в фонемы очень ответственна, поскольку совершенные на этом этапе ошибки не исправляются и полностью переходят в синтетическую речь, ухудшая разборчивость. Одна из весьма удачных систем состоит в записи произношения для некоторого множества слов и их сравнении с неизвестным словом путем относительного сдвига до достижения наибольшего сходства [90]. Слова при этом записываются в двоичном коде — с различением лишь гласных от согласных. Погрешность определения произношения неизвестного слова в этой системе равна 9%, что сравнимо с погрешностью других, более сложных систем. Все же эта ошибка довольно велика по сравнению с системой *DECTalk*, где ошибки преобразования равны только 3%.

В русском языке соотношения между буквенной и фонетической записью текста более простые, чем в английском, но более сложные, чем в белорусском. Соответственно, и правила преобразования также более простые, чем в английском языке. Сводка этих правил для полного стиля произношения может быть составлена по [17, 21, 52]. Первая группа правил относится к изменению признака звонкости/глухости согласных.

1. Сонорные звуки /М, Н, Л, Р/ становятся глухими между глухими согласными, а также в начале фразы перед глухими согласными и в конце фразы после глухого согласного.

Признак звонкости/глухости сохраняется у согласного, находящегося внутри слова или на стыке предлога со словом перед гласным или сонорным или конструкцией /В+гласный/, но принимает значение признака звонкости/глухости последующего согласного, даже отделенного звука /В/.

На стыке слов, слова с частицей или слова с предлогом, входящим в состав следующего слова, признак звонкости/глухости согласного или группы согласных принимает значение этого признака для последующего согласного, а перед гласным или сонорным (даже отделенным звуком /В/) звонкие согласные оглушаются.

В конце фразы перед паузой согласные или их группа всегда оглушаются.

2. Если согласный находится перед мягким согласным с тем же местом артикуляции, то он смягчается, кроме случая, когда он находится на стыке корневых морфем в сложном слове. Звуки /Ц, Ж/ — всегда твердые, а /Ч, Щ/ — всегда мягкие.

3. В окончаниях /ОГО, ЕГО/, а также в словах «ЕГО, НЕГО, СЕГО», согласный /Г/ заменяется на /В/ (но не в слове «МНОГО»).

4. Внутри слова в группе согласных выпадает внутренняя согласная в сочетаниях: /СТН/—/СН/, /ЗДН/—/ЗН/, /СТЛ/—/СЛ/, /РДЦ/—/РЦ/, /НТСК/—/НСК/, /СТСК/—/ССК/, а также /ЛНЦ/—/НЦ/.

5. Внутри фонетического слова происходят следующие преобразования: /СШ/—/ШШ/, /ЗЖ/—/ЖЖ/, /ЗШ/—/ШШ/, /СЖ/—/ЖЖ/, т. е. последующий согласный удваивается. Фонетическим словом считается односложный предлог и слово или слово и частица.

6. В корне и на стыке корня и суффикса конструкции /СЧ/, /ЗЧ/ и /ЖЧ/ переходят в /Ш/.

7. На стыке приставки или предлога с корнем /СЧ/ и /ЗЧ/ переходят в /ЩЧ/.

8. Окончания /ТСЯ, ТЬСЯ/ переходят в /ЦА/.

9. /ЧН/ переходит в /ШН/ в словах «скучно, конечно, пустячный, нарочно, яичница».

10. В морфеме /ЧТО/ /ЧТ/ переходит в /ШТ/.

11. В словах «легкий, легче» /ГЧ, ГК/ переходят в /ХК, ХЧ/.

12. Если двойной букве не соответствует долгий согласный, то одна буква вычеркивается.

13. Качество гласного меняется от окружения. Предударный /О/ переходит в /А/.

Особые правила фонетического транскрибирования текста требуется применять к словам, заимствованным из других языков. По [8], число таких слов достигает 10—20%.

Как видно, эти правила требуют выполнения морфологического анализа для определения приставок, предлогов, корней и суффиксов, что может быть осуществлено программами разной степени сложности. Например, в синтезаторе *DECTalk* слово разбивается на морфемы путем удаления суффиксов, оставшийся корень слова сравнивается со словами из словаря, и, если оно находится, то считывают его транскрипцию, ударение и часть речи. Если в словаре такого слова не находится, то применяются правила «буква — фонема».

В разговорном стиле часто возникает редукция (ослабление, уменьшение длительности, нейтрализация) или даже полное выпадение гласных, особенно безударных [54]. Редукция гласных часто встречается в заударных неконечных /О, И, Ы/ между двумя согласными /В/. Сильная редукция или выпадение гласного происходит между двумя одинаковыми согласными, после мягкого согласного и особенно между двумя мягкими согласными, в соседстве с группой согласных, в соседстве с сонорными, фрикативными и /В/. Гласный в первом предударном слоге обычно выпадает после мягкого согласного, а также если первый предударный слог не является начальным слогом слова. Наибольшая редукция гласного возникает в первом заударном слоге (см. по этому поводу § 4.1), тогда как гласный второго заударного слога редуцируется редко — при редукции часто встречающихся слов, либо когда второй заударный слог начинается с сонорного согласного.

Редуцированный /У/ теряет огубленность и может звучать как /Ы/, /И/ или /Ъ/ (неопределенный гласный открытого ряда). После мягких согласных ударные гласные меняют качество: /А/ переходит в /Э/, /Э/ переходит в /И/, /О/ может стать более закрытым и огубленным, как /У/. В первом предударном слоге /У/ переходит в /Ы, И/ или /Ъ/. После твердых согласных /И/ может перейти в /Ъ/. Две соседние редуцированные гласные могут быть представлены лишь одним звуком.

В редких случаях может редуцироваться и ударный слог, особенно когда это связано с требованиями ритмизации. Редукция гласных приводит к появлению долгих согласных, образованию групп согласных, к которым применимы вышеупомянутые правила ассимиляции признаков звонкости/глухости и твердости/мягкости, а также правила сочетаемости согласных. В конце слова перед начальными согласными следующего слова (но не перед паузой) происходит полная редукция гласного. В отдельных случаях при расположении гласного между двумя глухими согласными может возникнуть парадоксальное явление — оглушение гласного.

В разговорном стиле также заметна тенденция к переходу взрывных во фрикативные и озвончению глухих согласных. Это отмечается и в английском языке [78], так что причины этого явления носят неязыковый характер, отражая особенности функционирования системы управления артикуляцией в быстром темпе (см. § 4.1) — при ускорении темпа и снижении энергетических затрат происходит «недорегулирование», т. е. артикуляционные цели не достигаются. В артикуляторном синтезаторе это явление реализуется не путем явного задания преобразования звуков, а путем изменения длительности и, может быть, интенсивности, команд управления. В формантном синтезаторе эти правила должны быть заданы в явном виде.

Согласные в интервокальной позиции (между двумя гласными) могут редуцироваться, и образовавшаяся пара гласных

стягивается в один звук. Особенно часто это происходит с мягким /Д/ в глаголах. Звонкие согласные редуцируются чаще, чем глухие, фрикативные — чаще, чем взрывные. Согласные /М, Н/ и звонкие согласные часто теряют смычку. Особенно охотно редуцируются мягкие согласные и звуки группы /В, З, Ж/, тогда как наиболее устойчивы глухие взрывные и фрикативные. Группы из трех, иногда двух, согласных упрощаются, теряя какой-нибудь звук. В начале слова может исчезнуть лишь небольшая группа согласных: /Ф, В, З, j/.

Итак, в зависимости от разговорного стиля правила преобразования букв в фонемы различаются. Предстоит еще большая работа по определению как самих правил преобразования, так и условий их применения.

В процессе отладки системы преобразования буквенной записи текста в фонетическую большое значение имеет способ представления лингвистических знаний. Для того чтобы иметь возможность не менять всю систему правил при изменении какого-либо участка, в [10] предлагается параллельная многоуровневая структура в виде своеобразного компилятора, которая оказывается гораздо более удобной и гибкой, чем линейная (последовательная) структура.

Формирование команд управления с учетом коартикуляции в формантном синтезаторе заключено в отдельную систему правил, а управление артикуляторным синтезатором рассматривается в гл. 9.

ГЛАВА 5

ИСТОЧНИКИ ВОЗБУЖДЕНИЯ

§ 5.1. Влияние голосового источника на натуральность синтетической речи

Характеристики голосового источника являются наиболее важными среди прочих факторов для натурального звучания синтетической речи. В спектральной области наибольшее воздействие на восприятие сигнала формантного синтезатора оказывает форма спектра импульса возбуждения в полосе ниже 500 Гц при условии, что амплитуды формант подобраны с надлежащим убыванием в зависимости от частоты [117]. Спектральные характеристики голосового источника, однако, неполностью описывают его влияние на натуральность. Основная роль принадлежит временной форме импульса голосового источника и изменениям этой формы от импульса к импульсу. Непосредственные измерения площади голосовой щели показывают, что ее форма и амплитуда изменяются

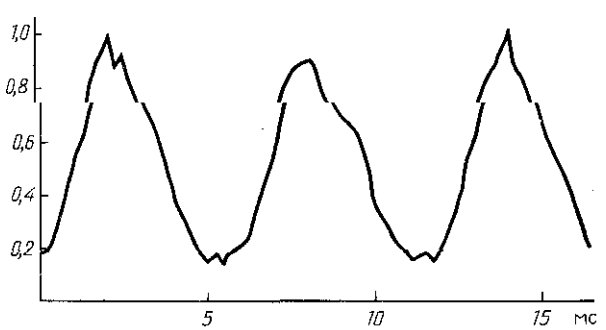


Рис. 5.1. Площадь голосовой щели по измерениям на скоростном фильме

от периода к периоду. Изменяется также длительность каждого периода и скважность — отношение длительности интервала с открытой голосовой щелью к периоду (см. рис. 5.1 по [87]).

В экспериментах с синтезаторами выяснилось, что эти изменения, хотя и малые по относительной величине (около

3%), играют важную роль в оценке натуральности синтетической речи человеком [59, 117, 181]. Имеется две причины этих быстрых, иногда называемых микропросодическими, вариаций параметров импульса голосового источника. Одна из них — физиологический тремор мышц гортани, в результате которого натяжение мышц изменяется на периоде колебаний голосовых складок. Параметрические эффекты, связанные с тремором, были исследованы с помощью формантно-артикулярного синтезатора в [59], где было найдено, что при относительном изменении периода основного тона менее 1,5% синтетическая речь воспринимается как явно машиноподобная, в диапазоне 2—4% натуральность звучания улучшается, а при 5% и более речь становится резкой и грубой. Другая причина быстрых изменений параметров импульса возбуждения состоит в случайном изменении давления над голосовой щелью вследствие вихревого (турбулентного) течения воздушного потока, вытекающего из голосовой щели с малой площадью в полость с большой площадью. Вихревое движение возникает при превышении числом Рейнольдса Re некоторой величины (1500—2000), а само это число определяется как

$$\mathrm{Re} = \rho_0 b v / \mu,$$

где v — скорость потока, ρ_0 — плотность воздуха, b — минимальный геометрический размер сужения (в нашем случае — ширина голосовой щели), μ — коэффициент вязкости воздуха. Эта турбулентность сопровождается возникновением шумов, присутствие которых является необходимым для повышения натуральности синтетической речи. Амплитуду и спектральный состав турбулентного шума мы оценим в одном из последующих разделов.

Форма голосовой щели и ее изменение во времени определяется соотношением амплитуд гармоник двухмерных упругих колебаний голосовых складок. Эти амплитуды, в свою очередь, зависят от начальных условий и распределения сил давления по поверхности голосовых складок. В изменениях этих факторов иногда наблюдается повторяемость с периодом, равным двум периодам основного тона — так называемая диплофония, при которой форма импульсов, следующих через один, более схожа, чем форма соседних импульсов.

На восприятие тембра голоса влияет поршневой источник, действующий синхронно с основным голосовым источником. Характеристики и роль поршневого источника были впервые исследованы в [59]. Поршневой источник — это источник объемной скорости, создаваемой вертикальными движениями голосовых складок. В отличие от основного голосового источника, который возбуждает акустическое колебание только во время открытой голосовой щели, поршневой источник действует непрерывно. Наибольшая объемная скорость потока, создаваемого поршневым источником, составляет около 10%

от максимальной объемной скорости потока в голосовой щели, и спектр импульсов поршневого источника спадает к высоким частотам гораздо быстрее спектра импульсов основного источника. Тем не менее, поршневой источник влияет на тембр голоса путем изменения соотношений амплитуд нижних и высших формант. Механизм этого влияния заключается в сдвиге фаз между импульсами основного голосового источника и поршневого источника — поршневой источник оказывается в противофазе с колебаниями давления на частоте первой форманты. Это приводит к уменьшению амплитуды первой форманты, т. е. относительному увеличению амплитуд высших формант.

Различие в тембральных характеристиках разных голосов, в частности, мужских и женских, зависит от амплитуды импульсов поршневого источника, а эта амплитуда, в свою очередь, определяется толщиной и жесткостью голосовых складок вдоль голосовой щели. По сравнению с мужскими, у женщин голосовые складки более тонкие и имеют меньший модуль упругости, так что податливость складок в вертикальном направлении больше и, следовательно, больше и амплитуда импульсов поршневого источника. Еще одно отличие женских голосов от мужских состоит в том, что у женщин отношение интервала открытой голосовой щели к интервалу закрытой голосовой щели больше, чем у мужчин. В ряде случаев у женщин наблюдалось отсутствие полного смыкания голосовых складок. Для женщин также является характерным увеличение доли турбулентных шумов, что связано с относительным удлинением интервала открытой голосовой щели. Д. Клатт обнаружил, что для женских голосов характерно неполное закрывание голосовой щели, сопровождающееся генерацией шумов в диапазоне частот выше 1800 Гц. Добавление этих шумов в синтетический сигнал значительно повышает натуральность. В последнее время, однако, появились экспериментальные данные о том, что неполное смыкание голосовых складок вблизи черпаловидных хрящей часто наблюдается и у мужчин, и у женщин.

Голосовой источник и речевой тракт представляют собой единую систему, и их разделение в значительной мере является искусственным приемом, помогающим выявить различные стороны их взаимодействия. Разновидность такого взаимодействия состоит в изменении частоты и затухания формант (особенно нижних) во время открытой голосовой щели. Синхронная с основным тоном вариация параметров формант улучшает натуральность звучания и, возможно, разборчивость синтетической речи.

Необходимо отметить, что каждый из обсуждающихся факторов по отдельности дает хотя и заметное, но малое улучшение натуральности синтетической речи, и лишь использование всех этих эффектов в совокупности существенно повышает натуральность.

Внимание к механизмам возбуждения акустических колебаний в речевом тракте связано с необходимостью создания синтетических голосов, обладающих индивидуальностью, т. е. отчетливо воспринимаемой разницей тембральных характеристик, поскольку человек может получать информацию от разных синтезаторов, или же смена тона голоса может потребоваться для разделения или выделения сообщения, передаваемого с помощью одного и того же синтезатора. Есть основание также полагать, что степень натуральности, т. е. естественности речи, влияет на сложность процессов принятия решений и задержку понимания синтетической речи, а этот фактор, в конечном итоге, определяет эффективность синтезатора и готовность человека к его использованию.

§ 5.2. Параметрические модели голосового источника

Механика и акустика голосового источника отличается большой сложностью. Поэтому естественно попытаться описать форму возбуждающих импульсов чисто формально, не вдаваясь в механизмы голосообразования. Основой для параметрических моделей голосового источника служит измерение площади голосовой щели, объемной скорости и производной во времени от объемной скорости, которая является импульсом, возбуждающим акустические колебания в речевом тракте.

Наиболее доступным способом измерения площади голосовой щели является метод трансиллюминации [147], в котором источник света прикладывается к наружной поверхности шеи ниже гортани, а световой поток, модулируемый колебанием голосовых складок, регистрируется фотоприемником, расположенным внутри голосового тракта над гортанью. Наиболее точными, хотя значительно более трудоемкими, являются измерения на скоростных фильмах со скоростью около 5000 кадр/с. Пример подобных измерений по данным работы [86] показан на рис. 5.1.

Воздушный поток (объемную скорость) измеряют с помощью безотражательной трубы, предложенной в [31]. Свободный конец этой трубы имеет такую форму, что акустические волны не отражаются от него, и поэтому в системе «речевой тракт — труба» резонансы не возникают. На рис. 5.2 показаны волны объемной скорости, измеренные таким образом. Методика безотражательной трубы достаточно трудоемка, поэтому для восстановления источника голосового возбуждения обычно пользуются обратной фильтрацией речевого сигнала [180]. Метод обратной фильтрации состоит в подавлении резонансов речевого тракта и выделении импульса возбуждения. Волны объемной скорости получают путем интегрирования этого импульса. Пример импульсов возбуждения и волн объемной скорости, полученных методом обратной фильтрации, показан на рис. 5.3.

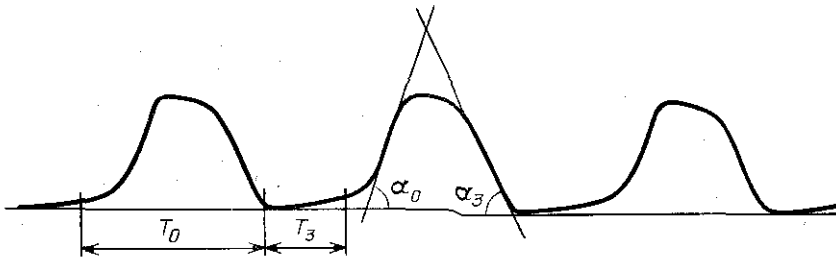


Рис. 5.2 Объемная скорость в речевом тракте, полученная методом безотражательной трубы

В отличие от предыдущих методов измерения, метод обратной фильтрации является косвенным. В нем имеется ряд погрешностей: предложение о независимости свойства источника возбуждения и акустических характеристик речевого тракта, эвристический подбор импеданса излучения, низкочастотные помехи и искажения при записи на магнитофон и т. д. Кроме того, сигнал-остаток, который обычно считается сигналом голосового возбуждения, при прослушивании дает достаточно разборчивую речь, что свидетельствует о сохранении формантных колебаний в этом сигнале, а следовательно, о его искажении относительно истинного сигнала возбуждения. Все же в силу своей доступности метод обратной фильтрации широко применяется для анализа голосового источника. Основное свойство импульсов голосового источника, которое наблюдают на сигналах после обратной фильтрации, состоит в том, что отрицательный импульс обычно имеет значительно большую амплитуду, чем положительный импульс, что соответствует более крутому спаду заднего фронта волны объемной скорости, чем нарастанию переднего фронта. Из рис 5.2 видно, что это не всегда так, поскольку в данном случае угол наклона касательной к наиболее крутому участку на переднем фронте больше, чем на заднем фронте. Трехмерная модель голосового источника, описанная в [59], также дает положи-

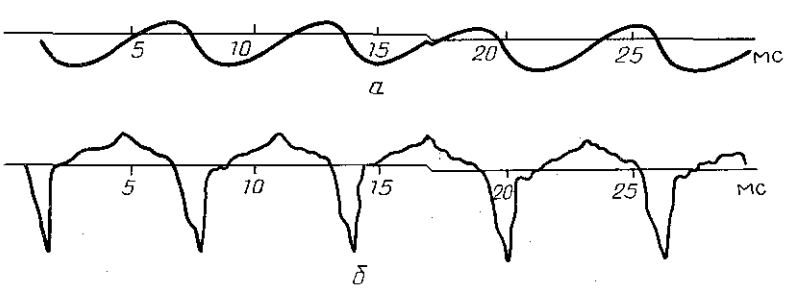


Рис. 5.3 Объемная скорость (*а*) и ее производная (*б*) воздушного потока в речевом тракте, полученные методом обратной фильтрации

тельные импульсы, сравнимые, или даже большие по амплитуде, чем отрицательные импульсы, тем не менее натуральность звучания такого источника в артикуляторном синтезаторе, учитывающем взаимодействие системы «источник-тракт», весьма высока. Большинство параметрических моделей голосового источника, однако, реализует случай преобладания амплитуды отрицательного импульса над амплитудой положительного импульса.

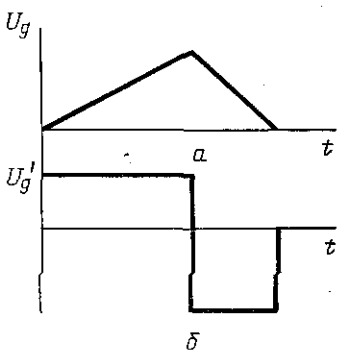


Рис. 5.4. Пилообразная аппроксимация импульса объемной скорости (*а*) и ее производной (*б*)

Простейшая параметрическая модель, аппроксимирующая волну объемной скорости пилообразным импульсом, уже не удовлетворяет требованиям натуральности синтетической речи, поскольку дает импульсы возбуждения, содержащие разрыв в производной (рис. 5.4). Часто применяется модель, предложенная в [95]. В этой модели передний фронт волны объемной скорости U_g описывается как

$$U_g(t)=\frac{U_{g_0}}{2}(1-\cos 2\pi F_0 t), \quad t<T,$$

а задний фронт, как

$$U_g(t)=U_{g_0}[k\cos 2\pi F_0(t-T)-k+1], \quad t\geqslant T,$$

где F_0 — частота, U_{g_0} — амплитуда волны, T — время от начала волны до ее максимума, $k>1$ определяет момент времени, в который отрицательная производная от волны объемной скорости достигает максимума по абсолютной величине. Дополнительно задается длительность интервала закрытой голосовой щели. Генерируемые этой моделью волны объемной скорости и сигнал голосового источника (т. е. производная от объемной скорости) показаны на рис. 5.5. Эта модель дает небольшой разрыв в непрерывности производной в момент перехода через нуль и, кроме того, большой разрыв в конце сигнала. Одна из модификаций параметрической модели, в которой такой разрыв ликвидирован, описана [70]. В этой модели аппроксимируется не волна объемной скорости, а сам сигнал голосового возбуждения (рис. 5.6). Восходящий участок положительного импульса представляется как

$$f(t)=A_1\sin\frac{\pi t}{2T_1}, \quad 0\leqslant t\leqslant T_1.$$

Ниспадающий участок сигнала описывается как

$$f(t)=(A_1+A_2)\cos\frac{\pi(t-T_1)}{4(T_2-T_1)}-A_2,\quad T_1<t\leqslant T_2.$$

Восходящий участок отрицательного импульса есть

$$f(t)=-A_2\left(\frac{T_3-t}{T_3-T_2}\right)^2,\quad T_2<t\leqslant T_3.$$

Имеется ряд других параметрических моделей, в том числе модель, используемая в серийных синтезаторах типа *DECTalk* [137]. В этой модели волна объемной скорости описывается как

$$U(t)=at^2-bt^3,$$

где a и b — параметры, зависящие от частоты основного тона и скважности импульсов (отношение длительности интервала открытой голосовой щели к периоду основного тона). Производная по времени есть перевернутая парабола.

Хотя параметрические модели голосового источника обеспечивают более или менее удовлетворительную натуральность синтетической речи, такая речь все еще безошибочно распознается как машинная, поскольку при таком подходе не моделируются многие эффекты, существенные для восприятия речи. В этих моделях трудно решить задачи создания индивидуальности голоса. В частности, женский голос звучит ненатурально. С другой стороны, эти модели практичны для использования в упрощенных синтезаторах, поскольку наряду с простотой реализации они обладают и некоторой гибкостью. Например, по заданной объемной скорости можно рассчитать момент возникновения и характеристики турбулентных шумов, сопут-

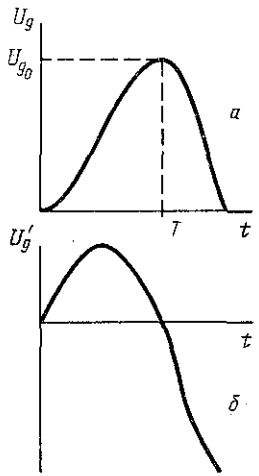


Рис. 5.5. Аппроксимация импульса объемной скорости (*а*) и ее производной (*б*)

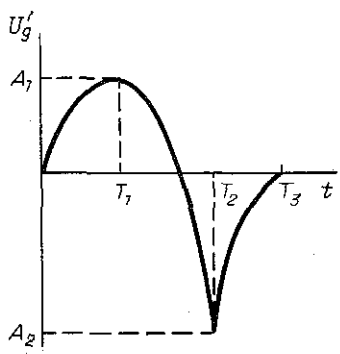


Рис. 5.6. Аппроксимация производной от объемной скорости голосового источника

ствующих голосовому источнику. Вариацией длительности интервалов, соответствующих открытой и закрытой голосовой щели, а также амплитуд положительного и отрицательного импульсов возбуждения можно имитировать микропросодические явления. При этом, однако, довольно трудно согласовать все изменения параметров так, чтобы они соответствовали соотношениям в естественной речи.

Поршневой источник также может быть представлен в виде параметрической модели, как это будет показано в § 5.6.

Поскольку в параметрических моделях не описывается площадь голосовой щели, то эффекты вариации частот и затуханий формант, хотя и могут быть отображены, но весьма приближенно.

Таким образом, параметрические модели могут использоваться для возбуждения формантных синтезаторов. Они обладают низкой вычислительной сложностью и предоставляют возможность имитации некоторых основных свойств голосового источника.

§ 5.3. Аэродинамика голосовой щели

Поскольку параметрические модели импульса голосового возбуждения не позволяют создать синтетическую речь с высокой натуральностью, приходится заниматься исследованием и математическим моделированием весьма сложных процессов голосообразования. В их число входит аэродинамика воздушного потока в голосовой щели, механика колебаний голосовых складок и влияние акустических колебаний на воздушный поток. В п. 5.4.1 этой главы мы увидим, что возбуждение акустических колебаний в речевом тракте осуществляется первой и третьей производной по времени от объемной скорости w на выходе голосового источника, $w=S(t)v(t)$, S — площадь поперечного сечения голосовой щели, v — скорость воздушного потока на выходе из голосовой щели. Поэтому для формирования источника голосового возбуждения нужно найти явный способ для вычисления скорости воздушного потока в зависимости от перепада давления и геометрических размеров голосовой щели.

Одномерное неустановившееся течение воздушного потока в канале с переменной площадью поперечного сечения $S(x, t)$ описывается нелинейным уравнением в частных производных:

$$\rho_0 \frac{\partial(vS)}{\partial t} + \rho_0 v \frac{\partial(vS)}{\partial x} + \frac{\bar{c}_x \rho_0}{2} v^2 S = -S \frac{\partial p}{\partial x}, \qquad (5.1)$$

где ρ_0 — плотность воздуха, v — скорость потока, p — давление в потоке, $\bar{c}_x$ — коэффициент динамического сопротивления.

Поскольку глубина голосовой щели $h_г$ мала по сравнению с длиной волны, соответствующей наинизшей частоте

основного тона F_0, т. е. $h_г \ll 330/F_0$, в (5.1) можно перейти к конечному интервалу h по оси x вместо бесконечно малого приращения dx. Тогда член в правой части примет следующий вид:

$$-S\frac{\partial p}{\partial x}=-S\frac{p(x+h)-p(x)}{h}=S\frac{\Delta p}{h},$$

где $\Delta p=p(x)-p(x+h)$, т. е. перепад давления на голосовой щели.

Воспользовавшись условием неразрывности потока по *пространственной координате* $w(x)=\text{const}$, *найдем*, что $\partial w/\partial x=\partial(vS)/\partial x=0$, так что для квазистационарного течения можно избавиться от второго члена в левой части (5.1). Добавляя член, соответствующий потерям на вязкое трение, который доминирует при малых раскрытиях голосовой щели, получим уравнение динамики воздушного потока в голосовой щели:

$$\rho_0 h_г \frac{d(vS_г)}{dt}+k_{тр}hvS_г+\frac{c_x\rho_0}{2}v^2S_г=\Delta p\,S_г, \qquad (5.2)$$

где коэффициент вязкого трения $k_{тр}$ определяется как для капиллярной трубки прямоугольного сечения с размерами $b_г$ и $h_г$:

$$k_{тр}=12\,\mu b_г^2/S_г^2,$$

а $c_x=\bar{c}_x b_г$ — коэффициент динамического сопротивления, который зависит от формы воздушного канала и числа Рейнольдса.

Обыкновенное дифференциальное уравнение (5.2) относится к типу уравнений Риккатти. В [59] было получено его решение в квадратурах в предположении постоянства площади голосовой щели, $S_г=\text{const}$:

$$v(t)=\frac{2h_г}{c_x}\left[g\,\text{th}\,(gt+\theta)-\frac{k_{тр}}{2\rho_0 h_г}\right], \qquad (5.3)$$

где

$$g=\frac{\Delta p\,c_x}{2\rho_0 h_г^2}+\left(\frac{k_{тр}}{2\rho_0 h_г}\right)^2,$$

$$\theta=\text{Arth}\left(\frac{\rho_0 c_x v_0+k_{тр}}{2\rho_0 h_г g}\right),$$

v_0 — начальная скорость воздушного потока в момент времени $t=0$.

При малых амплитудах раскрытия голосовой щели (меньше 0,1 см^2) и для частот, меньших 200 Гц, решение (5.3) довольно хорошо описывает динамику потока и при переменной во времени площади голосовой щели. Однако оно все же обладает погрешностью, возрастающей при больших амплитудах и частотах колебаний голосовых складок, и, к тому же, требует

сравнительно больших вычислительных затрат. Традиционные конечно-разностные методы (например, модифицированный метод Эйлера) имеют большую точность, но требуют и больше вычислений. Иногда предполагают, что тот или иной член в уравнении (5.2) доминирует, или задаются линейным законом изменения площади $S_г$ во времени. Это дает довольно простые решения, но их погрешность велика.

Ниже приводится метод решения уравнения (5.2), обладающий вычислительной простотой, и удовлетворяющий требованиям точности, необходимой для синтеза речи.

Перейдем в (5.2) к объемной скорости $w = vS_г$ и перенесем нелинейный член в правую часть:

$$\rho_0 h_г w' + k_{тр} w = F, \tag{5.4}$$

где

$$F = \Delta p S_г - \frac{c_x \rho_0}{2} \frac{w^2}{S_г},$$

а штрих означает производную по времени.

Игнорируя то обстоятельство, что функция возбуждения F содержит член, зависящий от переменной w, будем решать (5.4), как обыкновенное линейное дифференциальное уравнение с постоянными коэффициентами. Решение этого уравнения в конечных разностях, согласно § 2.4 есть

$$w(t+\Delta t) = w(t) + \alpha\left[\beta F(t+\Delta t) - w(t)\right], \tag{5.5}$$

где

$$\alpha = 1 - e^{-\Delta t/T}, \quad \beta = \frac{T}{\rho_0 h_г}, \quad T = \frac{\rho_0 h_г}{k_{тр}}.$$

Подставим теперь в F его значение, и получим квадратное алгебраическое уравнение относительно $w(t+\Delta t)$. Его решение есть

$$w(t+\Delta t) = \frac{\sqrt{1+4a_1 a_2} - 1}{2a_2}, \tag{5.6}$$

где

$$a_1 = w(t) + \alpha\left[\beta \Delta p(t+\Delta t) S_г(t+\Delta t) - w(t)\right],$$
$$a_2 = \frac{\alpha\beta c_x \rho_0}{2S_г(t+\Delta t)}.$$

Несмотря на то, что строго говоря, (5.5) служит решением только для уравнения с постоянными коэффициентами, решение (5.6) оказалось очень близко к решению модифицированным методом Эйлера, требуя в то же время значительно меньше вычислительных операций. Платой за эту простоту является невозможность вычисления за один шаг величины объемной скорости w в произвольный момент времени, поскольку (5.5) и (5.6) являются рекурсивными функциями, которые требуют начала вычислений от момента времени $t = 0$. В задачах синтеза речи, однако, это ограничение не играет роли, так

как все процессы рассматриваются развивающимися во времени от начальной точки $t=0$, причем выбор этой начальной точки, конечно, произволен.

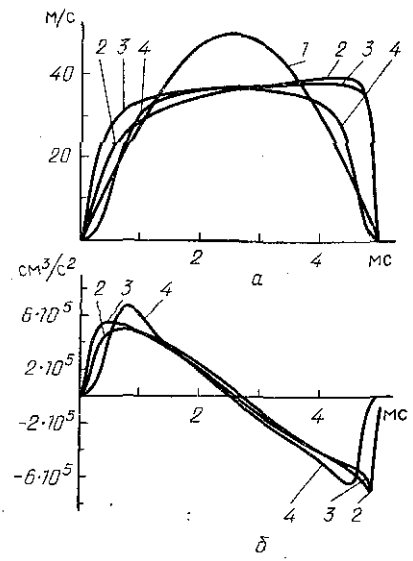


Рис. 5.7. а — скорость воздушного потока, б — производная от объемной скорости, 1 — площадь голосовой щели, 2 — решение с учетом концевой поправки, 3 — решение без учета концевой поправки, 4 — квазистатическое решение

На рис. 5.7 показано изменение скорости воздушного потока v в голосовой щели, площадь которой изменяется *по синусоидальному закону* с амплитудой в 0,2 см: $S_г(t)=0{,}2\sin 2\pi f t$, где $f=100$ Гц. Квазистатическое решение (5.4) имеет симметрическую форму, тогда как решение (5.6) обладает более медленным нарастанием и более крутым спадом, что проявляется, хотя и в меньшей степени, в форме производной от объемной скорости w'.

Если в воздушном канале имеется сужение, площадь сечения которого значительно меньше площади канала, то эффективная длина этого сужения $h_{эф}=h+\Delta h$ увеличивается вследствие сжатия потока в канале перед сужением и после него. Увеличение эффективной длины сужения зависит от многих факторов. В частности, в [95] приводится поправка, полученная Ингардом:

$$\Delta h=0{,}48\sqrt{S}\left[1-1{,}25\sqrt{\frac{S_г}{S}}\right],$$

где S — площадь канала и $S_г<0{,}16S$. В [37] дается более простая поправка

$$\Delta h=0{,}8\sqrt{S_г}.$$

При учете этой поправки ускорение потока при раскрытии голосовой щели уменьшается, тогда как ускорение при закрытой щели существенно не изменяется (рис. 5.7).

При скачкообразном раскрытии щели время переходного процесса для скорости потока v зависит от перепада давления Δp — чем меньше, тем длиннее переходный процесс (рис. 5.8). Для $\Delta p=1{,}5\cdot 10^3$ Па длительность переходного процесса составляет около 0,6 мс — величина, заметная по сравнению с реальной длительностью интервала открытия или закрытия голосовой щели. Вместе с тем, длительность переходного

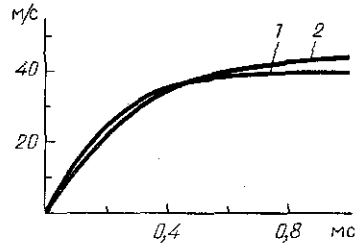


Рис. 5.8. Переходные процессы по скорости воздушного потока при скачке давления 10^3 Па. Площадь голосовой щели: *1* — 0,16 см^2, *2* — 0,32 см^2

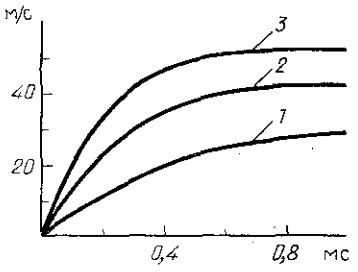


Рис. 5.9. Переходные процессы по скорости воздушного потока при скачке давления, площадь голосовой щели 0,32 см^2. Давление: *1* — $5 \cdot 10^3$ Па, *2* — 10^3 Па, *3* — $1{,}5 \cdot 10^3$ Па

процесса в меньшей степени зависит от площади раскрытия щели (рис. 5.9).

В связи с тем, что, как мы видели в предыдущем разделе, форма зависимости площади сечения голосовой щели от времени напоминает треугольник, интересно рассмотреть поведение воздушного потока и в этом случае. На рис. 5.10 видно, что объемная скорость, в общем, повторяет форму площади голосовой щели $S(t)$, однако наблюдается эффект, который в западной литературе называется «перекосом» (*skewing*). Этот эффект состоит в несимметрии импульса объемной скорости, причем нарастание потока происходит медленнее, чем его спад, а максимум объемной скорости смещается вперед по времени относительно максимума площади поперечного сечения голосовой щели. Степень этой несимметрии зависит также и от перепада давления Δp — чем он меньше, тем больше «перекос импульса» (рис. 5.11).

В уравнениях для воздушного потока в голосовой щели (5.2) имеется коэффициент c_x, отражающий влияние скорости потока на падение давления, т. е. коэффициент динамического сопротивления. Этот коэффициент состоит из двух величин: коэффициента падения давления перед щелью c_{x1} вследствие

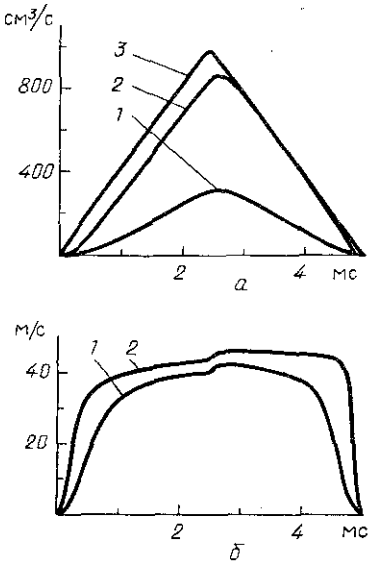


Рис. 5.10. Объемная (*а*) и линейная (*б*) скорость воздушного потока для площади голосовой щели, изменяющейся по линейному закону (*3*) с максимальной площадью: *1* — 0,16 см^2, *2* — 0,32 см^2. Давление 1000 Па

ускорения потока и коэффициента восстановления давления на выходе щели c_{x2} вследствие резкого расширения канала. Теоретическому и экспериментальному исследованию этих коэффициентов для голосовой щели был посвящен ряд работ [74, 76, 122]. В пионерской работе [74] на пластиковой модели гортани экспериментально было найдено, что $c_{x1} \approx 1{,}37$, а $c_{x2} = -0{,}5$. Дальнейшие исследования показали, что c_{x1} в этой работе преувеличен, и что теоретическое рассмотрение, приводящее к зависимости

$$c_{x2} = -\frac{S_г}{S}\left(1 - \frac{S_г}{S}\right),$$

находится в хорошем соответствии с экспериментальными данными. Здесь, как и раньше $S_г$ — площадь голосовой щели, S — площадь речевого тракта непосредственно над голосовой щелью. Поскольку $S \approx 3$ см^2, а максимальное значение $S_г$ обычно не превышает 0,3 см^2, то наибольшая величина коэффициента $c_{x2} \approx -0{,}2$. В [76] было уточнено, что коэффициент входного сопротивления c_{x1} зависит от формы голосовой щели и равен примерно 1,4 для чисел Рейнольдса, превышающих 5000. При меньших числах Рейнольдса c_{x1} находится в обратной линейной зависимости с Re в логарифмических координатах, т. е.

$$\ln c_{x1} = A - B \ln \mathrm{Re},$$

$$A = \ln 3, \quad B = \frac{\ln 2}{\ln 5000}.$$

Поскольку число Рейнольдса в голосовой щели обычно не превышает 5000—7000, то коэффициент c_{x1} зависит от скорости потока, хотя, как мы видим, эта скорость быстро достигает величин, близких к максимуму, так что c_{x1} близко к 1,4 большую часть времени. В результате изменения c_{x1} динамическое сопротивление оказывается большим для малых скоростей потока и меньшим — для больших. Форма импульса объемной скорости при этом становится более несимметричной.

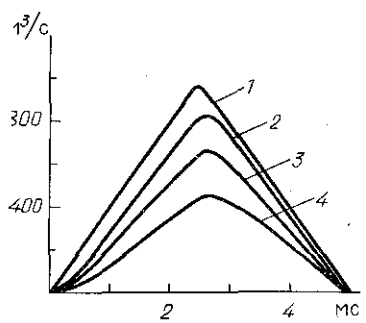


Рис. 5.11. «Перекос» изменения объемной скорости относительно изменения площади голосовой щели (*1*) с максимальным значением 0,32 см^2. Давление: *2* — 500 Па, *3* — 1000 Па, *4* — 1500 Па

На степень несимметрии импульса объемной скорости влияет также и масса воздуха в голосовой щели. До сих пор мы полагали, что площадь сечения голосовой щели не зависит от пространственной координаты вдоль оси речевого тракта. Это соответствует предположению, что все точки голосовых складок движутся синфазно. В действительности же наблюдается сдвиг по фазе движения

верхней и нижней поверхности голосовой складки (см. [59]), и поэтому голосовая щель вдоль координаты x может принимать различные формы. Если принять, что поверхность каждой голосовой складки представляет собой плоскость, наклоненную под различными углами к оси X, то можно приближенно оценить дополнительную массу воздуха, находящегося в голосовой щели в момент, когда нижние кромки уже разошлись,

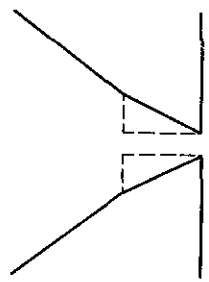


Рис. 5.12. Форма голосовой щели

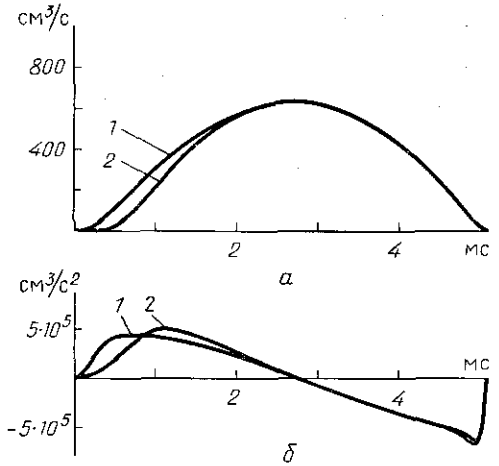


Рис. 5.13. Объемная скорость (*а*) и ее производная (*б*) без учета конуса (*1*) и с учетом конуса (*2*) голосовой щели

а верхние только начинают расходиться (см. рис. 5.12). Тогда масса воздуха в голосовой щели в этот момент не равна нулю, как в однородной модели, а есть $\rho_0 h S_0/2$, где S_0 — площадь на входе в голосовую щель. Полагая, что при дальнейшем движении голосовых складок эта дополнительная масса сохранится до тех пор, пока складки не начнут смыкаться, получим вместо первого члена уравнения (5.2), $\rho_0 h\,[(S_г + 0{,}5S_0)\,v\,]'$, где S_0 зависит от $S_г$ как $S_0 = A - S_г(t)$, $A = \text{const}$, $A > S_{г\max}$. Поэтому уравнение (5.2) переходит в уравнение

$$\rho_0 h\left(1+\frac{S_0}{2S_г}\right)w' + \left(k_{тр} - \rho_0 h\frac{S_0 S_г'}{2S_г^2}\right)w + \frac{\rho_0 c_x}{2S_г}w^2 = \Delta p\, S_г, \qquad (5.7)$$

которое решается так же, как и (5.2). Влияние дополнительной массы воздуха в голосовой щели приводит к увеличению несимметрии импульса объемной скорости, в результате чего максимум положительного импульса в производной от объемной скорости смещается вперед, и это больше соответствует тому, что можно наблюдать на сигналах, полученных методом обратной фильтрации (см. рис. 5.13), где изменение площади голосовой щели во времени было принято, как $S_г(t) = a\sin\omega t$. При этом отношение амплитуд отрицательного и положительного импульсов незначительно уменьшается.

Аэродинамическая модель является следующим шагом в направлении создания физически адекватного описания процессов голосообразования. Эта модель достаточно проста и в то же время позволяет учесть целый ряд эффектов, в том числе и взаимодействие с речевым трактом.

Изменения давления из-за акустических колебаний значительно больше сказываются на форме импульса объема скорости, чем на форме движения голосовых складок, поскольку вследствие инерционности складки не успевают отреагировать на быстрые колебания акустического давления. Это легко учесть путем добавления в первую часть уравнения (5.2) или (5.7) акустических компонент давления:

$$\Delta p = p_1 + p_{1a} - p_2 - p_{2a},$$

где p_1 и p_2 — медленно меняющиеся давления под и над голосовой щелью, p_{1a} и p_{2a} — акустическое давление на нижней и верхней кромках голосовой щели.

Аэродинамические свойства голосовой щели оказывают, пожалуй, большее влияние на характеристики голосового возбуждения, чем механика движения голосовых складок, и эти свойства достаточно хорошо описываются рассмотренной моделью. Для использования аэродинамической модели голосового источника в синтезаторах, в том числе и формантных, нужно задаться каким-либо законом изменения площади голосовой щели во времени. Этот закон можно сформировать так же, как это было сделано в параметрических моделях голосового источника для объемной скорости. В частности, это может быть несимметричный треугольник с регулируемой высотой и длиной каждой стороны. В [143] предлагается формула

$$S_{\text{г}}(t) = \begin{cases} 0{,}5 S_{\text{г max}} \left[1 - \cos \dfrac{\pi t}{t_1} \right], & 0 \leqslant t < t_1, \\ S_{\text{г max}} \cos \left[\dfrac{\pi (t - t_1)}{2(t_2 - t_1)} \right], & t_1 \leqslant t \leqslant t_2, \end{cases}$$

где $S_{\text{г max}}$ — амплитуда, t_1 — длительность интервала открытой голосовой щели, $t_2 - t_1$ — длительность интервала закрытия. В [200] предлагается другая форма

$$S(\theta) = \begin{cases} S_{\text{г max}} \left[\left(\dfrac{\theta}{\theta_m} \right)^{-\theta_m \operatorname{ctg} \theta_m} \dfrac{\sin \theta}{\sin \theta_m} \right]^{\beta}, & \theta \leqslant \pi, \\ 0, & \theta > \pi, \end{cases}$$

где

$$\theta = \frac{\pi t}{\gamma T}, \quad \theta_m = \frac{\pi \delta}{1 + \delta},$$

t — текущее время, T — период колебаний, γ — отношение дли-

тельности интервала открытой голосовой щели к периоду основного тона, δ — коэффициент симметрии импульса. Очевидно, можно предложить ряд простых законов для моделирования изменения площади голосовой щели во времени, которые совместно с аэродинамической моделью могут порождать импульсы голосового возбуждения, обеспечивающие довольно высокую натуральность звучания синтетической речи. Верно также и то, что такое увеличение натуральности формантного синтеза достигается весьма простыми средствами, без использования физически адекватных, но сложных моделей голосового источника, в которых учитываются не только аэродинамические и акустические колебания, но и упругие деформации тканей голосовых складок. Подвергая вариации параметры модели голосовой щели, можно получить не только интонационные контуры, но и создавать быстрые случайные изменения на каждом периоде основного тона (микровариации), а также имитировать различные типы голосов. Все же при таком подходе многие свойства голосового источника остаются не реализованными.

§ 5.4. Взаимодействие источника возбуждения и речевого тракта

Голосовой источник так же, как и остальные источники возбуждения, вообще говоря, неразделим с речевым трактом вследствие нелинейности процессов возбуждения и акустических колебаний. Однако исследование нелинейных уравнений в частных производных, описывающих колебания в речевом тракте, чрезвычайно трудно. Поэтому, огрубляя систему, рассматривают источники возбуждения и речевой тракт как раздельные, хотя и зависимые системы, и изучают их взаимное влияние. Говоря о взаимодействии источников возбуждения и речевого тракта, обычно подразумевают влияние акустических колебаний и артикуляторных движений на характеристики источника, но, как мы увидим ниже, эффекты взаимодействия этим не исчерпываются.

Синтезаторы, в которых используется взаимодействие источников возбуждения и речевого тракта, обладают лучшей натуральностью. Это отмечают многие исследователи, хотя количественная оценка улучшения натуральности приводится редко и сами эти оценки не всегда последовательны. Так, в [68] отмечается, что при использовании акустического влияния речевого тракта на источник, семь из восьми русских гласных в аудиторских экспериментах были приняты как более натуральные по сравнению с отсутствием взаимодействия, причем объективная оценка степени влияния акустических колебаний на источник и мнение аудиторов иногда расходились. Трудности в оценке натуральности связаны с тем, что при раздельном исследовании различных факторов взаимодействия

практически невозможно соблюсти естественные взаимоотношения между многими параметрами, тогда как система восприятия речи человеком чутко реагирует на нарушение этих взаимоотношений.

Приступая к рассмотрению таких аспектов взаимодействия источника и тракта, как зависимость амплитуды формант от частоты основного тона и конфигурации тракта, зависимость формы импульсов возбуждения от формантных частот, девиация частот и затуханий формант синхронно с колебаниями голосовых складок, внутренняя частота основного тона и другие эффекты, необходимо помнить о том, что эти явления в реальной речи взаимозависимы, и их суммарный эффект на восприятие речи не есть простая сумма частных эффектов.

5.4.1. Форма возбуждающего сигнала и амплитуда формант. Из общих соображений ясно, что чем медленнее спадает к высоким частотам спектр сигнала возбуждения, тем больше амплитуда верхних формант относительно первой форманты. Поскольку форма сигнала возбуждения меняется от импульса к импульсу, спектральное представление не очень удобно, и свойства источника возбуждения лучше рассматривать во временной области. При исследовании голосового источника физически наглядной переменной служит объемная скорость воздушного потока, но акустические колебания возбуждаются силой, которая пропорциональна ускорению, т. е. производной по времени от объемной скорости. Имеются и некоторые особенности в возбуждении акустических колебаний, которые выявляются при математическом анализе волнового уравнения для речевого тракта.

Если пренебречь влиянием подсвязочной области, то оказывается, что голосовой источник возбуждения находится в граничных условиях при $x=0$ и, таким образом, граничные условия неоднородны:

$$\left.\frac{\partial p}{\partial x}\right|_{x=0} = -\rho_0 \frac{\partial}{\partial t}\left(\frac{w}{S_т} + v_п\right) = f(t),$$

где w — объемная скорость воздушного потока в голосовой щели, $S_т$ — площадь речевого тракта непосредственно под голосовой щелью, $v_п$ — линейная скорость воздушного потока, создаваемая поршневым источником возбуждения. В [59] было показано, что избавляясь от этой неоднородности в граничных условиях, получаем вид действующего распределенного источника возбуждения:

$$F(x, t) = \left(\frac{1}{l} + \frac{1}{S}\frac{\partial S}{\partial x}\frac{x-l}{l}\right) f(t) - \frac{1}{c_0^2}\frac{(x-l)^2}{2l}\frac{\partial^2 f}{\partial t^2}, \qquad (5.8)$$

где l — длина речевого тракта, $S(x, t)$ — площадь поперечного сечения тракта.

Таким образом, акустические колебания в речевом тракте создаются источником, состоящим из трех компонент. Первая компонента пропорциональна первой производной от объемной скорости голосового источника и первой производной от линейной скорости поршневого источника. Вторая компонента, имея такую же зависимость от объемной и линейной скоростей источников, зависит еще и от формы речевого тракта; наконец, третья компонента зависит от третьей производной от объемной и линейной скоростей источников возбуждения и распределена вдоль тракта по параболическому закону. Наличие второй компоненты приводит к различию формы действующего возбуждения для разных конфигураций речевого тракта при одной и той же форме возбуждающего импульса, создаваемого голосовым и поршневым источниками. Третья компонента подчеркивает неоднородность в форме импульса источников. В результате действия всех компонент источник возбуждения содержит до четырех и более импульсов, и длительность интервала времени, на котором возбуждение равно нулю, обычно также равно нулю. Это означает, что в речевом тракте источник голосового возбуждения действует непрерывно (для звонких звуков) и период свободных колебаний практически отсутствует.

Из (5.8) можно определить амплитуду сигнала $f_k(t)$, возбуждающего колебания на частоте k-го резонанса:

$$A_k(t)=\frac{2\int\limits_0^l F(x,\,t)\,S(x,\,t)\,\psi_k(x)\,dx}{\rho_0 l\int\limits_0^l S(x,\,t)\,\psi_k^2(x)\,dx},$$

где $S(x,\,t)$ — площадь поперечного сечения речевого тракта, $\psi_k(x)$ — собственная функция, т. е. распределение акустического давления вдоль речевого тракта. Как видно, амплитуда $A_k(t)$ зависит от формы речевого тракта и формы собственных функций, а поскольку частота k-го резонанса также является функционалом от площади поперечного сечения речевого тракта, то в конечном счете $A_k(t)$ зависит от частоты k-го резонанса. Такая зависимость может быть приближенно вычислена, так что даже для формантных синтезаторов, не использующих информацию о площади поперечного сечения, можно учесть эффект зависимости амплитуды возбуждения от частоты форманты. Для артикуляторно-формантных синтезаторов $A_k(t)$ вычисляется непосредственно по $S(x,\,t)$ и $\psi_k(x,\,t)$. В артикуляторных синтезаторах, где голосовой источник является нераздельной частью речевого тракта, зависимость $A_k(S,\,\psi_k)$ реализуется автоматически.

Амплитуда акустических колебаний зависит и от частоты возбуждения. По известному свойству резонаторов в случае периодического возбуждения, например, в виде $A_k \sin \omega_0 t$

амплитуда колебаний

$$B_k = \frac{A_k}{\sqrt{(\omega_k^2 - \omega_0^2)^2 + 4g_k\omega_k^2}},$$

где ω_k — частота k-го резонанса, g_k — затухание на этой частоте. Если частота возбуждающего сигнала ω_0 зависит от времени, то амплитуда колебаний B_k изменяется, достигая максимума при $\omega_k = \omega_0$. Если же сигнал возбуждения имеет широкий спектр, то максимум B_k достигается при совпадении одной из гармоник частоты основного тона с частотой резонанса, $\omega_k = n\omega_0$. Пусть частота основного тона ω_0 меняется линейно во времени, т. е. $\omega_0(t) = \omega_0 + at$. Тогда амплитуда B_k будет подвергаться периодическому изменению по мере того, как очередная гармоника приближается и удаляется от ω_k. При этом частота таких колебаний амплитуды k-го резонанса зависит от его частоты ω_k, поскольку скорость изменения частоты n-й гармоники основного тона растет вместе с n: $\omega_n = n(\omega_0 + at) = n\omega_0 + nat$. Вследствие этого явления на сонограммах реального речевого сигнала при изменении частоты основного тона наблюдаются причудливые узоры, затрудняющие анализ формантных частот.

Поскольку сигнал возбуждения голосового источника содержит ряд положительных и отрицательных импульсов, то амплитуды колебаний на некоторой резонансной частоте зависит не только от периода следования сигналов возбуждения, но и от интервалов между импульсами, составляющими сигнал голосового возбуждения, а также от изменения этих интервалов во времени. Например, если интервал перед положительными и отрицательными импульсами равен T_1, то амплитуда акустических колебаний на частоте, пропорциональной $2\pi/T_1$, увеличится, так как импульсы возбуждения действуют в фазе с колебаниями, а на частоте, пропорциональной $3\pi/T_1$, амплитуда уменьшится, так как один из импульсов действует в противофазе с акустическими колебаниями.

Эффекты, связанные со сложной структурой сигнала возбуждения, существенны для восприятия синтетического сигнала. Это проявляется и в заметном улучшении качества вокодерной речи при использовании метода многоимпульсного возбуждения. Следовательно, нельзя ожидать высокой натуральности от синтезатора с чересчур простым источником голосового возбуждения. Имеются синтезаторы с однополярными импульсами, грубо соответствующими волнами объемных скоростей, а производные от этих импульсов не используются. Нечего и говорить о натуральности звучания таких синтезаторов.

5.4.2. Вариации формантных частот и затуханий. В однородной трубке длиной L, закрытой с одного конца и открытой с другого, резонансные частоты есть $\omega_k = \pi c_0(k - 0{,}5)/L$. Если посередине этой трубки поместить тонкую стенку, то резонанс-

ные частоты в оставшейся половине с открытым концом удвоятся, потому что длина трубки уменьшилась вдвое: $\omega_k^* = 2\pi c_0 (k-0{,}5)/L$. Если в перемычке имеется отверстие, то при увеличении его площади от нуля до площади трубки резонансные частоты ω_k стремятся к ω_k^*; переход этот, однако, не монотонен. При малых отношениях площади отверстия в перемычке к площади трубки ω_k^* сначала возрастает и лишь затем падает по мере расширения отверстия. Это явление хорошо известно в акустике [50]. Основываясь на этом явлении, в [58] было показано, что при раскрытии голосовой щели затухание резонансных колебаний речевого тракта увеличивается, а частоты резонансов либо возрастают, либо падают — в зависимости от соотношения импедансов подсвязочной области, голосовой щели и речевого тракта. Влияние раскрытия голосовой щели возрастает при уменьшении формантной частоты. Измерение вариаций частот и затухания формант в реальной речи затруднено из-за быстроты протекаемых процессов. Однако в [195] например, показано, что при открытой голосовой щели без изменения площади и без воздушного потока, частота первой форманты может возрасти на 200 Гц по сравнению с закрытой голосовой щелью. С помощью специально разработанных алгоритмов анализа речи иногда удается наблюдать повышение частоты первой форманты на 25%. Таким образом, и теоретический анализ и экспериментальные данные позволяют сделать вывод об изменении частот затуханий формант синхронно с колебаниями голосовых складок.

Увеличение затухания при раскрытии голосовой щели физически объясняется стоком акустической энергии в подсвязочную область. Если бы действовал только один фактор излучения в подсвязочную область, то потери возрастали бы пропорционально квадрату частоты. Однако в действительности имеются и потери на податливость стенок, возрастающие с уменьшением частоты; и потери на капиллярное трение, не зависящие от частоты. Для расчета этих явлений требуются достаточно подробные модели, о которых мы будем говорить в одном из последующих разделов. Пока же отметим, что приблизительно можно считать, что дополнительное затухание примерно пропорционально площади голосовой щели: $\delta_k = a(\omega_k) S_г$.

Следует упомянуть о том, что имеются противоречивые мнения о влиянии вариаций затухания формант на восприятие качества синтетической речи. Так, в [117] этот фактор указывается как один из основных, а в [162] делается вывод о малом влиянии вариаций затухания на улучшение качества синтетической речи. Опыт моей работы с артикуляторно-формантным синтезатором свидетельствует о несомненном повышении натуральности синтетической речи, если вариациям подвергается и частота, и затухание формант.

В этом синтезаторе использовались грубые модели, где принималось, что величина вариаций пропорциональна площади голосовой щели и линейно падает с ростом частоты. По-видимому, здесь мы имеем дело с комплексным фактором и для того, чтобы восприятие человека отметило улучшение качества синтетического сигнала, необходимо соблюсти определенные соотношения между вариациями затухания и частоты форманты.

Увеличение затухания на интервале открытой голосовой щели приводит не только к нарушению экспоненциального характера огибающей акустических колебаний, но и к уменьшению амплитуды, иногда весьма значительному. Вариации затухания и частот формант создают эффект смешанной амплитудно-частотной модуляции, к которому добавляется и фазовая модуляция как результат изменения импеданса голосовой щели. Величина этих модуляций достаточна велика для того, чтобы система восприятия человека их заметила.

Последние исследования по объективной оценке качества речевого сигнала в телефонных системах и архитектурной акустике указывают на важную роль этих модуляций. Поэтому не вызывает сомнения необходимость использования вариаций формантных параметров в синтезаторах речи. Вопрос заключается лишь в способе расчета этих вариаций. В чисто артикуляторном синтезаторе эти явления реализуются автоматически, тогда как в формантном и артикуляторно-формантном синтезаторах их нужно рассчитывать с помощью тех или иных моделей, к рассмотрению которых мы приступим в следующем разделе.

Имеется еще один эффект, который трудно реализовать в каком-либо синтезаторе, помимо артикуляторного. Начиная с некоторой площади голосовой щели, подсвязочная область и речевой тракт образуют единую акустическую систему. Длина этой системы примерно вдвое больше, чем длина речевого тракта, поэтому в пределе (при площади голосовой щели, равной площади тракта) в диапазоне частот примерно до 5 кГц существует вдвое больше резонансных частот, чем при $S_г = 0$. В [59] было показано, что при раскрытии голосовой щели создаются условия для развития дополнительных формант, затухание и частота которых в противоположность основным формантам, падает с увеличением $S_г$. В результате этого на интервале открытой голосовой щели в речевом сигнале появляются спектральные компоненты, которые отсутствуют при закрытой голосовой щели. Эти компоненты несут информацию о подсвязочной области. По-видимому, они влияют на тембральные характеристики голоса и создают индивидуальные особенности звучания в зависимости от геометрических размеров легких, бронхов и трахеи. Моделирование этого эффекта очень сложно и пока не известно работ по экспериментальной оценке его роли в качестве синтетической речи.

5.4.3. Акустическое взаимодействие. Акустические колебания в речевом тракте создают переменную компоненту давления на входе и выходе голосовой щели и, таким образом, влияют на объемную скорость воздуха. Это влияние тем больше, чем ниже частота резонансных колебаний в тракте, поэтому достаточно учесть два низших резонанса. Это относится к акустическим колебаниям и в подсвязочной области. Эквивалентную электрическую схему для расчета акустического взаимодействия обычно представляют в виде пары последовательно соединенных резонансных контуров в подсвязочной и надсвязочной областях (см. рис. 5.14). Частоты первых двух резонансов подсвязочной

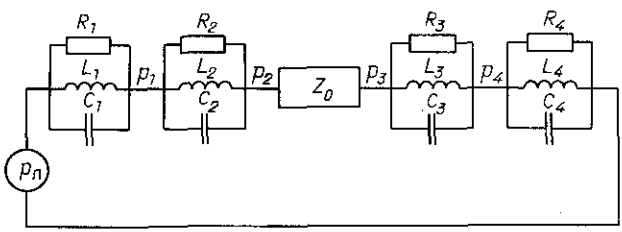


Рис. 5.14. Электрическая схема акустического взаимодействия голосового источника и речевого тракта

области в [125] были оценены как $F_{1п}=640$ Гц и $F_{2п}=1400$ Гц, их добротность $Q_{1п}=2{,}5$ и $Q_{2п}=8{,}7$, а сопротивление $R_{1п}=36{,}7$ и $R_{2п}=53{,}3$ акустических Ом. Из $Q=R/2\pi FL$ и $4\pi^2F^2LC=1$ можно найти электрические элементы: индуктивность L и емкость C. Поскольку вторая резонансная частота подсвязочной области довольно высока, то можно ограничиться лишь одним — первым резонансом. Если частота второй форманты в речевом тракте превышает 1000 Гц, то и для речевого тракта можно ограничиться лишь первой формантой. Конечно, для звуков типа /О/ и /У/ необходимо использовать две форманты, поскольку у них частота второй форманты низкая. Уменьшая число резонансов в схеме акустического взаимодействия, мы получаем более грубое описание эффектов, но при этом снижаем сложность вычисления.

Рассмотрим простейший случай, когда для подсвязочной области и речевого тракта используется лишь по одному резонансу. Тогда имеем следующую систему уравнений:

$$\begin{gathered}\rho_0 h w' + k_{тр} w + \frac{\rho_0 c_x}{2S_г} w^2 = (p_2 - p_3) S_г, \\ c_{1п}(p_л - p_2)'' + \frac{1}{R_{1п}}(p_л - p_2)' + \frac{1}{L_{1п}}(p_л - p_2) = w', \\ C_{1т} p_3'' + \frac{p_3'}{R_{1т}} + \frac{p_3}{L_{1т}} = w'.\end{gathered} \tag{5.9}$$

Первое уравнение в этой системе есть уравнение потока

в голосовой щели; оно совпадает с (5.2). $R_{1п}$, $L_{1п}$ и $C_{1п}$ — параметры электрического аналога первого резонанса подсвязочной области, $R_{1т}$, $L_{1т}$ и $C_{1т}$ — параметры электрического аналога первого резонанса речевого тракта. Площадь голосовой щели $S_г(t)$ выступает в качестве независимого переменного параметра. Если в каждой области учитывается по два резонанса, то в системе (5.9) добавляется два уравнения второго порядка.

Поскольку система (5.9) содержит нелинейное уравнение, то для ее решения можно пользоваться либо численными, либо приближенными методами. Для решения методом Рунге — Кутта каждое из уравнений этой системы нужно представить в виде $y'=f(y, t)$, что дает

$$w'=\frac{1}{\rho_0 h}\left[S_г(p_2-p_3)-k_{тр}w-\frac{\rho_0 c_x}{2S_г}w^2\right],$$

$$w'_{L1}=\frac{p_2}{L_{1п}},$$

$$(p_л-p_2)'=w-\frac{1}{C_{1п}}\left(\frac{p_л-p_2}{R_{1п}}+w_{L1}\right),$$

$$w'_{L2}=\frac{p_3}{L_{2п}},$$

$$p'_3=w-\frac{1}{C_{1т}}\left(\frac{p_3}{R_{2т}}+w_{L2}\right).$$

Метод Рунге — Кутта хорошо известен, для него имеются стандартные программы. Поэтому реализация его на микропроцессорах не вызывает затруднений. Обычно используют метод Рунге — Кутта четвертого порядка, поскольку он обеспечивает приемлемую точность и не слишком сложен. Более подробно свойства этого метода будут рассмотрены ниже.

Если от производных перейти к конечным разностям, то система (5.9) превратится в систему алгебраических уравнений. Для решения нелинейных алгебраических систем используется множество методов, различающихся по точности, сложности и скорости. Ниже будет описан итеративный метод, который полностью применим к системе (5.9). Достоинством этого метода является простота и быстрая сходимость (при некоторых условиях).

Преимущество численных методов, как известно, состоит в возможности решения сложных систем уравнений, недостаток же этих методов заключается в трудности оценки роли тех или иных факторов без проведения трудоемких расчетов. Поэтому стараются использовать приближенные аналитические методы, дающие решения, которые можно анализировать качественно, не прибегая к численным расчетам. Если в системе (5.9) пренебречь уравнением, описывающим акустические ко-

лебания в подсвязочной области, то к ней можно применить метод малого параметра. В этом случае давление в легких равно подсвязочному давлению и систему (5.9) можно представить в каноническом виде:

$$p_3'' + \omega_1^2 p_3 = \varepsilon f(p_3, p_3'),$$
$$f(p_3, p_3') = \frac{R_{1\text{т}} w' - p_3'}{R_{1\text{т}} C_{1\text{т}}}, \tag{5.10}$$

а для описания w воспользуемся аналитическим решением (5.3).

При $\varepsilon = 0$ решение (5.10), очевидно, есть

$$p_3 = a\cos\psi,$$
$$\psi = \omega_1 t + \theta_0.$$

Общее решение (5.10) будем искать в виде

$$p_3 = a\cos\psi + \varepsilon\varphi_1(a, \psi) + \varepsilon^2\varphi_2(a, \psi) + \dots,$$

где $\varphi_i(a, \psi)$ — периодические функции, а величины $a(t)$ и $\psi(t)$ определяются из дифференциальных уравнений:

$$da/dt = \varepsilon A_1(a) + \varepsilon^2 A_2(a) + \dots,$$
$$d\psi/dt = \omega_1 + \varepsilon B_1(a) + \varepsilon^2 B_2(a) + \dots \tag{5.11}$$

Для параметров первого приближения, по [4],

$$A_1 = -\frac{1}{2\pi\omega_1}\int_0^{2\pi} f(a\cos\psi,\ -a\omega_1\sin\psi)\sin\psi\, d\psi,$$

$$B_1 = -\frac{1}{2\pi a\omega_1}\int_0^{2\pi} f(a\cos\psi,\ -a\omega_1\sin\psi)\cos\psi\, d\psi,$$

где в $f(p_3, p_3')$ вместо p_3 и p_3' подставляются $a\cos\psi$ и $-a\omega_1\sin\psi$. Функцию $\varphi_1(a, \psi)$ найдем как

$$\varphi_1 = \frac{g_0(a)}{\omega_1^2} - \frac{1}{\omega_1^2}\sum_{n=2}^{\infty}\frac{g_n(a)\cos n\psi + h_n(a)\sin\psi}{n^2 - 1},$$

где

$$g_n(a) = \frac{1}{2\pi}\int_0^{2\pi} f(a\cos\psi,\ -a\omega_1\sin\psi)\cos n\psi\, d\psi,$$

$$h_n(a) = \frac{1}{2\pi}\int_0^{2\pi} f(a\cos\psi,\ -a\omega_1\sin\psi)\sin n\psi\, d\psi,$$

т. е. $g_n(a)$ и $h_n(a)$ являются коэффициентами разложения функции $f(p_3, p_3')$ в ряд Фурье. Из (5.11) видно, что если $\varepsilon > 0$ и $B_1(a) \neq 0$, то частота первого резонанса ω_1 подвергается

девиациям, а поскольку B_1 зависит от площади голосовой щели $S_г(t)$, то эти девиации синхронны с колебаниями $S_г(t)$.

Метод малого параметра применяется в [138] для изучения акустического взаимодействия источника возбуждения с речевым трактом, причем для аналитического решения дифференциального уравнения для воздушного потока в голосовой щели сопротивлением трения пренебрегали, хотя это дает большие ошибки при $S_г \to 0$. Все же результаты этой работы в общем соответствуют данным других исследователей.

При учете влияния акустических колебаний форма импульса объемной скорости голосового источника становится более несимметричной: передний фронт становится более пологим, а задний более крутым. Кроме того, акустические колебания накладывают на форму импульса объемной скорости так называемые «складки» (*ripples*). Обычно эти складки появляются на переднем фронте. Небольшим изменениям подвергаются частота основного тона, амплитуда и отношение длительности и переднего фронта к длительности заднего фронта импульса объемной скорости.

Как отмечается в работах по аудиторской экспертизе, акустическое взаимодействие улучшает натуральность звучания синтетической речи [68]. Сложность реализации акустического взаимодействия, однако, сравнима или даже превышает сложность реализации формантного синтезатора в его традиционном исполнении.

5.4.4. Взаимодействие по постоянному току. Термин «постоянный ток» не совсем удачен в применении к процессам в речевом тракте, где физические параметры — давление и объемная скорость, все время колеблются с той или иной частотой. Этот термин будет использоваться при исследовании тех процессов, при которых речевой тракт является системой с сосредоточенными параметрами, т. е. процессов, преимущественно связанных с протеканием воздушного потока. Сравнительно медленное изменение давления в легких, в трахее и в речевом тракте при колебаниях голосовых складок оказывает существенное влияние на условия самовозбуждения голосового источника и форму импульсов возбуждения. Характеристики

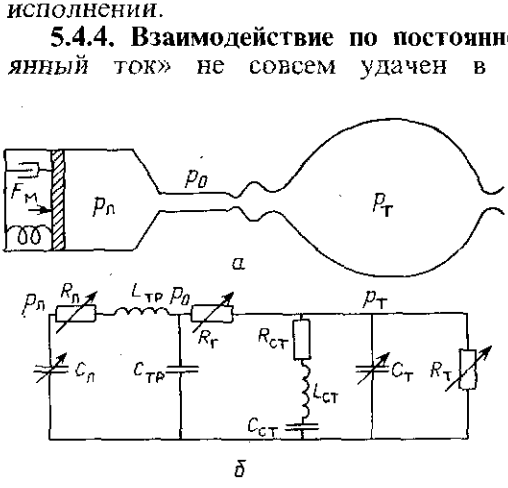


Рис. 5.15. Аэродинамическая (*а*) и электрическая (*б*) схемы речевого тракта

импульсного и турбулентного возбуждения также зависят от условий протекания воздушного потока.

Рассмотрим речевой тракт как аэродинамическую систему (рис. 5.15). В этой системе объем легких меняется вследствие опускания грудной диафрагмы, представленной на рисунке в виде поршня. Движение поршня зависит от разности силы, создаваемой напряжением мышц диафрагмы и давлением воздуха в легких. Кроме того, имеются элементы упругого и вязкого сопротивления. Уравнение смещения поршня запишем как

$$m_{\text{п}}x'' + r_{\text{п}}x' + k_{\text{п}}x = F_{\text{м}} - p_{\text{л}}S_{\text{д}}, \tag{5.12}$$

где x — смещение поршня, $m_{\text{п}}$ — масса, $r_{\text{п}}$ — коэффициент вязкости трения, $k_{\text{п}}$ — коэффициент жесткости, $F_{\text{м}}$ — сила сокращающихся мышц, $p_{\text{л}}$ — давление в легких, $S_{\text{д}}$ — площадь грудной диафрагмы.

В свою очередь, мышечная сила $F_{\text{м}}$ развивается не мгновенно, а является результатом решения параметрического уравнения

$$F''_{\text{м}} + g_{\text{м}}(t)F'_{\text{м}} + \lambda_k\left[E_0 + E_1(t)\right]F_{\text{м}} = 0,$$

где E_0 и E_1 — постоянная и переменная компоненты модуля упругости.

В данном случае не обязательно следовать строгой модели, поскольку зависимость $F_{\text{м}}$ от времени не очень существенна. Потому в качестве порождающей модели можно принять приближенное уравнение с постоянными коэффициентами:

$$F''_{\text{м}} + 2g_{\text{м}}F'_{\text{м}} + \omega_{\text{м}}^2 F_{\text{м}} = A,$$

где ступенчатое возбуждение

$$A = \begin{cases} 0, & t \leqslant 0, \\ A, & t > 0. \end{cases}$$

Собственная частота $f_{\text{м}} = \omega_{\text{м}}/2\pi$ примерно равна 5 Гц. Перепишем теперь (5.12) в следующем виде:

$$x'' + 2g_{\text{п}}x' + \omega_{\text{п}}^2 x = (p_{\text{м}} - p_{\text{л}})S_{\text{д}}/m_{\text{п}},$$

где $p_{\text{м}} = F_{\text{м}}/S_{\text{д}}$. Собственную частоту $f_{\text{п}} = \omega_{\text{п}}/2\pi$ примем близкой к 10 Гц, площадь диафрагмы $S_{\text{д}} \approx 600$ см^2, сопротивление $r_{\text{п}} \approx 1000$ г/(с·см^2), массу на единицу площади $m_{\text{п}}/S_{\text{д}} \approx 1{,}5$ г/см^2 (по [124]).

Скорость изменения объема легких есть

$$V'_{\text{л}} = S_{\text{д}}x'.$$

При отсутствии сопротивления воздушному потоку, вытекающему из легких, его объемная скорость равна $-V'$, так как размерность этого параметра см^3/с. Переменному объему легких в электрическом аналоге соответствует переменная

емкость

$$C_{л}(t)=\frac{V_{л}(t)}{\rho_0 c_0^2}=\frac{V_{0л}-V'_{л}t}{\rho_0 c_0^2},$$

где $V_{0л}$ — начальный объем легких. Очевидно, что объем легких не может уменьшаться до нуля и, таким образом, максимальное время $t_{м}$ от вздоха до вздоха в процессе речеобразования при постоянной скорости изменения объема $V'_{л}$ есть

$$t_{м}=(V_{0л}-V_{tл})/V'_{л},$$

где $V_{tл}$ — наименьший допустимый объем легких, при достижении которого в синтетической речи должна быть сделана пауза, по длительности соответствующая вдоху.

Дальнейшие процессы в речевом тракте удобно описывать с помощью электрического сигнала (рис. 5.15, *б*). Рассматривая легкие и трахею как резонатор Гельмгольца, запишем его уравнение:

$$\begin{gathered} w'_{л}L_{тр}+w_{л}(R_{л0}+k_{л}w_{л})=p_{л}-p_0, \\ (p_{л}C_{л})'=w_{л}. \end{gathered} \tag{5.13}$$

Здесь $w_{л}$ — поток, вытекающий из легких, $L_{тр}$ — индуктивность трахеи,

$$L_{тр}=\rho_0 l_{тр}/S_{тр},$$

$l_{тр}$ — длина трахеи (12—20 см), $S_{тр}$ — площадь трахеи (2—4 см^2), ρ_0 — плотность воздуха ($1{,}14\cdot 10^{-3}$ г/см^3).

Давление на выходе из трахеи, т. е. непосредственно под голосовой щелью, обозначается как p_0. В сопротивлении легких учитывается два вида сопротивления в бронхиолах и альвеолах: капиллярное трение $R_{л0}$ и динамическое сопротивление с коэффициентом $k_{л}$. Согласно данным [159], $R_{л0}\approx 3$, а коэффициент $k_{л}$ примем равным 0,02 — 0,05. Это означает, что при потоке $w_1=600$ см^3/с полное сопротивление легких для $k_{л}=0{,}02$ равно 15, а при $k_{л}=0{,}05$ сопротивление равно 33, тогда как среднее сопротивление на одну треть меньше.

Если голосовая щель закрыта, или в речевом тракте имеется смычка, то ток из легких не вытекает, $w_{л}=0$, и в системе (5.13) остается лишь последнее уравнение. Его решение

$$p_{л}(t)=p_{0л}\exp\left\{-\frac{C'_{л}}{C_{л}}t\right\},$$

где $p_{0л}$ — давление в легких перед началом сжатия, т. е. атмосферное давление. Поскольку $C'_{л}<0$, то давление в легких $p_{л}(t)$ растет до того момента, пока сила мышечного сокращения и сила легочного давления не станут равны.

Механическая податливость стенок трахеи представляется в электрическом аналоге в виде емкости

$$C_{тр} = h_{тр} S_{тр} / E_{тр},$$

где $h_{тр}$ — толщина стенок трахеи, $S_{тр}$ — площадь поверхности стенок, $E_{тр}$ — модуль упругости тканей. Принимая $h_{тр} = 1$ см, $S_{тр} = 100$ см^2, $E_{тр} = 10^6$ г/(см·с^2), получим, что емкость трахеи $C_{тр} \approx 10^{-4}$.

Ток через сопротивление голосовой щели, т. е. объемная скорость воздушного потока $w_г$ определяется из решения уравнения (5.2) или более точного уравнения (5.7). Податливость стенок речевого тракта создает дополнительный поток. Смещение стенок описывается уравнением

$$m_{ст} \ddot{x}''_{ст} + r_{ст} x'_{ст} + k_{ст} x_{ст} = p_т,$$

где $m_{ст}$ — поверхностная масса, $r_{ст}$ — коэффициент вязкого сопротивления, $k_{ст}$ — упругость, $p_т$ — давление в речевом тракте. По оценкам [124], $m_{ст} = 1{,}5-2$ г/см^2, $r_{ст} = 800-1100$ г/(с·см^2), $k_{ст} = (8-13)\cdot 10^4$ г/(см·с^2), так что резонансные частоты лежат в диапазоне 30—70 Гц. Такие низкие резонансные частоты означают большое взаимное влияние колебаний воздушного потока в тракте и колебаний стенок. Объемная скорость $w_{ст}$, создаваемая колебаниями стенок, есть $w_{ст} = S_{ст} x'_{ст}$, где $S_{ст}$ — площадь поверхности стенок в речевом тракте.

Объем воздуха в речевом тракте имеет упругость,

$$C_т = V_т / (\rho_0 c_0^2),$$

где $C_т$ зависит от времени, поскольку объем тракта $V_т$ меняется в зависимости от артикулируемого звука. При объеме $V_т = 100$ см^3 (при длине тракта $l_т = 20$ см и средней площади $S_т = 5$ см^2) емкость $C_т \approx 6\cdot 10^{-5}$ см^{4}с^2/г. Ток, протекающий через эту емкость, есть $w_{тв} = p_т C_т$. Переменная емкость речевого тракта играет большую роль в поддержании колебаний голосовых складок во время звонкой смычки, когда нарастание давления в речевом тракте частично компенсируется увеличением его объема.

Активное сопротивление в речевом тракте складывается из сопротивления вязкого трения $R_{тт}$ и динамического сопротивления $R_{тд}$:

$$R_т = R_{тт} + R_{тд} = \int_0^l \left[\frac{\rho_0}{2} \frac{c_x(x) v_т(x)}{S_т(x)} + r_т \right] dx,$$

где $v_т(x)$ — скорость воздушного потока, $r_т$ — погонный коэффициент вязкого трения. Обозначая давление на верхней поверхности голосовых складок $p_т$, из $p_т = R_т w_т$ найдем распределение установившейся скорости потока вдоль речевого

тракта [59]:

$$v_{\text{т}}(x)=\frac{p_{\text{т}}}{S_{\text{т}}(x)\int\limits_0^{l_{\text{т}}}\left[\frac{\rho_0}{2}\frac{c_x(x)v_{\text{т}}(x)}{S_{\text{т}}(x)}+r_{\text{т}}\right]dx}. \qquad (5.14)$$

При решении (5.14) итеративным способом для обеспечения сходимости нужно перейти к безразмерной величине $\bar{v}_{\text{т}}=v_{\text{т}}/v_{\text{т max}}$, где $v_{\text{т max}}\approx 5\cdot 10^3$ см/с. Коэффициент динамического сопротивления c_x зависит от того, сужается или расширяется речевой тракт в точке x. Приближенно можно записать:

$$c_x=\begin{cases} c_{x1}+c_{x2}, & S_{\text{т}}(x+\Delta x)>S_{\text{т}}(x),\\ c_{x1}, & S_{\text{т}}(x+\Delta x)\leqslant S_{\text{т}}(x),\end{cases}$$

где

$$c_{x1}=0{,}05;\ c_{x2}=M\left[\frac{S_{\text{т}}(x+\Delta x)}{S_{\text{т}}(x)}-1\right]^2;\quad M=\frac{v_{\text{т}}(x)}{c_0}.$$

При не слишком малых площадях поперечного сечения сопротивление трения зависит от частоты ω:

$$r_{\text{т}}(\omega)=\sqrt{\frac{\rho_0\mu\omega}{2}}\int\limits_0^{l_{\text{т}}}L(x)\,dx,$$

где L — периметр поперечного сечения. Полное сопротивление в полосе частот, например от 0 до 5 кГц, есть

$$r_{\text{т}}=\int\limits_0^{\Omega}\sqrt{\frac{\rho_0\mu\omega}{2}}\,d\omega\int\limits_0^{l_{\text{т}}}L(x)\,dx=\frac{2}{3}\left(\frac{\rho_0\mu\omega}{2}\right)^{3/2}\int\limits_0^{l_{\text{т}}}L(x)\,dx=0{,}0014\int\limits_0^{l_{\text{т}}}L(x)\,dx.$$

Принимая средний периметр равным $L=2\pi\sqrt{5/\pi}$ (т. е. для площади $S_{\text{т}}=5$ см^2) и длину речевого тракта $l_{\text{т}}=17{,}5$ см, оценим полное вязкое сопротивление, как $R_{\text{т}}\approx 0{,}2$. Это сопротивление невелико, но при сужении с площадью сечения меньше 0,1 см^2 действует закон для капиллярного трения в виде

$$r_{\text{т}}=12\mu b h_{\text{т}}^2/S_{\text{т}}^3,$$

где b — наименьший геометрический размер щели, $h_{\text{т}}$ — длина щели. Сопротивление капиллярного трения стремится к бесконечности при уменьшении площади сечения до нуля. Грубо оценивая динамическое сопротивление для тракта без сужения, получим $R_{\text{т д}}\approx 0{,}1$, т. е. величину того же порядка, что и вязкое сопротивление. Картина существенно изменяется при сужении с достаточно малой площадью, как это бывает при артикуляции фрикативных (щелевых) и взрывных звуков. В этом случае сопротивление в сужении значительно превысит сопротивление остальных участков речевого тракта и воздушный поток в сужении описывается тем же уравнением (5.2), что и для

голосовой щели. Итак, для процесса по постоянному току для давления и объемной скорости имеем следующую систему уравнений:

$$
\begin{gathered}
F''_{\text{м}} + 2g_{\text{м}} F'_{\text{м}} + \omega_{\text{м}}^2 F_{\text{м}} = A, \\
m_{\text{п}} x''_{\text{п}} + r_{\text{п}} x'_{\text{п}} + k_{\text{п}} x_{\text{п}} = F_{\text{м}} - p_{\text{л}} S_{\text{д}}, \\
V'_{\text{л}} = -S_{\text{д}} x', \\
C'_{\text{л}} = \frac{V'_{\text{л}}}{\rho_0 c_0^2}, \\
w'_{\text{л}} L_{\text{тр}} + w_{\text{л}} (R_{\text{л}0} + k_{\text{л}} w_{\text{л}}) = p_{\text{л}} - p_0, \\
(p_{\text{л}} C_{\text{л}})' = w_{\text{л}}, \\
p'_0 C_{\text{тр}} = w_{\text{л}} - w_{\text{г}}, \\
\rho_0 h_{\text{г}} \left(1 + \frac{S_0}{2S_{\text{г}}}\right) w'_{\text{г}} + \left(k_{\text{тг}} - \rho_0 h_{\text{г}} \frac{S_0 S'_{\text{г}}}{2S_{\text{г}}^2}\right) w_{\text{г}} + \frac{\rho_0 c_{\text{x}}}{2S_{\text{г}}} w_{\text{г}}^2 = (p_0 - p_{\text{т}}) S_{\text{г}}, \\
S_{\text{г}} = f_1 (p_0, p_{\text{т}}, w_{\text{г}}), \\
m_{\text{ст}} x''_{\text{ст}} + r_{\text{ст}} x'_{\text{ст}} + k_{\text{ст}} x_{\text{ст}} = p_{\text{т}}, \\
w_{\text{ст}} = S_{\text{ст}} x'_{\text{ст}}, \\
w_{\text{тв}} = (p_{\text{т}} C_{\text{т}})', \\
\rho_0 h_{\text{т}} w'_{\text{т}} + k_{\text{т}} w_{\text{т}} + \frac{\rho_0 c_{\text{xт}}}{2S_{\text{т}}^2} w_{\text{т}}^2 = p_{\text{т}} S_{\text{т}}, \\
w_{\text{г}} = w_{\text{ст}} + w_{\text{тв}} + w_{\text{т}}.
\end{gathered}
\tag{5.15}
$$

В системе (5.15) $S_{\text{т}}$ означает минимальную площадь сужения в речевом тракте. Площадь голосовой щели $S_{\text{г}}$ является функцией от перепада давления $p_0 - p_{\text{т}}$ и объемной скорости потока $w_{\text{г}}$. Эта площадь вычисляется с помощью модели механических колебаний голосовых складок. Различные способы описания колебаний складок будут рассматриваться в § 5.5, а в настоящем разделе для примера используем одномерную распределенную модель с тремя собственными функциями.

Для решения системы (5.15), содержащей нелинейные дифференциальные уравнения и уравнения с переменными параметрами, нужно найти такой способ, который был бы пригоден для реализации на микропроцессорах умеренной мощности. Это означает, что алгоритм решения (5.15) должен быть достаточно прост, не требовать большой памяти и слишком длительных вычислений. К счастью оказалось, что если от дифференциальных уравнений перейти к уравнениям в конечных разностях в некоторой специфической форме, то (5.15) превратится в систему алгебраических уравнений, которая решается итеративным способом, причем решения сходятся за 2—3 итерации.

Первое уравнение в системе (5.15) не зависит от других переменных, поэтому оно не участвует в итеративном процессе. Нелинейные уравнения для объемной скорости в голосовой щели и в сужении голосового тракта решаются методом, указанном в § 5.3.

Аналогично решается и уравнение для потока, вытекающего из легких. В итоге получаем следующую систему алгебраических уравнений в форме $y_{k+1}^{(i)} = f\left(a_1^{(i-1)}, a_2^{(i-1)}, \ldots, y_k^{(i-1)}\right)$, где индекс i означает номер итерации, а индекс k — отсчет во времени, $a_1^{(i-1)}$, $a_2^{(i-1)}$, ... — параметры системы, рассчитанные на предыдущем шаге итерации

$$x_{\text{п}k+1}^{(i)} = a_{\text{м}} b_1 \left(p_{\text{л}} - p_1^{(i-1)}\right) + b_2 x_{\text{п}k} + b_3 x_{\text{п}k-1},$$

$$C_{\text{л}}^{(i)} = 5 \cdot 10^{-4} \left(x_{\text{п}k} - x_{\text{п}k-1}^{(i-1)}\right) / \Delta t,$$

$$p_{1k+1}^{(i)} = p_{1k} + \left(1 - \exp\left\{-\Delta t \frac{C_{\text{л}}'^{(i-1)}}{C_{\text{л}}}\right\}\right)\left(\frac{w_{\text{л}}^{(i-1)}}{C_{\text{л}}'^{(i-1)}} - p_{1k} - p_{01}\right),$$

$$b_5 = w_{\text{л}k} + \left(1 - \exp\left\{-\Delta t \frac{R_{\text{л}0}}{L_{\text{тр}}}\right\}\right)\left(\frac{p_1^{(i-1)} - p_0^{(i-1)}}{R_{\text{л}0}} - w_{\text{л}k}\right),$$

$$w_{\text{л}k+1}^{(i)} = 0{,}5\left(\sqrt{1 + b_4 b_5} - 1\right) / b_4,$$

$$p_{0k+1}^{(i)} = p_{0k} + \Delta t \left(w_{\text{л}}^{(i-1)} - w_0^{(i-1)}\right) / C_{\text{тр}},$$

$$\Delta p_{k+1}^{(i)} = p_{0\,k+1}^{(i-1)} - p_{\text{т}\,k+1}^{(i-1)} + \Delta p_{\text{ак}\,k+1},$$

$$S_{\text{г}\,k+1}^{(i)} = f\left(\Delta p_{k+1}^{(i-1)}\right),$$

$$b_6 = \left(12\mu l_{\text{г}}^2 - \rho_0 S_{\text{г}\,k+1}'^{(i-1)} S_0\right) h_{\text{г}} / S_{\text{г}\,k+1}^{2\,(i-1)},$$

$$b_7 = \rho_0 h_{\text{г}} \left(1 + \frac{S_0}{S_{\text{г}\,k+1}'^{(i-1)}}\right) / b_6,$$

$$b_8 = 1 - \exp\left\{-\Delta t / b_7\right\},$$

$$b_9 = \frac{\rho_0 c_x b_8}{2 S_{\text{г}\,k+1}^{(i-1)} b_6},$$

$$b_{10} = w_{\text{г}k} + b_8 \left(\frac{\Delta p_{k+1}^{(i-1)} S_{\text{г}\,k+1}^{(i-1)}}{b_6} - w_{\text{г}k}\right),$$

$$w_{\text{г}\,k+1}^{(i)} = \left(\sqrt{1 + 4 b_{10}} - 1\right) / b_9,$$

$$b_{12} = w_{\text{т}k} + \left[1 + \exp\left\{-\frac{\Delta t\, 12\mu l_{\text{т}}^2}{\rho_0 S_{\text{т}}^2}\right\}\right]\left[\frac{p_{\text{т}\,k+1}^{(i-1)} S_{\text{т}}^3}{12\mu h_{\text{т}} l_{\text{т}}^2} - w_{\text{тк}}\right],$$

$$w_{\text{т}}^{(i)} = \frac{\sqrt{1 - 4 b_{12} b_{13}} - 1}{2 b_{13}},$$

$$x_{\text{ст}\,k+1}^{(i)} = b_{14} p_{\text{т}\,k+1}^{(i-1)} + b_{15} x_{\text{ст}\,k} + b_{16} x_{\text{ст}\,k-1},$$

$$w^{(i)}_{\text{ст}\,k+1}=\frac{x^{(i-1)}_{\text{ст}\,k+1}-x_{\text{ст}\,k}}{\Delta t},$$

$$p^{(i)}_{\text{т}\,k+1}=\begin{cases}\dfrac{p_{\text{тк}}+\Delta t\left[w^{(i-1)}_{\text{т}\,k+1}-w^{(i-1)}_{\text{ст}\,k+1}-w^{(i-1)}_{\text{т}\,k+1}\right]}{C_{\text{т}\,k+1}}, & C'_{\text{т}\,k+1}=0,\\ p_{\text{т}\,k}+\left(1-\exp-\dfrac{\Delta t\,C'_{\text{т}\,k+1}}{C_{\text{т}\,k}}\right)\left[\dfrac{w^{(i-1)}_{\text{т}\,k+1}-w^{(i-1)}_{\text{ст}\,k+1}-w^{(i-1)}_{\text{т}\,k+1}}{C'_{\text{т}\,k+1}}-\right.\\ \left.-p_{\text{тк}}-p_{0k}\right], & C'_{\text{т}\,k+1}\neq 0.\end{cases}\qquad(5.16)$$

Здесь коэффициент b_4 не зависит от текущих параметров и потому не участвует в итерациях:

$$b_4=\frac{\left(1-\exp\left\{-\Delta t\,\dfrac{R_{\text{ло}}}{L_{\text{тр}}}\right\}\right)k_{\text{л}}}{R_{\text{ло}}}.$$

В уравнении для потока через сужение в речевом тракте коэффициент b_{13} также не участвует в итерациях:

$$b_{13}=\left(1-\exp\left\{-\frac{\Delta t\,12\mu\,l_{\text{т}}^2}{\rho_0 S_{\text{т}}^2}\right\}\right)\frac{\rho_0 c_x S_{\text{т}}}{24\mu\,h_{\text{т}}\,l_{\text{т}}^2},$$

где $h_{\text{т}}$ — *глубина*, $l_{\text{т}}$ — *ширина*, $S_{\text{т}}$ — *минимальная площадь* сужения.

Величины p_{01} и $p_{0\text{т}}$ равны атмосферному давлению (10^6 г/(см·с^2)), а $p_{\text{ак}}$ есть разность давления, создаваемого акустическими колебаниями под и над голосовой щелью. Емкость речевого тракта $C_{\text{т}}$ рассчитывается как

$$C_{\text{т}}=V_{\text{т}}/(\rho_0 c_0^2),$$

где $V_{\text{т}}$ — объем речевого тракта от голосовой щели до смычки (если таковая имеется) или до губ:

$$V_{\text{т}}(t)=\int_0^{l_{\text{см}}} S_{\text{т}}(x,t)\,dx,$$

где $l_{\text{см}}$ — расстояние от голосовой щели до смычки или до губ.

Несмотря на кажущуюся громоздкость, система (5.16), по-видимому, является простейшей как в форме записи, так и в организации вычислительного процесса. Решениями этой системы являются объемная скорость потока в голосовой щели $w_{\text{г}}$ и объемная скорость в сужении речевого тракта после разрыва смычки $w_{\text{т}}$.

Производные по времени от этих величин служат источниками голосового и импульсного возбуждения. В процессе решения получаем величину изменения площади голосовой щели $S_{\text{г}}$ во времени. Представляет интерес и поведение

подсвязочного давления p_0 в зависимости от сопротивления голосовой щели. На рис. 5.16 показана зависимость от времени p_0, $S_г$, $w_г$ и $w'_г$ при возрастании от нуля до 10^4 г/(см·с^2) (или 10 см водяного столба) без учета влияния акустических

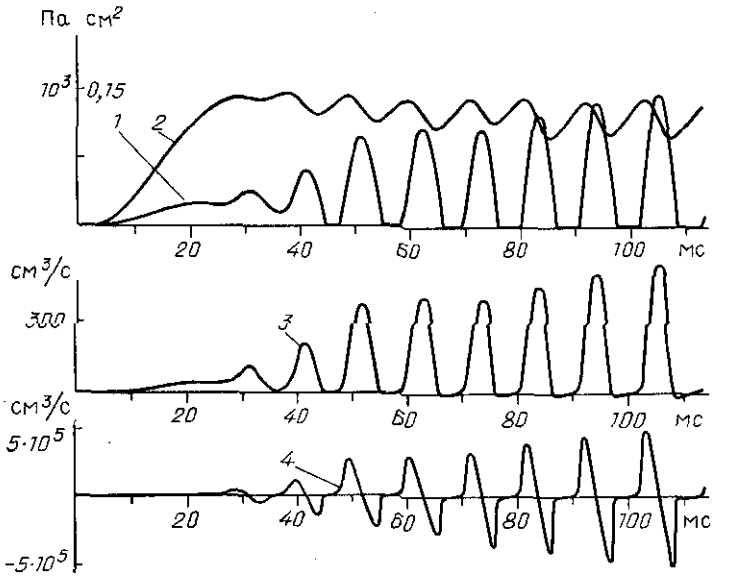


Рис. 5.16. Колебательные процессы в голосовом источнике: *1* — площадь голосовой щели, *2* — подсвязочное давление, *3* — объемная скорость воздушного потока, *4* — производная от объемной скорости

колебаний. Обращает на себя внимание то, что при раскрытии голосовой щели подсвязочное давление p_0 падает. Такое падение давления отмечается в работах, где производилось непосредственное измерение p_0 во время фонации с помощью миниатюрного микрофона, опускаемого под голосовую щель [125].

Величина падения p_0 при открытой голосовой щели зависит от отношения сопротивления легких и голосовой щели, в частности, от коэффициента динамического сопротивления легких. Минимум подсвязочного давления запаздывает относительно максимума площади голосовой щели, и это запаздывание является необходимым условием поддержания автоколебаний голосовых складок при отсутствии акустического влияния со стороны речевого тракта.

В возникновении автоколебаний существенную роль играет податливость стенок трахеи, создающая емкостную нагрузку. Без учета этой податливости автоколебания возникают только при наличии акустических колебаний в речевом тракте и трахее, тогда как с учетом податливости стенок трахеи, автоколебания

устойчивы и в отсутствие акустических колебаний. Емкостная нагрузка трахеи создает необходимое запаздывание между колебанием голосовых складок и колебаниями подсвязочного давления.

Для одномерной распределенной модели голосовых складок, которая была использована при расчете процессов, показанных на рис. 5.16, характерно то, что амплитуды положительного и отрицательного импульсов в производной от объектной скорости довольно близки. Частично это является следствием запаздывания минимума подсвязочного давления относительно максимальной площади голосовой щели, что приводит к снижению перепада давления на голосовой щели при смещении относительно фазы расхождения складок. При этом уменьшается крутизна заднего фронта импульса объемной скорости, а следовательно, и амплитуда отрицательного импульса сигнала источника голосового возбуждения. Необходимо отметить, что по мере развития колебаний голосовых складок отрицательный импульс становится все острее, и его задний фронт — все круче. Начиная с некоторого импульса, производная от объемной скорости терпит разрыв в момент схлопывания голосовой щели. Это означает, что спектр источника голосового возбуждения и амплитуды верхних формант возрастают, что способствует повышению качества синтетического речевого сигнала.

5.4.5. Механическое взаимодействие. Имеется ряд явлений, связанных с механическим взаимодействием гортани и остальных артикуляторных органов. Эти явления состоят в изменении частоты основного тона при изменении темпа артикуляции и типа гласного. Каждая гласная обладает некоторой характерной частотой основного тона, называемой собственной или внутренней частотой, причем эта частота повышается, если расположить гласные в следующий ряд: /А, И, У/.

По-видимому, существуют две причины непроизвольного изменения частоты основного тона в процессе артикуляции. Одна из них связана с увеличением напряжения мышц, непосредственно участвующих в выполнении какого-либо движения, но имеющих центры управления, которые находятся поблизости от центров управления мышцами, активными в данный момент. Подобное распространение возбуждения наблюдается во всех мышцах; оно также связано и с эмоциональным напряжением человека. Ускорение артикуляции, сопровождающееся повышенной мышечной активностью, может привести и к увеличению частоты основного тона. На основе моделирования этого механизма можно имитировать эмоциональное состояние и тип нервной системы человека в синтетической речи.

Другая причина непроизвольных изменений F_0 состоит в относительном движении щитовидного и перстневидного хрящей, которое приводит к изменению натяжения голосовых

складок. Для уяснения этого явления рассмотрим схему механических связей гортани (рис. 5.17).

Голосовые складки удлиняются и их натяжение возрастает при повороте черпаловидного хряща. Этот поворот может произойти либо вследствие сокращения перстне-черпаловидной мышцы, либо из-за увеличения расстояния между щитовидным и перстневидным хрящами. Подъем щитовидного хряща, приводящий к укорочению речевого тракта, очевидно, вызывается необходимостью формирования определенных частотных

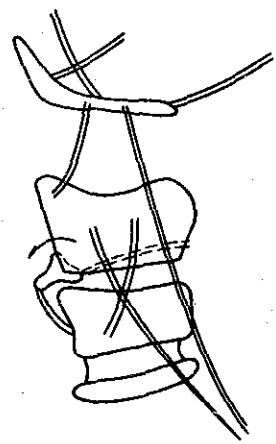


Рис. 5.17. Кинематическая схема гортани

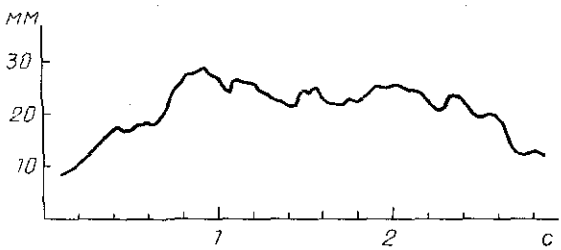


Рис. 5.18. Изменение высоты гортани во фразе

характеристик звуков речи. Поскольку перстневидный хрящ связан с трахеей более жестко, чем с щитовидным хрящом, то он поднимается на меньшую высоту, чем щитовидный хрящ, расстояние между ними увеличивается, черпаловидный хрящ поворачивается, голосовые складки натягиваются и частота основного тона повышается. Высота щитовидного хряща при артикуляции звука /И/ больше, чем при артикуляции /А/, поэтому увеличивается и собственная частота основного тона для /И/. При артикуляции звука /У/, однако, гортань ниже, чем при артикуляции /А/, поэтому повышение F_0 для /У/, по-видимому, связано с другими механизмами.

Есть основания полагать, что фразовая интонация формируется с помощью управления высотой гортани. Так, в [152] наблюдалась типичная картина быстрого подъема и постепенного опускания гортани, причем эти движения коррелировали с фразовой интонацией (рис. 5.18). Вполне возможно, что управление быстрыми и медленными изменениями частоты основного тона F_0 разделено между механизмами напряжения голосовых мышц и подъема — опускания гортани.

Таким образом, и в управлении голосовым источником возбуждения нужно предусмотреть распределение команд на натяжение и напряжение голосовых складок в зависимости от интонационного контура. Поскольку натяжение и напряжение складок несколько по-разному влияют на характеристики

колебаний складок, возможно, существуют и некоторые вторичные маркеры интонации, например, в виде формы импульсов возбуждения и длительности фаз открытой и закрытой голосовой щели.

§ 5.5. Модели механических колебаний голосовых складок

Объемная скорость воздушного потока в голосовой щели есть произведение скорости частиц воздуха на площадь поперечного сечения голосовой щели. Следовательно, форма импульса объемной скорости зависит от движения складок и от формы голосовой щели. Для максимального приближения качества синтеза к качеству естественной речи необходимо располагать, помимо аэродинамической модели, и моделью механических колебаний голосовых складок. Имеется два типа моделей: в одном из них каждая голосовая складка описывается как система с сосредоточенными параметрами, а в другом — как распределенная система. Представителями первого типа служат одномассовая и двухмассовая модели [97, 123], а второго — одно-, двух- и трехмерные распределенные модели [59, 72, 198, 199]. Качество источника голосового возбуждения улучшается по мере того, как используются все более физически адекватные модели механических колебаний складок, но, как и всегда, это улучшение покупается ценой усложнения вычислительных процедур.

Прежде чем приступить к описанию моделей механических колебаний, рассмотрим силы, действующие на поверхность голосовых складок, поскольку способ учета этих сил мало зависит от выбранной модели. На рис. 5.19 показана голосовая щель с ее окрестностями, линии тока воздуха и качественная картина изменения давления. В трахее на достаточном удалении от голосовой щели действует давление p_0. Если голосовая щель закрыта, то на скошенную поверхность голосовых складок в направлении оси речевого тракта действует сила $F_x = p_0 S_b \sin\alpha - p_т S_c$, а в перпендикулярном направлении — сила $F_y = p_0 S_b \cos\alpha$, где $p_т$ — давление в речевом тракте, S_b — площадь скошенного участка складок, α — угол наклона скошенного участка относительно оси речевого тракта, S_c — площадь верхней поверхности голосовой складки. Отсюда видно, что если в модели механических колебаний скос нижней поверхности складок отсутствует, т. е. эта поверхность расположена перпендикулярно оси тракта,

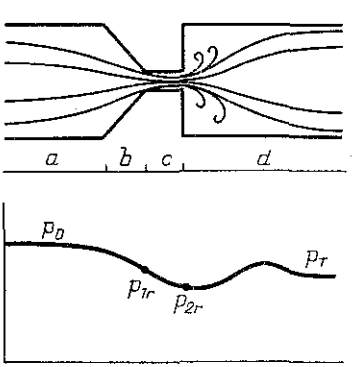


Рис. 5.19. Распределение давления в окрестности голосовой щели

то в закрытом состоянии голосовой щели сила F_y, раздвигающая складки, равна нулю, и из этого положения колебания начаться не могут. При раскрытой голосовой щели линии тока воздуха начинают сужаться еще до входа в голосовую щель, в результате чего поток ускоряется и, в силу закона Бернулли, давление $p_{1г}$ на входе голосовой щели ниже давления в трахее p_0. Величина падения давления пропорциональна квадрату скорости потока $c_{x1}\rho_0 v^2/2$, где коэффициент сопротивления c_{x1} зависит от формы щели, угла наклона скошенного участка складок и числа Рейнольдса. Эта величина определялась в экспериментах по продувке воздуха через модели гортани и, как обсуждалось в § 5.3, может быть приближенно описана как

$$\ln c_{x1} = \ln 3 - \frac{\ln 2}{\ln 5000}\ln \mathrm{Re}, \quad 1 \leqslant \mathrm{Re} \leqslant 5000,$$

$$c_{x1} = 1{,}4, \qquad \mathrm{Re} > 5000.$$

Внутри голосовой щели давление падает от входа к выходу по линейному закону, так как вследствие малых размеров щели сопротивление потоку не зависит от его скорости. Голосовая щель выходит в нижнюю часть речевого тракта, площадь которой значительно больше площади щели. Поэтому сечение воздушного потока на некотором протяжении сохраняет площадь голосовой щели, затем поток начинает завихряться, и лишь после этого линии тока расширяются, занимая все сечения речевого тракта. Вследствие этих явлений давление над голосовой щелью оказывается несколько ниже, чем давление в тракте, что отчасти компенсирует падение давления на входе в голосовую щель. Степень понижения давления относительно $p_т$ также пропорциональна квадрату скорости потока, но с другим коэффициентом c_{x2}:

$$c_{x2} = 2\frac{S_г}{S_т}\left(1 - \frac{S_г}{S_т}\right),$$

где $S_т$ — площадь тракта над голосовой щелью. В результате можно записать коэффициент динамического сопротивления как $c_x = c_{x1} - c_{x2}$.

Средняя сила, действующая на голосовые складки внутри голосовой щели, пропорциональна давлению в щели:

$$F_{ус} = (p_{1г} + p_{2г})\, h_г l_г/2,$$

где $h_г$ — глубина голосовой щели, $l_г$ — длина щели, т. е. длина голосовых складок. Аналогично можно принять, что средняя сила, действующая на скошенный участок поперек оси тракта есть $F_{ув} = (p_0 + p_{1г}) h_{ск} l_г/2$, а средняя сила, действующая вдоль оси тракта $F_{хв} = [0{,}5(p_0 + p_{1г}) - p_{2г}]\, b_{ск} b_г$, где $h_{ск}$ — длина скошенного участка вдоль оси тракта, а $b_{ск}$ — толщина голосовых складок.

Исходя из условия неразрывности воздушного потока, требующего равенства объемной скорости на всех участках при установившемся движении, можно записать:

$$p_0 - p_{2г} = \frac{c_{x1}\rho_0}{2}\frac{w_г^2}{S_г^2},$$

$$p_т - p_{2г} = \frac{c_{x2}\rho_0}{2}\frac{w_г^2}{S_г^2}.$$

В эту систему следует добавить уравнение аэродинамики, описывающее падение давления внутри голосовой щели:

$$p_{1г} - p_{2г} = (\rho_0 h_г w_г' + k_т w_г)/S_г.$$

Отсюда вновь получаем уравнение (5.2) в форме

$$\frac{\rho_0 h_г w_г'}{S_г} + (R_т + R_{тд} w_г) w_г = p_0 - p_т,$$

где

$$R_т = \frac{k_т}{S_г}, \quad R_{тд} = \frac{\rho_0 c_x w_г}{2S_г^2}.$$

Используя только давление в тракте $p_т$ и трахее p_0, найдем, что средняя сила, действующая на складки внутри голосовой щели, есть

$$F_{ус} = p_0 - \left[\frac{\rho_0 c_{x1}}{2}\frac{w_г^2}{S_г^2} + \frac{1}{2S_г}(\rho_0 h_г w_г' + k_т w_г)\right] l_г h_г,$$

или

$$F_{ус} = \frac{1}{2}\left(p_0 + p_т - \frac{\rho_0 c_x}{2S_г^2} w_г^2\right) l_г h_г,$$

а силы, действующие на скошенный участок складок:

$$F_{ув} = \left(p_0 + \frac{\rho_0 c_{x1}}{4S_г^2} w_г^2\right) h_{ск} l_г,$$

$$F_{хв} = \left(p_0 - p_т - \frac{\rho_0 c_x}{2S_г^2} w_г^2\right) b_{ск} b_г.$$

Рассчитанная таким образом средняя сила используется в сосредоточенных моделях механических колебаний. Для распределенных моделей нужно пользоваться распределенными силами $F_{ув}$ и $F_{ус}$:

$$F_{ув}(x) = \left(p_0 - \frac{p_0 - p_{1г}}{h_{ск}} x\right) h_{ск} l_г, \quad 0 \leqslant x \leqslant h_{ск},$$

$$F_{ус}(x) = \left(p_{1г} - \frac{p_{1г} - p_{2г}}{h_г} x\right) h_г l_г, \quad 0 \leqslant x \leqslant h_г,$$

или

$$F_{\text{уп}}(x)=\left(p_0-\frac{\rho_0 c_{x1}}{2}\frac{w_{\text{г}}^2}{S_{\text{г}}^2}x\right)h_{\text{ск}}l_{\text{г}},\quad 0\leqslant x\leqslant h_{\text{ск}},$$

$$F_{\text{ус}}(x)=\left[p_0-\frac{\rho_0 c_{x1}}{2}\frac{w_{\text{г}}^2}{S_{\text{г}}^2}-x\left(\frac{\rho_0 h_{\text{г}}}{S_{\text{г}}}w_{\text{г}}'+\frac{k_{\text{т}}}{S_{\text{г}}}w_{\text{г}}\right)\right]h_{\text{г}}l_{\text{г}},$$

$$0\leqslant x\leqslant l_{\text{г}}.$$

5.5.1. Сосредоточенные модели. *Одномассовая модель.* Располагая силами, действующими на поверхность голосовых складок, приступим к рассмотрению различных моделей механических колебаний. Одной из первых моделей голосового источника, в которой было достигнуто самовозбуждение, является одномассовая модель [97]. В этой модели каждая складка представлена в виде сосредоточенных массы $m_{\text{м}}$, упругости $k_{\text{м}}$ и вязкого трения $r_{\text{м}}$, причем обе складки считаются идентичными, так что все расчеты ведутся относительно одной складки. Уравнение движения есть

$$m_{\text{м}}x''+r_{\text{м}}x'+k_{\text{м}}x=F(t),$$

где x — отклонение от положения равновесия. В [97] сила $F(t)$ на интервале закрытой голосовой щели принималась равной $F(t)=p_0h_{\text{г}}l_{\text{г}}/2$, т. е. неявно вводился скос складок. По отклонению x рассчитывается площадь голосовой щели $S_{\text{г}}=2xl_{\text{г}}$ в предположении прямоугольной формы щели. При соударении рассматривались два варианта — абсолютно упругий удар и вязкий удар, сопровождающийся увеличением коэффициента трения.

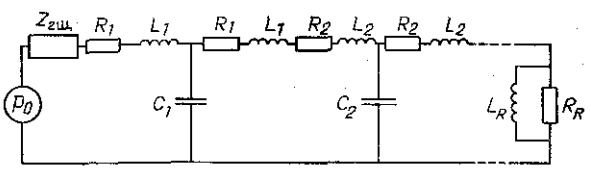


Рис. 5.20. Речевой тракт как длинная линия

Если предоставить речевой тракт в виде длинной линии, показанной на рис. 5.20, то система уравнений для одномассовой модели, нагруженной на речевой тракт есть:

$$m_{\text{м}}x''+r_{\text{м}}x'+k_{\text{м}}x=F(t),$$
$$S_{\text{г}}=2xl_{\text{г}},$$

$$R_{\text{тг}} = \frac{12\mu h_{\text{г}} l_{\text{г}}^2}{S_{\text{г}}^3},$$

$$R_{\text{тд}} = \frac{\rho_0 c_x}{2} \frac{w_{\text{г}}}{S_{\text{г}}^2},$$

$$(R_1 + R_{\text{тг}} + R_{\text{тд}}) w_{\text{г}} + \frac{\rho_0 h w_1'}{S_{\text{г}}} + (L_1 w_{\text{г}})' + \frac{1}{c_1}\int (w_{\text{г}} - w)\, dt = p_0,$$

$$(R_1 + R_2) w_1 + [(L_1 + L_2) w_1]' + \frac{1}{C_2}\int (w_1 - w) dt + \frac{1}{C_1}\int (w_1 - w_{\text{г}})\, dt = 0,$$

.....

$$[L_R (w_R + w_N)]' + R_R w_R = 0,$$

где

$$L_i = \frac{\rho_0 \Delta l_i}{2 S_i}, \quad C_i = \frac{S_i \Delta l_i}{\rho_0 c_0^2},$$

Δl_i — длина цилиндрической секции, S_i — площадь секции, R_R и L_R — сопротивление и индуктивность излучения, N — число секций. Переведенная в конечно-разностную форму эта система уравнений решается итеративным способом.

Одномассовая модель учитывает акустическое взаимодействие речевого тракта и голосового источника, а также влияние подсвязочного давления на частоту основного тона. Основным недостатком одномассовой модели является зависимость условий самовозбуждения от акустической нагрузки — когда нагрузка приобретает емкостный характер, например, при образовании смычки, автоколебания голосовых складок не возникают. Качество сигнала, синтезированного с одномассовой моделью, довольно низкое, поэтому несмотря на ее простоту, она в чистом виде почти не используется в синтезаторах.

Двухмассовая модель. Следующий шаг в приближении к действительности сделан в двухмассовой модели [123], которая учитывает очень важное явление — сдвиг по фазе между колебаниями верхней и нижней голосовых складок. В двухмассовой модели каждая голосовая складка представлена в виде элемента с сосредоточенными вязкостью и упругостью, причем эти элементы связаны между собой упругим соединением (рис. 5.21). Уравнения колебаний тканей механической системы есть

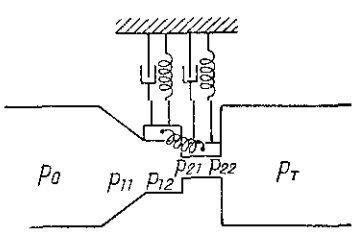


Рис. 5.21. Двухмассовая модель голосовых складок

$$m_1 x_1'' + r_1 x_1' + s_1(x_1) + k_c(x_1 - x_2) = F_1,$$

$$m_2 x_2'' + r_2 x_2' + s_2(x_2) + k_c(x_2 - x_1) = F_2,$$

где m и r — соответственно масса и коэффициент вязкого трения для каждого сосредоточенного элемента, x — смещение каждого элемента, k_c — коэффициент упругой связи между элементами. Упругая характеристика тканей голосовых складок $s(x)$ различна для случая расхождения и схлопывания. При расхождении складок

$$s_i(x_i)=k_i(x_i+\eta_i x_i^3), \quad x_i > -S_{r0i}/(2l_r),$$

тогда как при схлопывании принимается

$$s_i(x_i)=k_i(x_i+\eta_{1i}x_i^3)+\eta_{2i}\left[\left(x_i+\frac{S_{ri}}{l_r}\right)+\eta_{3i}\left(x_i+\frac{S_{ri}}{l_r}\right)^3\right],$$

$$x_i \leqslant -\frac{S_{r0i}}{l_r},$$

где k_i, η_{1i}, η_{2i}, η_{3i} — некоторые постоянные величины, S_{r0i} — начальная площадь голосовой щели для каждого сосредоточенного элемента.

При открытой голосовой щели, т. е. при $x_1 > -S_{r01}/l_r$ и $x_2 > -S_{r02}/l_r$, на каждый элемент действуют силы

$$F_1=\left[p_0-\frac{\rho_0 c_{x1}}{2}\frac{w_r^2}{S_{r1}^2}-\frac{1}{2}\left(R_{r1}w_r+L_{r1}w_r'\right)\right]l_r d_1,$$

$$F_2=F_1\frac{d_1}{d_2}-\left[\frac{R_{r1}+R_{r2}}{2}w_r+\frac{L_{r1}+L_{r2}}{2}w_r'-\frac{\rho_0}{2}\left(\frac{1}{S_{r2}^2}-\frac{1}{S_{r1}^2}\right)w_r^2\right]l_r d_2,$$

где d_1 и d_2 — протяженность каждого сосредоточенного элемента вдоль голосовой щели,

$$S_{ri}=S_{r0i}+2x_i l_r,$$
$$L_{ri}=\rho_0 d_i/S_{ri}.$$

В [123] использовались следующие значения: масса каждой складки $m_1+m_2=\rho_т l_r(d_1+d_2)=0{,}15$ г, $d_1+d_2=0{,}3$ см, коэффициенты для пружин: $\eta_{11}=\eta_{12}=100$, $\eta_{31}+\eta_{32}=500$, $k_1=8\cdot 10^4$ дин/см, $k_2=8\cdot 10^3$ дин/см, $k_c=2{,}5\cdot 10^4$ дин/см. Относительный коэффициент затухания

$$\xi_i=\frac{r_i}{2\sqrt{m_i k_i}}\approx 0{,}1—0{,}2.$$

Система уравнения механических колебаний объединяется с уравнениями колебаний в длинной линии в одну систему так же, как и для одномассовой модели. На рис. 5.22 показано изменение площади голосовой щели для каждого сосредоточенного элемента и объемная скорость, соответствующая эквивалентной площади, равная минимальному значению площадей S_{r1} и S_{r2}. Благодаря этому увеличивается задержка между изменениями скорости воздушного потока в голосовой щели и ее эквивалентной площади, что способствует развитию автоколебаний. В результате колебания сосредоточенных элементов возможны и в условиях емкостной нагрузки, когда частота основного тона выше частоты первого резонанса речевого тракта.

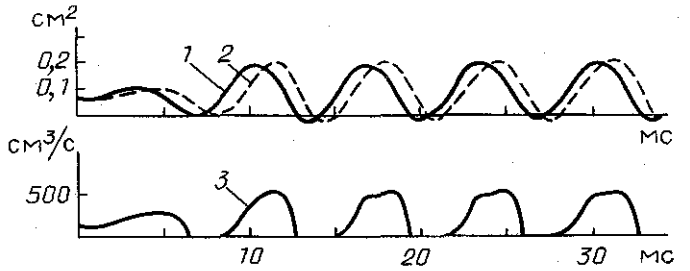


Рис. 5.22. Колебательные процессы в двухмассовой модели: *1* — площадь голосовой щели для первого элемента, *2* — площадь голосовой щели для второго элемента, *3* — объемная скорость на выходе голосовой щели

В [113] исследовались условия самовозбуждения двухмассовой модели и акустическое взаимодействие с речевым трактом. Согласно критерию Рауса—Гурвица, условия возбуждения автоколебаний есть

$$\omega_{01}^2-\omega_{02}^2+4(\beta_1+\beta_2)(\beta_1\omega_{02}^2+\beta_2\omega_{01}^2)\pm\frac{(\beta_1+\beta_2)^2}{\beta_1\beta_2}(\omega_{01}\omega_{02}k)^2<0,$$

где

$$\beta_i=r_i/m_i,\quad \omega_{01}=f(s_i,\ k_c,\ w_r)/m_i,\quad k\ll 1.$$

Индуктивная нагрузка (при высокой частоте первого резонанса тракта) облегчает возникновение автоколебаний и понижает частоту основного тона F_0, тогда как емкостная нагрузка затрудняет автоколебания и повышает F_0. Частота основного тона повышается и с увеличением активного сопротивления. Хотя акустическое взаимодействие речевого тракта с двухмассовой моделью меньше, чем с одномассовой моделью, оно демонстрирует эффект обратный тому, какой наблюдался в действительности: при переходе от закрытых гласных (типа /И/) к открытым (типа /А/) частота основного тона для двухмассовой модели возрастает, тогда как в естественной речи она падает.

Несмотря на этот недостаток, двухмассовая модель обеспечивает приемлемое качество звучания синтезаторов, поскольку она лучше описывает изменение массы воздуха в голосовой щели. Поэтому она используется не только в формантных, но и в артикуляторных синтезаторах. Сложность реализации двухмассовой модели не намного превышает сложность реализации одномассовой модели и эта разница совершенно незначительна, если голосовой источник используется для возбуждения артикуляторного синтезатора, сложность реализации которого весьма высока.

5.5.2. Распределенная модель. Каждая голосовая складка представляет собой упругое тело, состоящее из двух сильно различающихся по механическим характеристикам структур: довольно жестких мышечных волокон и сравнительно мягких тканей. Имеется также два механизма изменения жесткости голосовых складок: один из них связан с напряжением

внутрискладочных мышц, а другой — с натяжением складок путем поворота черпаловидных хрящей. Оба механизма могут действовать независимо, и в результате этого геометрические параметры и механические характеристики голосовых складок меняются в широких пределах. Может показаться, что механика упругих колебаний складок слишком сложна для того, чтобы распределенные модели голосового источника использовались помимо чисто исследовательских синтезаторов, однако ниже будут приведены приемы для существенного упрощения вычислений без потери физической адекватности.

В данном разделе мы рассмотрим одномерные и двухмерные модели голосового источника, описывающие упругие колебания вдоль складок и вдоль голосовой щели, а вертикальные колебания складок обсудим в следующем разделе. Напомним, что в [59] была получена система дифференциальных уравнений, описывающих форму пространственных деформаций X, Y и характер колебаний во времени T:

$$J_y \frac{d^4X}{dx^4} - \frac{NS_x}{E+N} \frac{d^2X}{dx^2} - \alpha^2 X = 0, \tag{5.17}$$

$$J_x \frac{d^4Y}{dy^4} - \frac{NS_y}{E+N} \frac{d^2Y}{dy^2} - \beta^2 Y = 0, \tag{5.18}$$

$$\rho_т \frac{d^2T}{dt^2} + r\frac{dT}{dt} + (c_п + \lambda^2)\, T = 0, \tag{5.19}$$

где x — координата вдоль голосовой складки, y — координата вдоль оси речевого тракта (вдоль голосовой щели), t — время, J_x, J_y — моменты инерции сечения складок относительно осей x и y, $c_п$ — упругость тканей, на которые опирается голосовая складка, α и β — собственные числа, $\lambda^2 = \alpha^2 + \beta^2$, E — модель упругости тканей в расслабленном состоянии, N — натяжение или напряжение тканей.

Решение системы (5.17), (5.18), (5.19) представляется в виде

$$U(x, y, t) = X(x)\, Y(y)\, T(t),$$

где U — смещение поверхности голосовых складок. Пользуясь разложением в ряд Фурье по собственным функциям $X_i(x)$, $Y_j(y)$, получаем

$$U(x, y, t) = \sum_{i,j} X_i(x)\, Y_j(y)\, T_{ij}(t).$$

Вид собственных функций X_i и Y_j определяется граничными условиями и в общем случае выражается как

$$X_i(x) = A_i \operatorname{sh} p_{1i}x + B_1 \operatorname{ch} p_{1i}x + C_i \sin p_{2i}x + D_i \cos p_{2i}x.$$

Решение уравнения (5.17) и (5.18), конечно, возможно численными методами, особенно в случае, когда учитывается изменение коэффициентов уравнений в зависимости от координат x и y. Несколько огрубляя решение, однако, можно

избежать необходимости использования численных методов, и получить достаточно хорошие аналитические решения, положив все коэффициенты постоянными и задавшись простыми граничными условиями.

Поскольку голосовые складки своими передними концами прикрепляются к щитовидному хрящу, то в процессе их колебаний смещение этого конца невозможно, т. е. $X(0)=0$. С некоторым приближением можно принять, что и угол наклона складки постоянен, т. е. $X'(0)=0$. Эти условия соответствуют так называемой жесткой заделке конца упругого тела. Задний конец складок прикрепляется к черпаловидным хрящам, а степень податливости этих хрящей зависит от усилий, развиваемых мышцами, которые сводят и разводят, и поворачивают эти хрящи. Если эти усилия велики, то и на заднем конце можно принять условия типа жесткой заделки, т. е. $X(l_r)=X'(l_r)=0$, где l_r — длина голосовых складок.

Для условий типа жесткой заделки получены решения для собственных функций $X_i(x)$:

$$X_i(x)=a_{1i}(\operatorname{ch} p_{1i}x-\cos p_{2i}x)+a_{2i}\left(\operatorname{sh} p_{1i}x-\frac{p_{1i}}{p_{2i}}\sin p_{2i}x\right),$$

где

$$p_{11}=4{,}7299/l_r,\quad p_{12}=7{,}853/l_r,\quad p_{13}=3{,}5\pi/l_r,$$

$$p_{2i}^2=p_{1i}^2-\frac{S\eta}{J_y(1+\eta)},\quad \eta=\frac{N}{E},\quad a_{2i}=a_{1i}p_{2i}\frac{\cos p_{2i}l_r-\operatorname{ch} p_{1i}l_r}{p_2\operatorname{sh} p_{1i}l_r-p_{1i}\sin p_{2i}l_r},$$

$$\alpha_i=p_{1i}p_{2i},$$

$$\omega_i=\sqrt{\frac{c_n+\alpha_i^2(E+N)J_y}{\rho_r}}. \qquad (5.20)$$

Из (5.20) видно, что имеется несколько возможностей для управления собственными частотами колебаний голосовых складок. Первая возможность состоит в изменении напряжения, голосовой мышцы, т. е. отношения N/E. Если упругостью подстилающего слоя c_n можно пренебречь по сравнению со вторым членом в (5.20), то собственная частота ω_i изменяется пропорционально корню квадратному из N/E и пропорционально собственному числу α_i:

$$\omega_i\approx\alpha_i\sqrt{\frac{J_yE}{\rho_r}\left(1+\frac{N}{E}\right)},$$

а поскольку в нормальном режиме $N/E\gg1$, то

$$\omega_i\approx\alpha_i\sqrt{J_yN/\rho_r}.$$

Другая возможность управления частотой ω_i состоит в изменении формы поперечного сечения голосовой складки. При этом меняется момент инерции J_y. Для прямоугольного и трапециевидного сечения момент инерции сильно зависит от высоты

голосовой складки $b_г$. Изменение площади сечения S_x при сохранении формы также влияет на ω_i, поскольку погонная плотность $\rho_т = \rho_{0т} S_x$, где $\rho_{0т} = 1{,}05$ г/см^3. Уменьшение площади S_x приводит к возрастанию ω_i.

Если длина голосовых складок увеличивается вследствие поворота черпаловидных хрящей, то модель упругости возрастает по экспоненциальному закону

$$E = E^{(0)} \exp\{c_E \delta l_г\},$$

где $\delta l_г$ — относительное удлинение голосовых складок. Удлинение складок и увеличение жесткости противоположным образом влияют на собственные частоты, однако увеличение жесткости преобладает. Возрастание жесткости при натяжении складок влияет на собственные частоты тем больше, чем меньше натяжение голосовых мышц N. Оно наибольшее для тканей, окружающих голосовые мышцы. Полный учет этого эффекта возможен лишь в многослойной модели.

Удлинение голосовых складок может достигать 25%, что соответствует 3 мм при длине $l_г = 1{,}2$ см. Такое удлинение приводит к увеличению жесткости складок вдвое [59]. При удлинении складок увеличивается жесткость $c_п$ подстилающего слоя. Совместное действие всех этих факторов приводит к повышению частоты основного тона и приближенно можно записать

$$\omega_i^2 \approx \frac{c_п^{(0)} \exp(c_E \delta l_г) + p_1^2 p_2^2 J_y [E^{(0)} \exp(c_E \delta l_г) + N]}{\rho_т}.$$

Зависимость собственных частот ω_i от напряжения N/E может быть вычислена и проще. Из рис. 5.23 видно, что в диапазоне от $4 \leqslant N/E \leqslant 12$, $\omega_i(N/E)$ хорошо аппроксимируется линейной функцией

$$\omega_i = a_i + b_i \frac{N}{E},$$

причем a_i и b_i, хотя и зависят от длины голосовых складок, но при фиксированной $l_г \omega_i$ линейно зависит от N/E.

Имеется еще одна возможность воздействия на собственные частоты ω_i. Возьмем идеализированный случай, когда смещение заднего конца голосовых складок не встречает никакого сопротивления, т. е. задний конец свободен. Тогда граничные условия есть $X''(l_г) = X'''(l_г) = 0$, и собственные функции имеют следующий вид:

$$X_i(x) = a_{1i}(\operatorname{ch} p_i x - \cos p_i x) + a_{2i}(\operatorname{sh} p_i x - \sin p_i x),$$

где

$$p_1 = 1{,}8751/l_г, \quad p_2 = 4{,}6941/l_г, \quad p_3 = 2{,}5\pi/l_г.$$

Сравнивая эти собственные числа с теми, которые имеются для жесткой заделки, видим, что изменяя тип граничных

условий, можно изменить собственные числа α_i, а вслед за ними и собственные частоты ω_i. Поскольку жесткость и усилия мышц, управляющих положением черпаловидных хрящей, могут быть различными для разных дикторов и изменяться

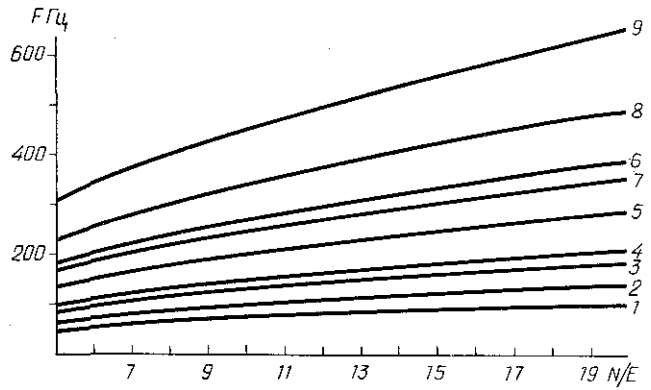


Рис. 5.23. Зависимость собственных частот голосовых складок от натяжения и длины складок. Первая гармоника: *1* — $l_г = 2{,}2$ см, *2* — $l_г = 1{,}6$ см, *3* — $l_г = 1{,}2$ см. Вторая гармоника: *4* — $l_г = 2{,}2$ см, *5* — $l_г = 1{,}6$ см, *6* — $l_г = 1{,}2$ см. Третья гармоника: *7* — $l_г = 2{,}2$ см, *8* — $l_г = 1{,}6$ см, *9* — $l_г = 1{,}2$ см

в процессе речеобразования, то управление граничными условиями на заднем конце голосовых складок приводит к изменению частоты основного тона и тембра голосового источника, т. е. соотношения между амплитудами собственных функций.

Все вышесказанное с некоторыми изменениями относится и к анализу свойств двухмерной модели.

Вследствие неполного сглаживания импульсов двигательных единиц в мышцах гортани, натяжение голосовых складок меняется по случайному закону, причем спектр этого шума не распространяется выше 15—20 Гц. В модели голосового источника для синтезатора речи эти случайные колебания можно генерировать разными способами.

Практически все библиотеки программного обеспечения для ЭВМ имеют стандартную подпрограмму для формирования шума ξ с равномерным распределением на интервале $[0, 1]$. Пользуясь центральной предельной теоремой, из шума с равномерным распределением нетрудно получить шум с нормальным распределением η путем суммирования не слишком большого числа реализаций ξ и последующего вычитания постоянной составляющей. Практика показывает, что в задачах синтеза речи достаточно использовать 16 компонент:

$$\eta = \sum_{i=1}^{16} \xi_i - 8.$$

Для того чтобы получить колебания натяжения голосовых складок с соответствующим спектром, случайный сигнал

η пропускается через фильтр низкой частоты, реализованный в виде дифференциального уравнения второго порядка:

$$\gamma'' + 2g\gamma' + \omega_0^2\gamma = \eta,$$

где $2\pi\omega_0 = 15$ Гц. Случайная величина γ, умноженная на некоторый коэффициент, добавляется к натяжению голосовых складок N, имитируя физиологический тремор.

Д. Клатт использует другой способ формирования случайных отклонений частоты основного тона, в котором вместо генерирования случайного процесса суммируются три синусоидальные компоненты со специально подобранными частотами:

$$\Delta F_0 = kF_0(\sin\omega_1 t + \sin\omega_2 t + \sin\omega_3 t),$$

где $\omega_1/2\pi = 4{,}7$ Гц, $\omega_2/2\pi = 7{,}1$ Гц, $\omega_3/2\pi = 12{,}7$ Гц. Вследствие некратности частот синусоидальных компонент результирующий сигнал выглядит в достаточной степени случайным.

Одномерная модель. Опишем свойства одномерной модели голосового источника, т. е. такой модели, в которой уравнение (5.16) отсутствует, и предполагается существование колебаний только вдоль голосовых складок. Пусть на поверхность голосовых складок действует распределенная сила $F(x, t)$. Тогда каждая мода собственных колебаний возбуждается парциальной силой

$$F_i(t) = \frac{2}{l_г}\int_0^{l_г} F(x, t)\, X_i(x)\, dx,$$

и, если принять, что $F(x, t) = F(t)$, т. е. равномерно распределена по поверхности складок, то

$$F_i(t) = \frac{2F(t)}{l_г}\left[a_{1i}\left(\frac{\operatorname{sh} p_{1i}x}{p_{1i}} - \frac{\sin p_{2i}x}{p_{2i}}\right) + a_{2i}\left(\frac{\operatorname{ch} p_{1i}x}{p_{1i}} + \frac{\cos p_{2i}x}{p_{2i}}\right)\right]\Bigg|_0^{l_г}.$$

Учитывая зависимость a_{1i} от корней p_{1i} и p_{2i}, найдем, что $F_i(t)$ уменьшается пропорционально $1/p_{1i}^3$, так что амплитуда третьей моды почти в десять раз меньше амплитуды первой моды. Поэтому для описания формы голосовых складок при жестких граничных условиях на обоих концах складок нет смысла использовать больше, чем три собственные функции.

При жестких граничных условиях на обоих концах складок вторая собственная функция $X_2(x)$ нечетна относительно середины голосовых складок. Следовательно, если сила $F(x, t)$ симметрична относительно $x = l_г/2$, то интеграл от произведения $F(x, t)\, X_2(x)$ равен нулю, возбуждающая сила $F_2(t) = 0$, и в такой модели $X_2(x)$ не участвует в образовании формы голосовых складок — остаются лишь четные собственные функции $X_1(x)$ и $X_3(x)$. В действительности же задний конец голосовых складок обладает некоторой податливостью, вследствие чего $X_2(x)$ несимметрична относительно середины складок и ее

нужно учитывать в колебаниях. Даже для четной $X_2(x)$ могут создаться условия для ее возбуждения, если сила $F(x, y)$ не симметрична относительно $x = l_г/2$.

Колебания голосовых складок могут происходить и тогда, когда они разведены на некоторое расстояние y_1. При этом голосовая щель имеет форму треугольника, так как передние концы складок соприкасаются друг с другом. В такой щели эффект вязкого трения в пограничном слое приводит к уменьшению скорости воздушного потока в некоторой окрестности переднего конца складок и к смещению равнодействующей аэродинамических сил относительно середины складок в сторону их заднего конца.

Наконец, реальные голосовые складки не однородны по своим геометрическим и механическим параметрам, так что на самом деле вторая собственная функция не симметрична относительно $x = l_г/2$ даже для жесткой заделки заднего конца складок. Все это приводит к необходимости учитывать $X_2(x)$ в одномерной модели.

Присутствие второй и третьей мод в колебаниях голосовых складок создает эффект не одновременного схлопывания. При схождении голосовых складок для первой собственной функции область их контакта движется от концов к середине, так как для всех высших мод контакт складок происходит сначала вблизи одного конца, а затем вблизи другого. При этом силы удара несимметрично распределены по поверхности складок, как бы перемещаясь от одного конца складок к другому. Тот участок складок, который раньше вступил в контакт с другой складкой, начнет и раньше расходиться. В зависимости от остальных условий этот сдвиг может либо сохраняться в процессе фонации, либо изменяться, в результате форма импульсов голосового возбуждения меняется во времени. Известно явление диплофонии, при котором форма импульсов голосовых складок более похожа не для соседних импульсов, а через один. Возможно, что это явление связано со сдвигом фаз в развитии колебаний разных мод.

Сдвиг фаз в контакте переднего и заднего участков голосовых складок аналогичен сдвигу фаз в колебаниях верхней и нижней кромки складок. В том случае, когда частота второй моды ω_2 близка к удвоенной частоте основного тона, возбуждающая сила $F_2(t)$ действует в фазе с колебаниями и создаются условия для резонанса на частоте ω_2. В результате амплитуда колебаний на этой частоте возрастает и становится сравнимой с амплитудой колебаний на частоте ω_1. Вторая мода иногда отчетливо просматривается на скоростных фильмах.

Поскольку частота основного тона непрерывно изменяется, условия резонанса выполняются лучше или хуже, и вблизи резонанса можно ожидать заметные отклонения в амплитуде второй моды, вследствие чего и форма колебаний голосовых складок также подвержена изменениям.

Длительность интервала, в течение которого складки соприкасаются друг с другом, оказывает большое влияние на восприятие характеристик голоса. В частности, у женщин отношение длительности интервала открытой голосовой щели к длительности интервала закрытой голосовой щели меньше, чем у мужчин. При ударе складок происходят сложные явления, рассчитать которые довольно трудно, если не прибегнуть к упрощениям.

В [59] параметры удара были определены в предположении, что на процесс удара оказывает влияние только первая мода колебаний. В этих условиях частота первой моды определяется как

$$\omega_{1c}=\frac{4(E+N)}{\rho_т}\left\{\frac{h_г}{b_г}+J_y l_г\left[\frac{\alpha_1}{8X_1(0{,}5l_г)}\right]^2\right\}, \qquad (5.21)$$

где α_1 то же, что и в (5.17). Если выбирается такая схема управления голосовым источником, в которой изменяется лишь напряжение голосовых мышц N, то все остальные параметры в (5.21) постоянны и ω_{1c} вычисляется очень просто. При управлении натяжением складок отношение $h_г/b_г$ остается постоянным, а меняются лишь E, $\rho_т$ и J_y. Во всех случаях α_1 и $X_1(l_г/2)$ неизменны, и вычисление ω_{1c} не представляет никакого труда.

Длительность удара и начальные условия для расходящихся складок определяются решением уравнения

$$z''+\frac{2r_{эк}}{\rho_т}z'+\omega_c^2 z=-\frac{p_0 h_{гс}}{\rho_т},$$

где z — смещение центра масс каждой складки после удара, $r_{эк}$ — эквивалентный коэффициент вязкого трения, который возрастает с момента начала соприкосновения складок, p_0 — подсвязочное давление. Этот коэффициент зависит и от площади соприкосновения, и от z.

Начальными условиями для z в момент удара служат $z(0)=0$, $z'(0)=U'(l_г/2,\ T_{от})$, т. е. $z'(0)$ равно скорости движения середины голосовых связок, где

$$U'(l_г/2,\ T)=\sum_{i=1}^{3}X_i(l_г/2)\,T_i'(t),$$

$T_{от}$ — длительность интервала открытой голосовой щели, начало отсчета которого происходит в момент $z=0$, $z'>0$. Величина $z'(\pi/\omega_{1c})$ служит начальным условием для колебаний на частоте первой моды. Начальные условия по скорости для второй и третьей мод определяются при неизменных частотах ω_2 и ω_3, но с увеличенным коэффициентом затухания, в результате чего скорость колебаний в момент расхождения складок всегда меньше, чем их скорость в момент удара.

Если голосовые складки различаются по механическим и геометрическим характеристикам, то их моды также различны, и процессы колебаний нужно рассчитывать для каждой складки. Тогда площадь голосовой щели

$$S_{\text{г}}(t)=\int_0^{l_{\text{г}}}\left[U_1(x,\,t)-U_2(x,\,t)\right]dx,$$

где U_1 и U_2 — форма каждой складки. Неодинаковость складок создает дополнительные вариации формы колебаний. Она особенно важна при исследовании патологии голосового источника с помощью синтезаторов. Кроме этого, она влияет и на восприятие индивидуальности голоса. При идентичных складках достаточно рассчитать процессы колебаний только для одной складки и если складки разведены настолько, что их колебания не сопровождаются соприкосновением, то площадь голосовой щели есть

$$S_{\text{г}}(t)=2\sum_{i=1}^{3}T_i(t)\int_0^{l_{\text{г}}}X_i(x)\,dx+S_{\text{г}0},$$

где $S_{\text{г}0}$ — площадь щели в нейтральном состоянии.

В этом, наиболее простом случае, собственные функции не нужно ни хранить, ни рассчитывать, так как интегралы от каждой функции есть величины постоянные, они могут быть вычислены один раз, и $S_{\text{г}}(t)=2\sum_{i=1}^{3}c_iT_i(t)$, где

$$c_i=\int_0^{l}X_i(x)\,dx=\text{const},$$

причем для антисимметричной второй собственной функции $c_2=0$, и в простейшем случае эта мода не влияет на величину площади голосовой щели.

Поскольку собственные функции X_i ортонормированны, то даже при изменении длины складок $l_{\text{г}}$ не нужно пересчитывать интегралы — изменяется лишь коэффициент a_{1i}. Вычисление незначительно усложняется, если в нейтральном состоянии задние концы складок разведены на расстояние z_l и колебания площади голосовой щели происходят относительно некоторой постоянной составляющей $S_{\text{гп}}=z_ll_{\text{г}}/2$. В этом случае к функции $U^{(0)}(x,\,t)$, описывающей форму каждой голосовой складки, добавляется линейная функция от x: $U(x,\,t)=U^{(0)}(x,\,t)+\alpha(t)\,x$, где $\alpha(t)=z_l(t)/2l_{\text{г}}$.

Сложнее обстоит дело, когда складки соприкасаются не всей поверхностью, а лишь на отдельных участках. Тогда интеграл берется только для тех участков, где $U^{(0)}(x,\,t)>\alpha(t)\,x$, т. е. нужно выполнить операцию сравнения $u^{(0)}$ в каждой точке x с порогом αx, где в частном случае $\alpha=0$. В силу нелинейности этой операции уже нельзя порознь интегрировать

собственные функции, а нужно сначала их сложить с текущими амплитудами:

$$S(t)=2\int_0^{l_г}\sigma\left[\sum_{i=1}^{3}X_i(x)\,T_i(t)-\alpha(t)\,x\right]dx, \qquad (5.22)$$

где функция σ может быть названа характеристической, так как

$$\sigma=\begin{cases}1, & \xi>0,\\ 0, & \xi\leqslant 0.\end{cases}$$

Пользуясь функцией σ, найдем те участки, где складки соприкасаются, и вычислим возникающие силы удара как

$$F_{уд}(x,\,t)=\begin{cases}mU^{(0)''}+rU^{(0)'}+cU^{(0)}, & \sigma=1,\\ 0, & \sigma=0,\end{cases}$$

где m, r и c — погонные масса, коэффициент трения и упругость складки.

При решении системы уравнений (5.15), описывающих неакустические процессы колебания давления и объемной скорости воздушного потока в речевом тракте, для одномерной модели голосового источника используется (5.22), где коэффициенты T_i при собственных функциях вычисляются как

$$T_i''+2g_iT_i'+\omega_i^2T_i=F_i(t)-F_{уд\,i}(t),$$

а сила $F_i(t)$ определяется, как было показано выше. Одномерная модель обеспечивает достаточно высокую натуральность синтетической речи. С ее помощью можно моделировать явления физиологического тремора мышц путем добавления случайной компоненты к натяжению N и длине голосовых складок $l_г$:

$$N(t)=N^{(0)}(t)+\alpha_1\xi(t),$$
$$l_г(t)=l_г^{(0)}(t)+\alpha_2\xi(t),$$

где $\xi(t)$ — случайная функция с гауссовским распределением, пропущенная через фильтр второго порядка с центральной частотой 10—15 Гц. Коэффициент α может изменяться в небольших пределах, $\alpha=0{,}02$—$0{,}04$.

Основной недостаток одномерной модели состоит в том, что она не описывает изменение массы воздуха в голосовой щели из-за изменения формы щели вдоль оси речевого тракта. Этот недостаток, однако, можно частично исправить специальными приемами. Например, задаваясь формой щели в виде конуса с переменным наклоном стенок, и пользуясь максимальным значением площади щели на предыдущем периоде $S_м$, определим дополнительную массу воздуха в голосовой щели как

$$m_в=\rho_0h_г\left[S_м-S(t)\right]/2.$$

Этот прием позволяет приблизить форму импульса голосового источника к той, которая наблюдается в естественной речи.

Двухмерная модель. Аналогично тому, как был осуществлен переход от одномассовой сосредоточенной модели голосовых складок к двухмассовой, можно перейти от одномерной распределенной модели к двум упруго связанным одномерным системам.

Пусть каждая голосовая складка состоит из двух распределенных упругих элементов, смещающихся относительно друг друга. Тогда система уравнений для одной складки есть

$$J_{y1}(E_1+N_1)\frac{\partial^4 U_1}{\partial x^4}-N_1 S_1\frac{\partial^2 U_1}{\partial x^2}+c_{n1}U_1+c_3(U_1-U_2)+$$
$$+r_1\frac{\partial U_1}{\partial t}+r_3\frac{\partial(U_1-U_2)}{\partial t}+\rho_{r1}\frac{\partial^2 U_1}{\partial t^2}=F_1(x,\ t),$$

$$J_{y2}(E_2+N_2)\frac{\partial^4 U_2}{\partial x^4}-N_2 S_2\frac{\partial^2 U_2}{\partial x^2}+c_{n2}U_2+c_3(U_2-U_1)+$$
$$+r_2\frac{\partial U_2}{\partial t}+r_3\frac{\partial(U_2-U_1)}{\partial t}+\rho_{r2}\frac{\partial^2 U_2}{\partial t^2}=F_2(x,\ t),$$

где индексы 1 и 2 относятся к первому и второму упругому элементу, c_3 — коэффициент упругой связи между элементами, r_3 — коэффициент вязкого трения для относительного движения элементов. Перенося в правую часть каждого уравнения члены с переменной другого уравнения, получим

$$J_{y1}(E_1+N_1)\frac{\partial^4 U_1}{\partial x^4}-N_1 S_1\frac{\partial^2 U_1}{\partial x^2}+(c_{n1}+c_3)U_1+(r_1+r_3)\frac{\partial U_1}{\partial t}+$$
$$+\rho_{r1}\frac{\partial^2 U_1}{\partial t^2}=\bar{F}_1(x,\ t),$$

$$J_{y2}(E_2+N_2)\frac{\partial^4 U_2}{\partial x^4}-N_2 S_2\frac{\partial^2 U_2}{\partial x^2}+(c_{n2}+c_3)U_2+(r_2+r_3)\frac{\partial U_2}{\partial t}+$$
$$+\rho_{r2}\frac{\partial^2 U_2}{\partial t^2}=\bar{F}_2(x,\ t),$$

где

$$\bar{F}_1(x,\ t)=F_1(x,\ t)+c_3U_2+r_3\frac{\partial U_2}{\partial t},$$

$$\bar{F}_2(x,\ t)=F_2(x,\ t)+c_3U_1+r_3\frac{\partial U_1}{\partial t}.$$

Если упругие элементы складок выбраны одинаковыми по своим геометрическим размерам, то все параметры в системе уравнений также одинаковы, т. е. $J_{y1}=J_{y2}=J_y$, $S_1=S_2=S$ и т. д.

Разделяя переменные $U_1=X_1(x)T_1(t)$ и $U_2=X_2(x)T_2(t)$, получим систему обыкновенных дифференциальных уравнений:

$$J_y\frac{d^4X_1}{dx^4}-\frac{NS}{E+N}\frac{d^2X_1}{dx^2}-\alpha_k^2X_1=0,$$

$$J_y \frac{d^4 X_2}{dx^4} - \frac{NS}{E+N} \frac{d^2 X_2}{dx^2} - \alpha_k^2 X_2 = 0,$$

$$\rho_т \frac{d^2 T_1}{dt^2} + (r + r_3) \frac{dT_1}{dt} + (c_п + c_3 + \alpha_k^2) T_1 = \bar{F}_{1k}(t),$$

$$\rho_т \frac{d^2 T_2}{dt^2} + (r + r_3) \frac{dT_2}{dt} + (c_п + c_3 + \alpha_k^2) T_2 = \bar{F}_{2k}(t), \qquad (5.23)$$

где

$$\bar{F}_{1k}(t) = \frac{2}{l_г} \int_0^{l_г} \left\{ F_1(x, t) + \right.$$

$$\left. + \sum_{i=1}^{3} \left[c_3 X_{2i}(x) T_{2i}(t) + r_3 X_{2i}(x) T'_{2i}(t) \right] \right\} X_{1k}(x)\, dx,$$

$$\bar{F}_{2k}(t) = \frac{2}{l_г} \int_0^{l_г} \left\{ F_2(x, t) + \right.$$

$$\left. + \sum_{i=1}^{3} \left[c_3 X_{1i}(x) T_{1i}(t) + r_3 X_{1i}(x) T'_{1i}(t) \right] \right\} X_{2k}(x)\, dx.$$

Больше того, если элементы каждой складки полностью идентичны, то и их собственные функции одинаковы, т. е. $X_{1i}(x) = X_{2i}(x)$, и вследствие ортонормированности этих функций выполняется равенство:

$$\frac{2}{l_г} \int_0^{l_г} X_{1i}(x) X_{2j}(x)\, dx = \begin{cases} 0, & i \neq j, \\ 1, & i = j, \end{cases}$$

отсюда

$$\bar{F}_{1k}(t) = \frac{2}{l_г} \int_0^{l_г} F_1(x, t) X_k(x)\, dx + c_3 T_{2k}(t) + r_3 T'_{2k}(t),$$

$$\bar{F}_{2k}(t) = \frac{2}{l_г} \int_0^{l_г} F_2(x, t) X_k(x)\, dx + c_3 T_{1k}(t) + r_3 T'_{1k}(t).$$

Таким образом, мы получим систему уравнений, очень похожую на систему для двухмассовой модели:

$$\rho_т T''_{1i} + (r + r_3) T'_{1i} + (c_п + c_3 + \alpha_i^2) T_{1i} - r_3 T'_{2i} - c_3 T_{2i} = F_{1i}(t),$$

$$\rho_т T''_{2i} + (r + r_3) T'_{2i} + (c_п + c_3 + \alpha_i^2) T_{2i} - r_3 T'_{1i} - c_3 T_{1i} = F_{2i}(t),$$

где

$$F_{1,2}(t)=\frac{2}{l_r}\int_0^{l_r} F_{1,2}(x,\ t)\,X_i(x)\,dx.$$

Силы $F_1(x,\ t)$ и $F_2(x,\ t)$ — те же, что и в двухмассовой модели, только они распределены вдоль голосовых складок. Все проблемы, связанные со столкновением отдельных участков голосовых складок, решаются так же, как и в одномерной распределенной модели.

Модель, состоящая из двух распределенных элементов обладает тем же преимуществом перед одномерной распределенной моделью, что и двухмассовая модель перед одномассовой — она описывает такое принципиально важное явление, как переменная масса воздуха в голосовой щели. Эта модель несколько сложнее, чем одномерная модель, однако, учитывая тот факт, что ее уравнения включены в систему (5.15), состоящую из 14 уравнений, можно сказать, что относительное возрастание сложности не слишком велико. Вместе с тем, эта модель реализует все основные процессы голосообразования, и с ее помощью можно достичь натуральности синтетической речи, имитировать дикторскую индивидуальность и исследовать патологические явления.

Естественно сделать следующий шаг в развитии моделей голосового источника — перейти к описанию непрерывных деформаций голосовой складки вдоль оси речевого тракта, т. е. попытаться решить систему уравнений (5.17), (5.18), (5.19). Основная трудность в поиске решения состоит в сложной форме складки и неполностью известных граничных условиях. В [59] исследовались следующие граничные условия: $Y(0)=Y''(0)=0$, т. е закрепление типа шарнирного на нижней кромке голосовой складки, и

$$\gamma Y(l_y)+(1-\gamma)\,Y''(l_y)=0,\quad Y'''(l_y)=0,$$

т. е. отсутствие перерезывающих сил и комбинации сил, сопротивляющихся смещению и повороту верхней кромки складки. Исходя из этих граничных условий, найдем собственные функции $Y(y)$ для деформации поверхности складки вдоль оси y:

$$\bar{Y}_j(y)=b_j\left(\operatorname{sh} q_{1j}y+\frac{q_{1j}^3}{q_{2j}^3}\frac{\operatorname{ch} q_{1j}l_y}{\cos q_{2j}l_y}\sin q_{2j}y\right),$$

где $\bar{Y}_j(y)$ — ненормированные собственные функции. Известно, что

$$\int_0^{l_y}\bar{Y}_j^2(y)\,dy=\frac{l_y}{4}\left(\bar{Y}_j^2-2\bar{Y}_j'\bar{Y}_j'''^2+Y_j''^2\right)\Big|_{y=l_y}.$$

Учитывая граничные условия, получим

$$b_j = \frac{1}{\overline{Y}_j(l_y)} \sqrt{\frac{2}{1+\left(\frac{1-\gamma}{\gamma}\right)^2}}.$$

В [59] коэффициент γ был принят равным 0,1, т. е. граничные условия на верхней кромке складки почти соответствуют свободному концу. Представляется, что для описания поперечных деформаций голосовых складок достаточно двух собственных функций.

Форма голосовой складки с учетом поперечных деформаций описывается рядом

$$U(x, y, t) = \sum_{i,\ j} X_i(x)\, Y_j(y)\, T_{ij}(t),$$

где T_{ij} есть решение дифференциального уравнения

$$T''_{ij} + 2g_{ij}T'_{ij} + \omega_{ij}^2 T_{ij} = F_{ij}(t),$$

и

$$\omega_{ij}^2 = \frac{c_{\text{п}} + (J_y\alpha_i^2 + J_x\beta_j^2)}{\rho_{\text{т}}}.$$

Принимая, что жесткость подстилающего слоя такая же, как и жесткость поверхностного слоя складки, можно записать

$$\omega_{ij}^2 = \frac{1}{\rho_{\text{т}}}\left[\frac{ES_{\text{п}}}{l_{\text{т}}} + (p_{1i}p_{2i}J_y + q_{1j}q_{2j}J_x)(E+N)\right],$$

где $S_{\text{п}}$ — площадь поперечного сечения подстилающего слоя. Для трапецевидного сечения центральный момент инерции J_y (т. е. относительно оси y, проходящей через центр тяжести сечения) при $h_{\text{ск}} = h_{\text{г}}$ есть

$$J_y = \frac{13}{108} h_{\text{г}} b_{\text{г}}^3.$$

Момент инерции относительно оси x, есть

$$J_x = \frac{11}{81} l_{\text{г}} b_{\text{г}}^3.$$

Вместо трех компонент, имевшихся в одномерной модели, теперь форма голосовых складок определяется шестью компонентами. Возбуждающие силы для каждой компоненты находим как

$$F_{ij}(t) = \frac{4}{\rho_{\text{т}} l_y l_{\text{г}}} \int_0^{l_y}\int_0^{l_{\text{г}}} F(x, y, t)\, X_i(x)\, Y_j(y)\, dx\, dy.$$

Если пренебречь изменением формы голосовой щели вдоль

оси речевого тракта и принять $F(x, y, t)=F(t)$, т. е. допустить равномерное распределение сил по поверхности складок, то двойной интеграл берется, и для каждой комбинации ij он равен некоторой постоянной величине. Для фиксированной длины голосовых складок $l_г=\text{const}$ эти интегралы вычисляются лишь один раз и в процессе синтеза вычисления не повторяются. Для того чтобы полностью реализовать преимущества двухмерной модели, необходимо рассчитать распределенные силы давления внутри голосовой щели. С этой целью приходится решать систему уравнений (5.23). Возможны, конечно, и *упрощенные модели*. Например, можно принять, что голосовая щель имеет площадь, равную максимальной площади проекции голосовой щели на плоскость, перпендикулярную оси речевого тракта, т. е.

$$S_г(t)=\min_y S_г(y, t)=\min_y 2\int_0^{l_г}\sigma\left[\sum_{i, j} X_i(x)\, Y_j(y)\, T_{ij}(t)-\alpha(t)\, x\right]dx,$$

где σ — характеристическая функция, равная 1 или 0, а дополнительную массу воздуха в голосовой щели отсюда найдем, как

$$m_в=\rho_0\left[2\int_0^{l_1}\int_0^{l_y} U(x, y, t)\,dx\,dy-S_г(t)\,l_г\right].$$

Итак, мы располагаем различными распределенными моделями голосовых складок, и в зависимости от задачи и имеющихся вычислительных средств можем использовать ту или иную модель. Распределенные модели, конечно, сложнее, чем сосредоточенные — двухмассовая и, тем более, одномассовая. Однако без распределенных моделей невозможно достичь требуемой натуральности и гибкости в управлении характеристиками голосового источника возбуждения.

§ 5.6. Поршневой источник

Поршневой источник возбуждения создается упругими деформациями голосовых складок вдоль оси речевого тракта и, *таким образом, строгая модель колебаний складок должна* быть трехмерной. Однако вертикальные колебания складок в значительной степени независимы от упругих деформаций в остальных направлениях, поэтому их можно рассматривать порознь.

Представим голосовую складку как короткую консоль, пренебрегая закреплением на переднем и заднем концах. Со стороны хряща граничные условия соответствуют жесткой заделке: $z(0)=z'(0)=0$, а со стороны голосовой щели граничные условия соответствуют свободному концу, т. е. $z''(b_г)=z'''(b_г)=0$. Тогда собственные функции запи-

сываются как

$$Z_k(z)=\left[c_{1k}\left(\operatorname{ch}\frac{p_k z}{b_\text{г}}-\cos\frac{p_k z}{b_\text{г}}\right)+c_{2k}\left(\operatorname{sh}\frac{p_k z}{b_\text{г}}-\sin\frac{p_k z}{b_\text{г}}\right)\right]\sqrt{\frac{2}{b_\text{г}}},$$

где $p_1=0{,}597\pi$, $p_2=1{,}494\pi$, $c_{11}=c_{12}=0{,}707$, $c_{21}=-0{,}518$, $c_{22}=-0{,}721$.

Вынуждающая сила для каждой моды вертикальных колебаний определяется аналогично предыдущему параграфу как

$$F_k=\frac{2(p_{1\text{г}}-p_{2\text{г}})}{\rho_\text{т}}\int_0^{b_\text{г}} Z_k(z)\,dz,$$

где $p_{1\text{г}}$ и $p_{2\text{г}}$ — *подсвязочное и надсвязочное давления*, определяемые решениями системы (5.15), $\rho_\text{т}$ — погонная плотность тканей вдоль оси z. Как видно, интеграл от $Z_k(z)$ есть величина постоянная, не изменяющаяся в процессе голосообразования. Вертикальное смещение складки получается как сумма решений обыкновенных дифференциальных уравнений

$$T_k''+2g_k T_k'+\omega_k^2 T_k=F_k,$$

где собственные частоты

$$\omega_k=\frac{p_k}{b_\text{г}}\sqrt{\frac{J_z(E+N)}{\rho_\text{т}}}=p_k\sqrt{\frac{E+N}{12\rho_{\text{т}0}}}$$

зависят от начального модуля упругости E и натяжения складок N.

Амплитуда вертикальных смещений голосовых складок в 2—4 раза больше амплитуды горизонтальных смещений, а максимальное отклонение смещений по вертикали от смещений по горизонтали сдвигается по фазе примерно на $\pi/2$, *вследствие чего каждая точка на верхней поверхности* в процессе фонации описывает эллипс, у которого вертикальная ось в 2—4 раза больше горизонтальной.

Объемная скорость воздушного потока, создаваемого вертикальными *колебаниями складок, есть*

$$w_\text{п}(t)=\sum_k T_k(t)\,l_\text{г}\int_0^{b_\text{г}} Z_k(z)\,dz.$$

Поршневой источник порождает сигнал возбуждения акустических колебаний, равный производной по времени от объемной скорости $w_\text{п}(t)$.

На рис. 5.23 показаны *максимальные горизонтальные* и вертикальные смещения голосовых складок, объемная скорость от голосового и поршневого источников, а также суммарный сигнал голосового источника (производная от объемной скорости воздушного потока в голосовой щели и поршневого источника). Вследствие того, что объемная скорость поршне-

вого источника создается только вертикальными движениями складок, форма импульсов поршневого источника гораздо более гладкая, чем импульсов голосового источника. Поэтому вклад поршневого источника в возбуждения высших формант

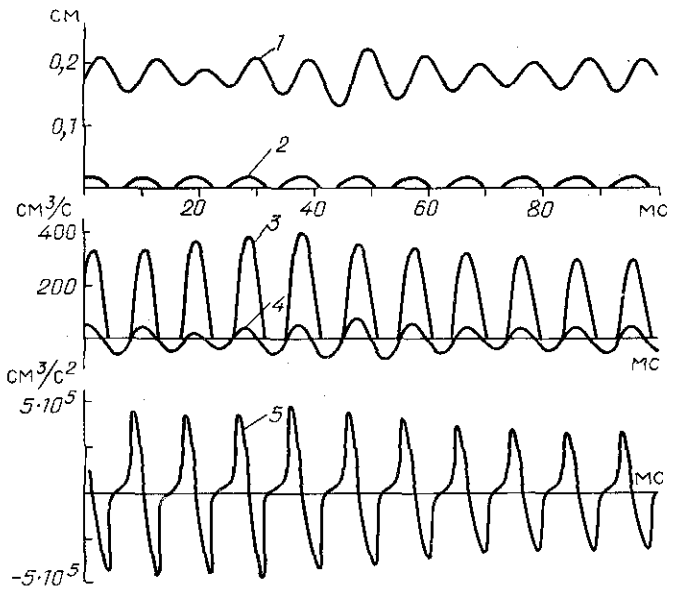


Рис. 5.24. Голосовой и поршневой источники: *1* — вертикальное смещение голосовых складок, *2* — горизонтальное смещение, *3* — объемная скорость в голосовой щели, *4* — объемная скорость поршневого источника, *5* — производная суммарной объемной скорости. Частота первой гармоники вертикальных колебаний $F_0 = 80$ Гц

невелик. Однако поршневой источник может уменьшить амплитуду акустических колебаний низкочастотных резонансов из-за того что поршневой источник действует со сдвигом фазы во времени относительно голосового источника. Если этот сдвиг попадает в интервал времени, находящийся между τ_i и $\tau_i/2$, где τ_i — период акустических колебаний, то амплитуда колебаний уменьшается, что равносильно относительному увеличению уровней высших формант.

Фаза вертикальных колебаний голосовых складок относительно их горизонтальных колебаний зависит от механических и геометрических параметров складок. На высоких собственных частотах вертикальных колебаний максимум объемной скорости поршневого источника приходится на интервал сомкнутой голосовой щели, как это показано на рис. 5.24. На низких собственных частотах этот максимум смещается в область открытой голосовой щели (рис. 5.25). Разнообразие форм объемной скорости для разных факторов на интервале сомкнутой голосовой щели наблюдается на сигналах, полученных

методом обратной фильтрации. Поскольку изменение жесткости складок из-за натяжения и напряжения голосовых мышц влияет на собственные частоты вертикальных колебаний, то можно ожидать изменения амплитуды и сигнала от поршневого источника у одного и того же диктора при изменении частоты основного тона.

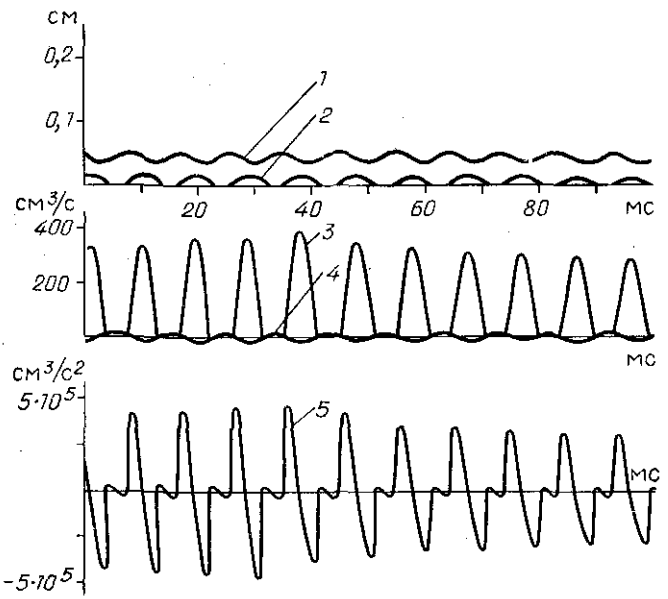


Рис. 5.25. То же, что и на рис. 5.24. Частота первой гармоники вертикальных колебаний $F_0 = 160$ Гц

Вариации параметров поршневого источника влияют на воспринимаемый тембр голоса [59], и, таким образом, поршневой источник является одним из средств создания индивидуальности синтетического голоса.

§ 5.7. Импульсный источник

Если в течение некоторого времени площадь поперечного сечения в речевом тракте где-либо выше голосовой щели равнялась нулю, то давление в тракте возрастает, а после взрыва смычки быстро падает, создавая так называемый импульсный источник возбуждения. Механика этих процессов детально описана в [59], а также в предыдущих разделах этой главы, поэтому в данном разделе мы рассмотрим в основном только свойства самого импульсного источника. Процессы накопления и спада давления в речевом тракте описываются системой (5.15), и зависят, главным образом,

от сопротивления голосовой щели и сопротивления в месте наибольшего сужения в тракте. Другие параметры, такие как податливость стенки тракта, объем тракта, сопротивление в легких и т. д., также влияют на характеристики импульсного источника.

Рассмотрим искусственную ситуацию, показанную на рис. 5.26. Здесь площадь голосовой щели, начально равная

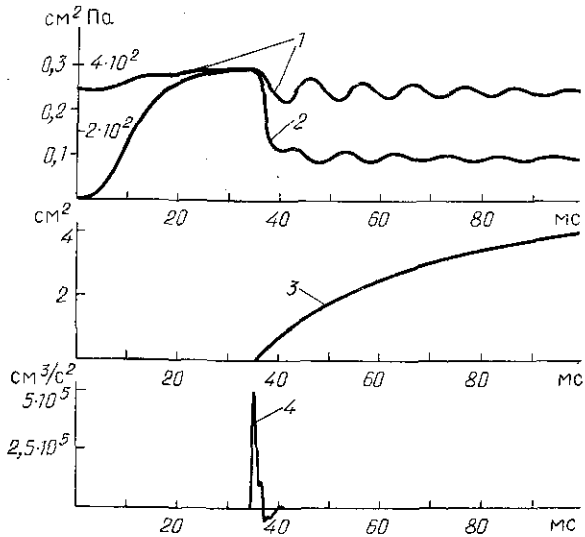


Рис. 5.26. Импульсный источник: *1* — площадь голосовой щели, *2* — подсвязочное давление, *3* — площадь сужения в речевом тракте, *4* — производная от объемной скорости воздушного потока

0,3 см^2, не уменьшается до нуля в области взрыва сигнала, а сохраняется постоянной с некоторыми колебаниями, вызванными взрывом. Подсвязочное давление во время смычки растет от нуля до $4 \cdot 10^2$ Па, а затем быстро спадает после взрыва смычки. Производная от объемной скорости воздушного потока в месте наибольшего сжатия речевого тракта и является сигналом импульсного источника возбуждения. Амплитуда этого сигнала определяется скоростью изменения давления (или объемной скорости) и сравнима с амплитудой голосового источника. Сигнал имеет две фазы — с положительным и отрицательным значениями, причем амплитуда положительных значений значительно больше амплитуды отрицательных. Длительности положительной и отрицательной фаз примерно одинаковы и равны в данном случае 80 мс.

В реальной речи характеристики импульса несколько отличаются от тех, которые представлены на рис. 5.26. Во время звонкой смычки колебания складок не прекращаются и момент взрыва может прийтись на интервал открытой либо интервал

закрытой голосовой щели. В первом случае импульсный источник действует либо в фазе, либо в противофазе с голосовым источником. Во втором случае импульсный источник существует отдельно лишь короткое время до раскрытия голосовой щели. В русской речи импульсный источник довольно редко наблюдается на звонких взрывных.

Перед взрывом глухой смычки голосовые складки сближаются и в зависимости от фазы их движений относительно движений основного артикулятора меняется время от сигнала импульсного источника до первого импульса голосового источника. Как указывалось в гл. 4, среднее время между взрывом и началом фонаций примерно одинаково для /П/ и /Т/ (около 15 мс) и заметно больше для /К/ (от 20 до 45 мс). При этом для взрыва /К/ характерно наличие двух или более импульсов, следующих друг за другом через 4—24 мс. Это явление, скорее всего, связано с податливостью языка и мягкого неба и сравнительно медленным движением языка, в результате чего площадь тракта в месте наибольшего сужения колеблется аналогично тому, как колеблются голосовые складки после взрыва на рис. 5.26.

Другая возможная причина появления повторных импульсов после взрыва /К/ заключается в изменении объема воздуха, участвующего в формировании упругости (емкости) речевого тракта. При заднеязычной смычке объем речевого тракта между голосовой щелью и смычкой примерно вдвое меньше общего объема. После взрыва в создании упругости воздуха участвует и объем, заключенный между местом наибольшего сужения и губами. Поэтому в электрической схеме для процессов по постоянному току для заднеязычных согласных необходимо включать дополнительную емкость последовательно с сопротивлением речевого тракта.

Рассмотрим, как импульсный источник влияет на характеристики речевого сигнала. Пусть отклик речевого тракта на частоте k-го резонанса описывается обыкновенным дифференциальным уравнением второго порядка. Тогда возбуждающий сигнал есть

$$F_k(t)=\frac{2\int_0^t F(x,t)S(x,t)\psi_k(x)\,dx}{\rho_0 l\int_0^l S(x,t)\psi_k^2(x)\,dx},$$

где x — координата вдоль оси речевого тракта, l — длина тракта, ψ_k — собственная функция акустических колебаний, $F(x,t)$ — распределенный источник возбуждения, $S(x,t)$ — площадь поперечного сечения речевого тракта. Импульсный источник сосредоточен в месте наибольшего сужения и может быть представлен с помощью функции Дирака как $F(x,t)=f(t)\delta(x-x_0)$, где δ — дельта-функция ($\delta=\infty$ при $x_0=x$

и $\delta=0$, $x_0 \neq x$), x_0 — координата наибольшего сужения. Тогда возбуждающий сигнал для k-го резонанса есть

$$F_k(t)=\frac{2}{\rho_0 l} f(t) S(x_0, t) \psi_k(x_0).$$

Отсюда видно, что если x_0 совпадает с одним из нулей собственной функции ψ_k, то и возбуждение k-го резонанса равно нулю, и k-я форманта отсутствует в спектре речевого сигнала в момент взрыва. В общем случае $x \neq x_0$, и тогда амплитуда отклика и фаза колебаний на k-м резонансе зависят от положения x_0 относительно нулей и знака собственной функции ψ_k. Из проведенного анализа следует, что влияние импульсного источника на характеристики речевого сигнала реализуется автоматически в артикуляторном синтезаторе, может быть легко рассчитано в артикулярно-формантном синтезаторе и требует детальной спецификации для каждого звука в формантном синтезаторе. Тем не менее, для формантного синтезатора нетрудно найти аппроксимацию формы сигнала от импульсного источника и изменять в зависимости от места артикуляции его амплитуду и интервал времени до начала работы голосового источника. Для согласного /К/ может потребоваться двойной импульс возбуждения; что же касается частот и амплитуд формант в момент включения импульсного источника, то они должны быть подобраны для каждого сочетания «согласный-гласный». Следует учитывать также, что при достаточно большой площади голосовой щели в момент взрыва в спектре речевого сигнала присутствуют дополнительные форманты вследствие того, что фактическая длина речевого тракта в это время увеличена почти вдвое за счет подсвязочной области. Наиболее заметен в спектре взрыва согласного /П/ резонанс на частоте около 1700 Гц и поэтому в формантном синтезаторе может оказаться необходимым использование дополнительных формант в момент взрыва. Как мы увидим в следующем разделе, точно такая же ситуация возникает и для глухих фрикативных.

§ 5.8. Турбулентный источник

Известно свойство воздушных струй генерировать звук, однако протекающие при этом процессы настолько сложны, что теория аэроакустики имеет полуэмпирический характер, т. е. многие зависимости и коэффициенты устанавливаются экспериментально, а теоретические положения заимствуются из описания струй, вытекающих из некоторого сопла в свободную среду, или расчета обтекания различных объектов. В речи существуют звуки, фонетическое качество которых определяется наличием и спектральными характеристиками шумов в речевом тракте. До сих пор при описании процессов шумообразования

в речевом тракте учитывался только эффект турбулизации потока при числах Рейнольдса, превышающих 1800. Однако и при достаточно низких (докритических) числах Рейнольдса, больших 100, воздушный поток неустойчив к возмущениям [12] и генерирует шумы типа шипения, спектр которых имеет максимум на частоте

$$\omega_д = 0{,}11\,\pi\,\frac{v}{h}. \tag{5.25}$$

Другая оценка [1], учитывающая преобладание вязкости, дает

$$\omega_д = \mathrm{sh}\,\frac{\nu}{h^2},$$

где sh — число Струхаля, ν — кинематическая вязкость, ω — круговая частота. Такие колебания возникают в очень узких трубах, где доминирует капиллярное трение, в частности на этапах открытия и закрытия голосовой щели. В силу этого явления можно ожидать появления случайной компоненты в импульсе объемной скорости голосового источника. Напомним, что число Рейнольдса определяется, как

$$\mathrm{Re} = \frac{\rho_0}{\mu} vh,$$

где ρ_0 — плотность воздуха, μ — коэффициент вязкости воздуха, v — скорость воздушного потока, h — характерный геометрический размер канала или объекта, обтекаемого струей: $h \approx 4S/L$, S — площадь поперечного сечения, L — периметр сечения [33]. Очевидно, что для круглого сечения h равно его диаметру.

Оценим частоту докритического шума в речевом тракте, приняв сечение круговым и перейдя к объемной скорости $w = vS$. Тогда циклическая частота

$$f_д = 0{,}055\,\frac{w}{Sd},$$

где d — диаметр сечения. Допустим, что объемная скорость $w = 500$ см^3/с, а площадь сечения $S = 1$ см^2. Тогда $f_д \approx 24$ Гц, что, конечно, является очень низким значением. Следует иметь в виду, что оценка средней частоты шума по (5.24) относится к истечению струй в свободное пространство, а в применении к речевому тракту может потребоваться коррекция этой зависимости. Экспериментально также следует определить и уровень генерируемых шумов.

Докритические шумы непрерывно распределены вдоль речевого тракта, поэтому их вклад в возбуждение того или иного резонанса следует рассчитывать по (5.24) из предыдущего раздела. Источник такого шума $\xi(x, t)$ может быть получен

как решение обыкновенного дифференциального уравнения

$$\xi'' + 2g\xi' + w^2\xi = F(x, t)\eta(t), \qquad (5.26)$$

где η — случайная функция, равномерно распределенная на интервале $[0,1]$, $F(x, t)$ — интенсивность шума в точке тракта с координатой x. В результате такого воздействия амплитуда каждой форманты в речевом сигнале будет иметь аналогичную случайную компоненту.

Как только число Рейнольдса достигает критической величины, характер воздушного потока резко меняется — возникает турбулентность. Для расчета турбулентных явлений в речевом тракте обычно пользуются результатами обтекания потоком некоторого тела. При этом возникают случайные колебания с частотами

$$f_n = \mathrm{sh}\,(\mathrm{Re})\frac{v}{d}n,$$

где $\mathrm{sh}\,(\mathrm{Re})$ — число Струхаля. Для круглого сечения $\mathrm{sh} \approx 0{,}2$ при $10^3 < \mathrm{Re} < 3 \cdot 10^4$, а для пластины $\mathrm{sh} = 0{,}16 - 0{,}18$ при $10^3 < \mathrm{Re} < 2 \cdot 10^5$ [2]. Амплитуда первого обертона оценивается как

$$A_1 = k_1(\mathrm{Re}^2 - \mathrm{Re}_{\text{кр}}^2),$$

где k_1 — коэффициент порядка $10^{-5} - 10^{-6}$ в зависимости от расположения источника в речевом тракте, и амплитуда второго обертона, по [2], в десять раз меньше.

Как было показано в [59], число Рейнольдса превышает критическую величину $\mathrm{Re}_{\text{кр}}$ лишь в узких областях. Один турбулентный источник всегда сопровождает колебания голосовых складок и расположен на выходе из голосовой щели. Другие турбулентные источники зависят от степени сужения и скорости воздушного потока в тракте.

Оценим параметры турбулентного источника на выходе из голосовой щели, аппроксимируя форму каждой голосовой складки в виде равнобедренного треугольника с длиной $l_г$ и высотой $h/2$.

Тогда, учитывая, что максимальное отклонение складок $h/2$ от положения равновесия гораздо меньше длины складок $l_г$, для числа Рейнольдса в голосовой щели получим

$$\mathrm{Re}_г = 12{,}5\, w/l_г.$$

Число $\mathrm{Re}_г$ превышает критическую величину при объемной скорости $w > 180 - 250$ см^3/с. Как мы видели ранее, объемная скорость в голосовой щели нарастает и спадает довольно быстро, так что в течение большей части времени, занятого открытой голосовой щелью, действует турбулентный источник.

Циклическая частота его первого обертона

$$f_{1г} = 0{,}085 \frac{w l_г}{S_г^2}.$$

При $w = 500$ см^3/с, $l_г = 1{,}5$ см, $S_г = 0{,}2$ см^2, $f_{1г} \approx 1600$ Гц. Эта частота соответствует моменту времени с максимальным раскрытием голосовой щели. Если в процессе колебаний происходит не полное смыкание голосовых складок, то турбулентный шум генерируется непрерывно, как это характерно *для женских голосов*.

В [2] приводятся измерения шумов струи, набегающей на цилиндр диаметра 0,5 см со скоростью 33 м/с, где частота первого обертона равнялась 1330 Гц, и ширина первого спектрального максимума была 275 Гц. Характерный геометрический размер препятствия близок к размерам щелей в речевом тракте, поэтому частоты первых обертонов в том и другом случае оказываются близкими. В речевом тракте ширина спектра шума гораздо больше, если судить по частотным характеристикам фрикативных звуков, приближенно ее можно представить, как $g \approx 0{,}5\,\omega$.

Турбулентный источник является источником давления. Поэтому давление над голосовой щелью изменяется не только вследствие процессов по постоянному току, которые рассматривались в п. 5.4.4, но также из-за вихревого движения воздушного потока. Это случайная компонента надсвязочного давления $p_т$ должна учитываться при решении системы (5.15).

Частоты обертонов турбулентного шума примерно одинаковы для разных фрикативных звуков и мало зависят от места артикуляции. Однако спектры речевых сигналов этих звуков сильно различаются, причем наблюдается эффект полосового усиления или ослабления. Это является результатом обсуждавшегося в предыдущем разделе свойства источников, сосредоточенных в узкой пространственной области — в спектре речевого сигнала отсутствуют те частоты, которые соответствуют собственной функции, имеющей нуль в области действия источника возбуждения.

Если аппроксимировать ромбом форму сечения щели с турбулентным источником, то для расчета характеристики шума можно воспользоваться соотношениями, полученными для голосовой щели. Так, при ширине щели 1—2 см, число Рейнольдса $\mathrm{Re} \approx (6-12)\,w$, а частота первого максимума в спектре шума $f_1 = (0{,}08-0{,}17)\,w/S_г^2$, что при реально наблюдающихся площадях $S_г = 0{,}1-0{,}4$ см^2, соответствует частотам $f_1 = (0{,}5-17)\,w$, т. е. примерно 250—8500 Гц при $w = 500$ см^3/с, что, очевидно, вполне соответствует диапазону частот фрикативных согласных.

Турбулентный источник возникает не только при артикуляции фрикативных согласных, но также и на фазах смыкания

и размыкания артикуляторных органов при образовании взрывных согласных. Звуки с шумами при подходе к смычке называются преаспиративными. Образование шума в этом случае зависит от координации движений расходящихся голосовых складок и сближения артикуляторов. В русском языке преаспирация наблюдается редко, хотя, например, в исландском это явление играет фонеморазличительную роль. При взрыве слитных глухих согласных турбулентные шумы действуют совместно с импульсным источником, давая возможность перцептивной системе слушателя оценить распределение энергии по спектру речевого сигнала, характерное для того или иного места артикуляции.

ГЛАВА 6

ФОРМАНТНЫЙ СИНТЕЗ

Измеряя акустические характеристики гласных звуков, Гельмгольц обнаружил в их спектрах максимум, находящийся на разной частоте в зависимости от звука. Он назвал этот максимум формантом. В дальнейшем в спектре звуков было найдено два и более максимумов. В русской литературе термин «формант» приобрел женский род — «форманта». Обычно форманты связываются с резонансными частотами речевого тракта, хотя частоты формант далеко не всегда позволяют определить частоты резонансов тракта — иногда они сближаются настолько, что максимумы в спектре сливаются.

Представления о модуляционной природе речевого сигнала, состоящие в том, что информация в речевом сигнале передается путем управления резонансными характеристиками речевого тракта, получили подтверждение после изобретения динамического спектрографа. Этот прибор, состоящий из гребенки полосовых фильтров, дает возможность наблюдать изменения спектра во времени в форме так называемой видимой речи или спектрограмм (сонограмм). На спектрограммах хорошо видно изменение частот и амплитуд формант, а также появление широких областей шумов турбулентного потока воздуха при артикуляции фрикативных звуков. На рис. 6.1 показана спектрограмма фразы «Ну, больше ничего, наверное, не надо говорить», на которой видим эти явления.

Следующий шаг в изучении свойств речевого сигнала был сделан в работах Хаскинских лабораторий. Используя спектрограммы, нарисованные на прозрачной подложке, и синтезируя речь с помощью суммирования выходов гребенки фильтров, на входы которых с помощью фотодатчика подавался сигнал, обратно пропорциональный степени прозрачности подложки, удалось получить весьма разборчивую речь. На основе этих экспериментов в [91] были сформулированы правила для управления частотами формант с целью синтеза звуков английской речи.

Физически ясный смысл связи между артикуляцией и резонансными частотами речевого тракта, наглядность пред-

ставления текущего спектра в виде спектрограмм и возможность генерирования разборчивой речи путем управления частотами формант послужили основой для разработки формантных синтезаторов. Многолетние усилия привели к созда-

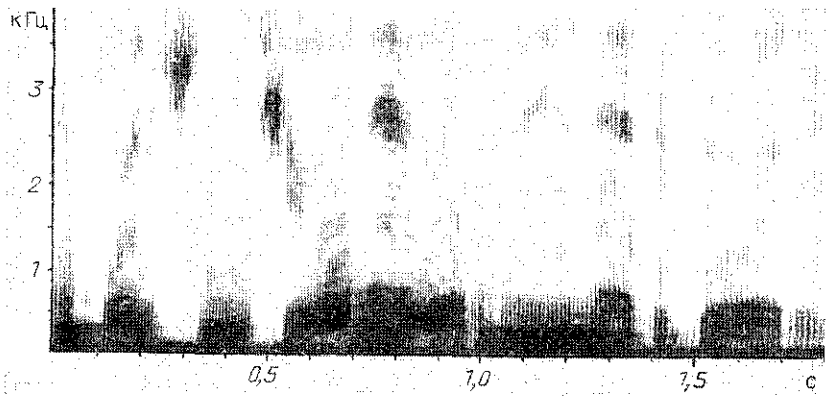


Рис. 6.1. Спектрограмма фразы «Ну, больше ничего, наверное, не надо говорить»

нию систем с высоким уровнем разборчивости, что определило широкое распространение формантных синтезаторов. Однако по мере того, как проходило наивное восхищение самой возможностью имитации человеческого голоса, становилось очевидным, что для обеспечения приемлемости синтетической речи для широкого потребителя необходимо учитывать все более и более тонкие свойства речеобразования и восприятия. При этом усилия по улучшению качества синтеза концентрируются в двух направлениях: наиболее точная имитация акустических процессов в речевом тракте и формирование правил управления синтезом с учетом коартикуляции, т. е. взаимного влияния звуков в слитном потоке речи.

Начнем описание формантных синтезаторов с анализа различных способов представления акустических процессов.

§ 6.1. Каскадная схема

В формантном синтезаторе каждой форманте в спектре речевого сигнала соответствует отдельный резонатор, обычно представляемый в виде системы второго порядка. Прежде всего нужно определить, сколько таких резонаторов требуется для обеспечения надлежащей разборчивости синтетической речи. Затем возникает вопрос о способе соединения этих резонаторов — последовательном или параллельном.

Из экспериментов по ограничению полосы частот речевого сигнала известно, что расширение полосы выше 5 кГц практически не улучшает качество речи на слух, а полоса

телефонного канала 3,4 кГц еще обеспечивает достаточную разборчивость. Это означает, что в спектре сигнала нужно сохранить не менее 5 резонансов речевого тракта. Это связано с большим затуханием выше 5 кГц, в основном из-за потерь на излучение, а также быстрого снижения уровня спектра сигнала голосового возбуждения. Кроме того, чувствительность и различительная способность слуха в этой области частот быстро падает. Изучение спектрограмм и опыты по формантному синтезу показали, что управлять необходимо первыми тремя формантами, а четвертая и пятая, расположенные в области 3—5 кГц, могут быть фиксированы. И хотя из дальнейшего выяснится необходимость управления и высшими формантами, в первом приближении их можно считать постоянными.

Вопрос о способе соединения резонаторов решается путем анализа акустической модели речеобразования. В [64] было показано, что если речевой тракт представить в виде неоднородной длинной линии, то его передаточная функция $H(s)$ записывается в виде произведения

$$H(s) = \prod_{n=1}^{\infty} \frac{s_n s_n^*}{(s-s_n)(s-s_n^*)},$$

где s — комплексная частота, s_n и s_n^* — комплексно-сопряженные полюсы длинной линии. Как известно, произведение в s-области соответствует последовательному (каскадному) соединению резонаторов.

Было обнаружено, что при каскадном соединении резонаторов огибающая спектра синтезированного сигнала имеет примерно тот же наклон, который наблюдается в спектрах гласных звуков. Это означает, что каскадная схема не требует раздельного управления амплитудами резонаторов. Это соображение оказалось решающим в те времена, когда синтез речи осуществлялся с помощью аналоговой техники и всякое сокращение объема управления было крайне необходимо.

Однако каскадная схема создает другую проблему — проблему учета влияния бесконечного числа отброшенных резонансов за пределами полосы, например, в 5 кГц. Если длина однородной акустической трубы равна 17,5 см, то ее четвертьволновые резонансы расположены через 1 кГц и величина требуемой коррекции для каскадной схемы на частоте в 5 кГц очень велика — 57 дБ [118]. В [64] приводятся формулы для поправок на высшие резонансы, но обычно коррекция осуществляется добавлением двух — трех резонаторов на частотах 5,5 кГц, 6,5 кГц и т. д. Дополнительная погрешность возникает при цифровой реализации резонаторов, если частота дискретизации равна 10 кГц. При этом вблизи частот среза (5 кГц) степень коррекции становится зависимой от расположения высшей форманты. Для исключения этого эффекта частота дискретизации должна быть равной 20 кГц.

Выбор частот корректирующих резонаторов выше 5 кГц нуждается в обосновании, поскольку не очевидно, что эти частоты должны быть равны частотам резонансов речевого тракта длиной в 17,5 см. Кроме того, известно, что в процессе артикуляции длина речевого тракта изменяется в пределах от 16 см до 20 см даже у одного и того же диктора. Диапазон изменений еще больше у разных дикторов.

Подробные исследования свойств каскадного синтезатора показали, что даже для гласных эта схема недостаточно точно описывает наклон спектра. В частности, для улучшения качества синтеза Р. Клатт вынужден был разделить гласные на два класса — передние и задние, и для одного класса использовать последовательное соединение резонаторов, а для другого — параллельное.

Каскадная схема не позволяет описать влияние нулей, возникающих при разветвлении речевого тракта и при возбуждении турбулентным или импульсным источниками. Поэтому любой формантный синтезатор должен содержать, помимо каскадной цепи, и параллельную, причем дополнительная сложность состоит в определении класса звука, т. е. в решении, в какой схеме этот звук должен синтезироваться. Например, при переходе от гласного к фрикативному согласному /С/ некоторое время продолжает действовать источник голосового возбуждения, и формантная структура гласного постепенно переходит в широкополосный шум фрикативного. Ошибки в выборе момента переключения каскадной и параллельной схем вызывают скачки в спектральной структуре речевого сигнала.

Каскадная схема не позволяет имитировать изменение голосовых усилий, проявляющееся в относительном смещении уровня высших формант (например, при подъеме уровня первой форманты на 10 дБ, уровень высших формант может увеличиваться на 30 дБ).

§ 6.2. Параллельная схема

Недостатки каскадной схемы на самом деле связаны с ее неадекватным описанием акустических свойств речевого тракта. Если к волновому уравнению для однородной трубы (с постоянным сечением)

$$\frac{\partial^2 \Phi}{\partial x^2} = \frac{1}{c_0^2} \frac{\partial^2 \Phi}{\partial t^2}$$

применить метод разделения переменных $\Phi(x, t) = X(x)\, T(t)$, то из системы

$$X'' + \lambda^2 X = 0,$$
$$T'' + \lambda^2 c_0^2 T = 0$$

находим, что решение волнового уравнения представляется

в виде суперпозиции

$$\Phi(x, t)=\sum_{m=1}^{\infty} X_m(x)\, T_m(t), \tag{6.1}$$

где

$$X_m(x)=A_m\cos\lambda_m x+B_m\sin\lambda_m x,$$
$$T_m(x)=C_m\cos c_0\lambda_m t+D_m\sin c_0\lambda_m t.$$

Потери добавляют в (6.1) экспоненту $e^{-g_m t}$ и, таким образом, акустический сигнал на выходе однородной трубы формируется как сумма затухающих колебаний на разных резонансных частотах $\omega_m=c_0\lambda_m$. Это соответствует параллельному соединению резонаторов. Тот факт, что речевой тракт является трубой с переменной площадью поперечного сечения, не меняет свойства суперпозиции затухающих колебаний.

Как будет показано в гл. 7, величина отклика речевого тракта на возбуждение зависит от формы собственных функций $X_m(x)$, но в формантном синтезаторе эта информация отсутствует, в связи с чем возникает проблема стабилизации уровня низкочастотных компонент речевого сигнала для достижения независимости от расположения первых двух формант. Это вызывается тем, что фазы колебаний на четных и нечетных резонансах сдвинуты на 180°, т. е. отклики соседних формант имеют разные знаки.

Возможность ограничения полосы телефонного канала снизу частотой 300 Гц говорит о малом вкладе более низкочастотных компонент в разборчивость речи. Однако эти компоненты влияют на восприятие тембра, и ухо более чувствительно к неестественным вариациям их амплитуды, чем к аналогичным ошибкам в уровне формант. Для того чтобы избежать нежелательных изменений в низкочастотной области, в [118] предлагается менять амплитуду голосового возбуждения обратно пропорционально корню квадратному из частоты основного тона ($A_0\approx 1/\sqrt{F_0}$). Кроме того, полезно за голосовым источником поместить резонатор с частотой 250—300 Гц и шириной полосы около 150 Гц, так что он скорее представляет собой фильтр низкой частоты. Меняя амплитуду этого резонатора в соответствии с частотой первой форманты F_1, можно добиться стабилизации отклика системы на низких частотах. Отметим, что коэффициент ослабления передаточной функции речевого тракта равен единице, поскольку для всех артикуляционных состояний, кроме смычки, воздушный поток через голосовую щель равен сумме потоков через рот и нос. Изменения наклона спектральной огибающей, создаваемые дополнительным резонатором, компенсируются реальным интегратором с частотой среза 600—700 Гц, поставленным на выходе синтезатора после суммирования откликов резонаторов.

Дополнительный низкочастотный резонатор играет и самостоятельную роль при формировании назальных звуков, но этот вопрос мы обсудим несколько ниже.

Естественная зависимость отклика резонаторов на голосовое возбуждение достигается, если учесть элементарные свойства дифференциальных уравнений второго порядка, которыми описываются формантные резонаторы. Как было показано в гл. 2, для уравнения

$$y'' + 2g y' + \omega^2 y = x, \tag{6.2}$$

рекурсивная форма есть

$$y_i = b_i x_i + a_{1\,i} y_{i-1} + a_{2i} y_{i-2}, \tag{6.3}$$

где x_i — возбуждение в i-й момент времени, y_i — отклик резонатора, y_{i-1}, y_{i-2} — значения отклика в предыдущие моменты времени. Коэффициент b_i обратно пропорционален квадрату мгновенного значения частоты ω_i:

$$b_i = (1 - a_{1i} - a_{2i})/\omega_i^2.$$

Поэтому при прочих условиях огибающая спектра формантных максимумов имеет наклон -12 дБ/окт (каждое деление на ω соответствует наклону в -6 дБ/окт), что примерно соответствует наблюдающемуся наклону спектра для реальных речевых сигналов.

Несколько выше мы говорили о том, что разборчивость звуков обеспечивается изменением частот первых трех формант, а высшие форманты лишь улучшают натуральность и могут быть фиксированы. Действительно, в первом приближении можно принять, что частоты высших резонансов речевого тракта почти не зависят от его формы, а определяются лишь его длиной $l_{тр}$. Тогда для достаточно большой площади ротового отверстия и $i > 3$ формантные частоты можно рассчитать как

$$F_i \approx (2i - 1)\, c_0 / l_{тр},$$

и при $l_{тр} = 17{,}5$ см, $F_4 = 3500$ Гц, $F_5 = 4500$ Гц и т. д. Однако при изменении длины речевого тракта изменяются и частоты высших резонансов. В табл. 6.1 показаны расчетные частоты высших формант F_4, F_5 и F_6 для гласных звуков, которым свойственна разная длина тракта.

Таблица 6.1. Формантные частоты для гласных звуков

Звук	Длина тракта, см	F_4, Гц	F_5, Гц	F_6, Гц
Нейтраль	17,5	3500	4500	5500
У	20,7	2910	3630	4420
Э	19,8	2770	3570	4220
А	19,4	2710	3800	4730
О	19,4	2630	4030	4730
Ы	19,4	2780	3830	4590
И	18	3080	3790	5050

Из этой таблицы видно, что разность между значениями формантных частот гласных звуков может достигать величин 450 Гц для F_4, 460 Гц — для F_5 и 830 Гц — для F_6. Эти величины сопоставимы с шириной критической полосы, равной 500 Гц в диапазоне частот 3—5 кГц, так что разница в положении высших формант вполне может быть замечена системой слухового восприятия. Кроме того, соотношения между формантными частотами для некоторых звуков сильно отличаются от гармонических. Например, для /О/ $F_5-F_4=1400$ Гц, а $F_6-F_5=700$ Гц, тогда как для /И/ $F_5-F_4=710$ Гц, а $F_6-F_5=1200$ Гц. В то же время для /У, Э, А/ эти разности близки к 800 Гц. Как показывают последние работы по теории слухового анализатора, система восприятия очень чувствительна к отклонению компонент сигнала от гармонических отношений. Поэтому, по крайней мере для некоторых звуков, наряду с первыми тремя формантами необходимо управление и высшими формантами.

Информацию о частотах высших формант можно получить либо непосредственно измерениями на реальных звуках речи, либо расчетным путем. Следует, однако, обратить внимание на то, что при частотах выше 3 кГц при наличии достаточно больших площадей поперечного сечения в речевом тракте возникают радиальные колебания и расчет резонансов затрудняется. Так, в опытах с физической моделью речевого тракта в [118] были обнаружены дополнительные резонансы в области 3—5 кГц и сдвиг частот основных резонансов, если площадь модели где-либо превышала 6 см^2. Дальнейшее развитие теории акустических процессов речеобразования и улучшение методов анализа речевых сигналов дадут необходимые сведения о поведении высших резонансов, а пока можно пользоваться приближенными данными с дальнейшей коррекцией методом проб и ошибок при участии аудиторских испытаний.

В параллельной схеме, реализуемой аналоговыми резонаторами, возникает проблема согласования фаз колебаний разных формант, однако было найдено, что фазовые соотношения важны только при восприятии через высококачественный головной телефон, а при прослушивании в реверберирующей комнате через громкоговоритель несогласование фаз маскируется и не замечается слушателем.

§ 6.3. Акустические процессы в формантном синтезаторе

Основные нелинейные и параметрические явления в акустике речевого тракта обсуждались в Введении. В соответствии с перечнем этих явлений рассмотрим возможности их реализации в формантном синтезаторе.

Взаимодействие голосового источника с трактом. Имеется ряд явлений, связанных с разведением и сведением голосовых складок, которые влияют на акустические характеристики

речевого сигнала и которые нетрудно реализовать даже в формантном синтезаторе, конечно, при условии соответствующего изменения используемой модели голосового источника.

При артикуляции глухих фрикативных согласных, находящихся между гласными, голосовые складки еще продолжают колебания в течение 20—40 мс после образования щели в тракте и возникновении турбулентных шумов. На интервале перехода от фрикативного к гласному синхронно с движениями артикуляторных органов голосовые складки начинают сближаться. При этом обычно имеется такой интервал времени, когда турбулентный источник уже прекратил свою работу, а голосовые складки еще не начали колебаться, что проявляется в виде паузы длительностью в 10—30 мс между фрикативным согласным и последующим гласным. Длительность сегмента с фонацией в начале фрикативного и длительность паузы в его конце зависит от сдвига по времени и скорости движения голосовых складок и артикуляторных органов. В начале высказывания (после паузы) колебания голосовых складок не сразу выходят на стационарный режим, и спектр импульсов возбуждения расширяется постепенно. Поэтому начальные звонкие слабее последующих звуков и имеют меньше высших гармоник. Это явление отмечается для начальных /B, D, G, M, N, L/ в английском языке.

Конечно, все эти эффекты можно измерить на уровне акустических параметров, и записать в виде правил управления голосовым и турбулентным источниками возбуждения в явном виде. Однако гораздо проще и значительно более естественно задать лишь сдвиг по времени между командами на формантные переходы и командой на сведение-разведение голосовых складок. При этом голосовой источник должен обладать возможностью моделирования явлений, связанных с изменениями расстояния между голосовыми складками.

Аналогичные эффекты наблюдаются и при артикуляции взрывных согласных, причем в разных языках используются различные тактики управления. Например, в английском языке начало сведения голосовых складок задержано относительно раскрытия смычки, в результате чего между импульсным источником возбуждения и началом фонации проходит довольно значительное время, которое играет роль основного различительного признака между английскими звонкими и глухими взрывными. Величина этого времени в английском языке зависит и от контекста. Так, по [137] для /M, N, L, R/ это время увеличивается в группах согласных (кластерах) с последующим сонорантом и в начале слова и уменьшается в кластерах с предшествующим /S/, если последний сегмент предыдущего слова звонкий, а также в функциональных словах. В русском языке сдвиг фазы между артикуляцией и движениями голосовых

складок значительно меньше. Количественные характеристики этого явления обсуждались в § 4.1.

С разведением голосовых складок связано и прекращение возбуждения формант, начиная с высших, при довольно длительном сегменте звучания голосового источника в конце высказываний. Это вызывается тем, что при разведении голосовых складок их колебания приобретают все более синусоидальный характер, в результате чего спектр импульсов сужается и формантные колебания перестают возбуждаться. В этом случае, как и ранее, вместо явных правил управления формантными резонаторами для моделирования этого явления целесообразно использовать адекватный голосовой источник.

В теоретических основах формантного синтеза обычно принимается, что импеданс голосового источника значительно больше импеданса речевого тракта, и никакого взаимодействия между ними, кроме возбуждения акустических колебаний, не имеется. Однако в действительности импеданс голосового источника не столь велик, особенно на низких частотах. В результате этого при открытой голосовой щели происходит изменение формантных частот и затуханий и появляются дополнительные резонансы. В [59] было показано, что знак и величина девиации формантных частот зависят от соотношения импедансов подсвязочной области, голосовой щели и речевого тракта. Поскольку в формантном ситезаторе импедансы тракта и подсвязочной области не рассчитываются, то приближенно можно принять, что дополнительная ширина k-й форманты σ_k и приращение k-й формантной частоты δF_k описываются как

$$\delta_k = \frac{a_0}{a_1 + F_k} S_г,$$

$$\delta F_k = \frac{b_0}{b_1 + F_k} S_г,$$

где $S_г$ — площадь голосовой щели, F_k — частота форманты. На интервале сомкнутых голосовых складок $S_г = 0$ и, соответственно, $\delta_k = 0$ и $\delta F_k = 0$, т. е. параметры формант определяются только формой речевого тракта, а в формантном синтезаторе — медленно меняющимися командами, соответствующими синтезируемому звуку. При открытой голосовой щели $S_г > 0$ и девиация параметров возрастает по мере снижения формантной частоты. Наибольшая девиация наблюдается на первой форманте. Величины коэффициентов a_0, a_1, b_0, b_1, а также знак b_0 определяют тембральные характеристики синтетического голоса.

При $S_г = 0{,}2$ см^2 величина δF_k составляет около 10—20% от F_k на частоте 500 Гц, т. е. около 50—100 Гц. На этой же частоте $\delta_k \approx 100$ Гц.

Для моделирования эффекта вариации формантных параметров синхронно с колебаниями голосовых складок голосовой

источник, помимо собственного импульса возбуждения, должен выдавать и значение площади голосовой щели $S_r(t)$. При фиксированных коэффициентах a_0, a_1, b_0 и b_1 девиации δ_k и δF_k будут изменяться в соответствии с изменениями формантных частот. Но все же полной зависимости от формы речевого тракта, в которой меняются сами эти коэффициенты, при таком подходе достичь не удается. Кроме того, необходимо учитывать задержку во времени влияния изменения площади голосовой щели на изменение формантных параметров. В [59] было показано, что для тракта длиной в 17,5 см никакие изменения площади голосовой щели не могут сказаться на акустических характеристиках речевого тракта раньше, чем через 1 мс. Это означает, что в формантном синтезаторе нужно экспериментально подбирать величину задержки девиации формантных параметров относительно колебаний голосовых складок для достижения наибольшего перцептивного эффекта.

Поскольку δ_k и δF_k меняются примерно пропорционально S_r, т. е. весьма быстро, то необходимо предусмотреть возможность такого же быстрого изменения параметров g и ω в (6.2) для формантного синтезатора. Это требует пересмотра тактики управления параметрами формант в синтезаторе. Так, в [135] принимается, что значения формантных частот должны обновляться через 5 мс. При этом некоторое сглаживание параметров достигается вследствие естественных переходных процессов в уравнении (6.2). Наиболее экономная тактика управления девиацией параметров состоит в скачкообразном изменении δ_k и δF_k и сохранении их постоянными в течение всего периода открытой голосовой щели, и установлении $\delta_k = 0$, $\delta F_k = 0$ при закрытой голосовой щели. Степень точности в отслеживании δ_k и δF_k, однако, нуждается в дополнительных исследованиях с оценкой перцептивного эффекта различных способов.

Среди эффектов влияния речевого тракта на голосовой источник в формантном синтезаторе можно реализовать падение частоты основного тона во время звонкой смычки и увеличение периода между первыми импульсами голосового источника вслед за взрывом смычки, особенно глухой.

Влияние подсвязочной области. Рассмотренные выше девиации частот и затуханий резонансных колебаний синхронно с колебаниями голосовых складок являются следствием влияния подсвязочной области — трахеи, бронхов и легких, на акустические характеристики речевого тракта при раскрытой голосовой щели. Это влияние оказывается наибольшим при артикуляции глухих согласных (взрывных и фрикативных), поскольку в это время голосовые складки разведены достаточно широко, и взаимодействие подсвязочной области и речевого тракта максимально. Подсвязочная область и речевой тракт образуют единую акустическую систему, длина которой почти вдвое

больше длины речевого тракта от голосовой щели до губ, и вследствие этого число резонансов в полосе до 5 кГц также почти вдвое больше. Один из этих резонансов на частоте около 1000 Гц отчетливо проявляется в спектре взрыва глухих согласных.

Дополнительные резонансы появляются и при достаточно большой амплитуде колебаний голосовых складок во время фонации. Число таких дополнительных формант, скорее всего, может быть ограничено двумя — одна около 500 Гц, а другая около 1000 Гц, как это следует из моделирования эффекта на однородной акустической трубе. Поэтому усложнение формантного синтезатора по числу резонаторов не слишком велико. Проблема заключается в другом — как определить частоты этих дополнительных формант для реальных форм речевого тракта. Для сегментов взрыва глухих согласных и фрикативных эти частоты могут быть определены путем анализа реальных речевых сигналов. Измерение параметров дополнительных формант на интервале открытой голосовой щели во время колебаний голосовых складок представляется довольно трудной задачей, и здесь должна помочь хотя бы упрощенная, но адекватная модель явления.

Стробоскопические измерения показывают, что примерно у 80% женщин и 20% мужчин задняя часть голосовой щели остается открытой даже на интервале схлопывания голосовых складок, и, таким образом, связь между ротовой полостью и подсвязочной областью не прерывается. Д. Клатт считает, что моделирование дополнительных формант существенно для создания женского голоса, причем нужно использовать три дополнительные форманты на частотах 600 Гц, 1400 Гц и 2200 Гц. Для мужчин частоты этих формант должны быть несколько ниже.

Низкочастотный радиальный резонанс. Податливость стенок речевого тракта создает низкочастотные радиальные колебания, которые повышают значения формантных частот и препятствуют уменьшению до нуля частоты первой форманты при стремлении минимальной площади поперечного сечения тракта к нулю. На интервале звонкой смычки частота первой форманты равна частоте низкочастотного радиального резонанса $F_{\text{рад}}$. Величина $F_{\text{рад}}$ зависит от объема речевого тракта, заключенного между голосовой щелью и смычкой, а также от степени податливости стенок тракта (см. также § 1.3). Для одного и того же диктора $F_{\text{рад}}$ может меняться до 50%. В формантном синтезаторе $F_{\text{рад}}$ может быть фиксированной, например, на частоте 200 Гц, либо изменяться согласно таблице, рассчитанной для разных мест артикуляции согласного на фоне различных гласных звуков.

Назализация. Опускание небной занавески создает сток акустической энергии через носовую полость, что обычно трактуется как появление нулей в передаточной функции

речевого тракта. В формантном синтезаторе *DECTalk* назализация моделируется парой полюс — нуль, причем полюс фиксирован на частоте 270 Гц, а частота нуля равна 270 Гц для неназализованных звуков (полностью компенсируя полюс), и увеличивается таким образом, чтобы попасть в середину интервала между 270 Гц и значением частоты первой форманты гласного звука, увеличенной на 100 ГЦ на назализованном сегменте. Для назальных /М, Н/ частота нуля равна 450 Гц.

Антиформантный резонатор, реализующий нуль, конструируется в виде рекурсивной схемы второго порядка

$$y_i = b'_i x_i + a'_{1i} y_{i-1} + a'_{2i} y_{i-2},$$

где коэффициенты b'_i, a'_{1i} и a'_{2i} связаны с коэффициентами b_i, a_{1i} и a_{2i} рекурсивной схемы для дифференциального уравнения второго порядка (см. гл. 2) как

$$b'_i = 1/b_i, \quad a'_{1i} = -a_{1i}/b_i, \quad a'_{2i} = -a_{2i} b_i.$$

Эффект назализации наиболее заметно сказывается в низкочастотной области, и использование пары полюс — нуль действительно позволяет его частично имитировать. Однако назализация имеет и другие проявления. На рис. 6.2 показаны расчетные треки формант для слога/АМА/ по [59]. Как видно, если на сегменте назализованной первой гласной /А/ появление дополнительного резонанса между первой и второй формантами еще можно трактовать как расщепление нулем исходного

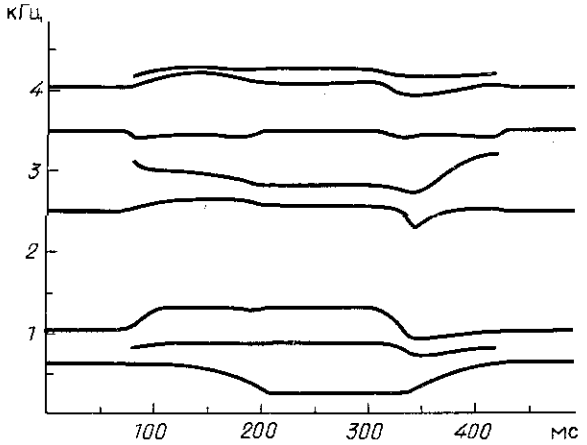


Рис. 6.2. Изменение резонансных частот речевого тракта в синтезированном слоге АМА

первого резонанса, то на сегменте губной смычки частоты первого резонанса и дополнительного резонанса разошлись слишком далеко. Поэтому более естественным было бы не оперировать понятием нулей, а просто вводить дополнительные

резонансы. Это соображение подкрепляется и поведением частоты F'_2 второго дополнительного резонанса, расположенного между третьей и четвертой формантами — перед раскрытием губной смычки F'_2 близка к F_3, а по завершении переходного процесса движений губ и небной занавески F'_2 приближается к F_4.

Другой существенный аспект назализации состоит в значительном расширении полос формант вследствие увеличения потерь на колебания стенок носовой полости, вязкого трения и излучения через ноздри. Полосы формант расширяются при этом примерно в два раза.

Хотя существующие формантные синтезаторы моделируют звук /Л/ лишь движениями формант, известно, что при артикуляции /Л/ временно возникает разветвление речевого тракта в виде каналов по бокам языка. Это сопровождается, в принципе, теми же явлениями, что и назализация, так что при синтезе /Л/ может оказаться целесообразным использование дополнительных формант так же, как и при синтезе /М, Н/.

Фрикативные звуки. Сужение в речевом тракте до площади 0,2 –0,4 см^2 при определенной скорости воздушного потока вызывает его завихрения в расширяющейся части тракта и появление турбулентного источника возбуждения. Характеристики этого источника анализировались в гл. 5. Спектр сигнала возбуждения от турбулентного источника имеет один или несколько максимумов с довольно пологими склонами. В зависимости от расположения источника возбуждения в речевом тракте, он возбуждает только те резонансы, нули собственных функций которых находятся далеко от источника. Если же нули какой-либо собственной функции попадают в окрестность источника, то и в спектре акустического сигнала на выходе речевого тракта появляются нули на тех же частотах. В результате спектр некоторых фрикативных имеет полосовой характер. Если источник возбуждения находится на губах, то нулей нет, и спектр звуков /В, Ф/ почти равномерен.

Моделирование фрикативных звуков в формантном синтезаторе осуществляется просто управлением амплитуд формант (для параллельной схемы), что позволяет вводить и нули. Синтез губных фрикативных обеспечивается непосредственной передачей сигнала шумового источника на выход, минуя набор резонаторов, поскольку здесь нет необходимости в формировании нулей.

Фрикативный /С/ часто имеет спектральный максимум на частоте, лежащей выше 5 кГц. Во избежание неприятных переходных процессов при быстром изменении частоты высшей форманты, для /С/ специально вводится дополнительный резонатор, частота которого выбирается в соответствии с частотой дискретизации. При частоте дискретизации $F_{отс} = 10$ кГц, соответствующей полосе сигнала в 5 кГц, частота дополнительного резонанса для /С/ в [135] устанавливается равной 4,9 кГц.

Частоты формант при генерации фрикативных звуков сохраняют непрерывность движения с учетом коартикуляции. Иногда, особенно на звонких фрикативных, формантные частоты видны совершенно отчетливо на спектрограммах. Вследствие дополнительных потерь на завихрения воздуха ширина формант фрикативных звуков увеличивается.

Если фрикативный — звонкий, т. е. действуют сразу два *источника возбуждения* — голосовой и турбулентный, то пульсирующая скорость воздушного потока модулирует шум. При этом необходимо различать сигнал голосового возбуждения, который пропорционален первой производной по времени от объемной скорости потока в голосовой щели, и просто скорость (не объемную) потока. Возможно, несоблюдение этих соотношений привело к заключению о перцептивной эквивалентности суммирования сигналов голосового и шумового возбуждения и модуляции шума сигналами голосового источника в [118].

Излучение речевого сигнала. Характеристики излучения речевого сигнала, например, из ротового отверстия, получаются *путем аппроксимации импеданса излучения поршня малого* диаметра, помещенного в бесконечный экран. В [37] получены следующие приближенные выражения для импеданса излучения

$$Z_l = \frac{1}{2}\left(\frac{\omega r_l}{c_0}\right)^2 + j\frac{8\omega r_l}{3\pi c_0}, \tag{6.4}$$

где ω — круговая частота, c_0 — скорость звука, r_l — радиус поршня (в нашем случае — эквивалентный радиус излучающего ротового отверстия). Поскольку умножение на $j\omega$ соответствует однократному дифференцированию во временной области, то в формантных синтезаторах для моделирования эффекта излучения обычно берут первую производную по времени от сигнала на выходе сумматора формантных колебаний.

Необходимо, однако, обратить внимание на переменный параметр $r_l = \sqrt{S_l/\pi}$, где S_l — площадь излучающего отверстия. Площадь в процессе артикуляции меняется от довольно большой величины для открытых гласных до нуля для взрывных согласных. При малых площадях излучающего отверстия сигнал практически не дифференцируется, так что при движении губ к смычке и от нее наклон спектра излучаемого сигнала меняется на 6 дБ/окт и этим обстоятельством нельзя пренебрегать.

Игнорируя пока зависимость потерь излучения от частоты, запишем соответствующую (6.4) временную форму:

$$y = Rx + \frac{8r_l(t)}{2\pi c_0}x',$$

где x — входной сигнал, y — сигнал, излучающийся в пространство. Отсюда видно, что это уравнение с переменным значением

«постоянной» времени T_l:

$$\frac{y}{R} = x + T_l x',$$

где

$$T_l(t) = \frac{8r_l(t)}{3\pi R c_0}.$$

Более подробный анализ характеристик излучения с учетом экранирующего влияния головы проводится в гл. 7. Поскольку в формантном синтезаторе отсутствует информация о площади излучающего отверстия, то необходима спецификация постоянной времени T_l для всех звуков. Нужно также позаботиться о выборе правильной зависимости T_l от времени при переходе от одного целевого значения к другому.

Излучение через стенки речевого тракта на частоте радиального резонанса во время звонкой смычки моделируется установлением нижнего предела в 150—300 Гц для значений частоты первой форманты. Следует подчеркнуть, что звонкая смычка не должна имитироваться простым излучением сигнала голосового возбуждения — его сначала следует пропустить через резонатор с частотой, равной частоте радиального резонанса.

До сих пор не исследовались эффекты, возникающие при одновременном излучении через рот и нос, как это бывает у назализованных гласных. Очевидно, что в зависимости от акустических характеристик носовой и ротовой полостей, степени их связи и расстояния между ноздрями и губами, должны возникать разнообразные интерференционные явления.

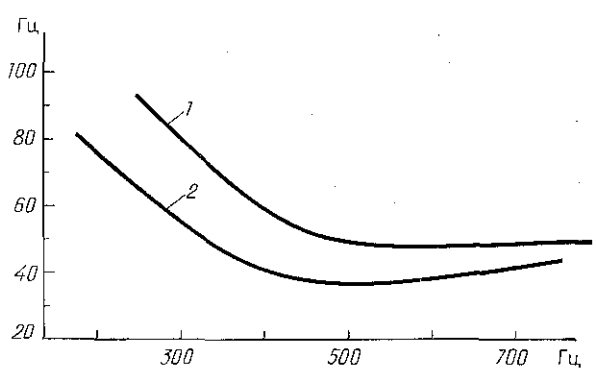


Рис. 6.3. Ширина полосы формант в низкочастотной области: *1* — женский голос, *2* — мужской голос (по [137])

Их роль в восприятии назализованных сегментов гласных заслуживает изучения.

Ширина полосы формант. Ширина полосы формант, хотя и слабо (в пределах одного и того же голоса), но влияет

на восприятие синтетических гласных [137]. Более существенным оказывается влияние на тембр, в особенности, на противопоставление мужского и женского голосов (см. рис. 6.3). Наконец, изменения ширины полосы формант в процессе синтеза меняет и амплитуду отклика, что особенно заметно на низких и высоких частотах.

На низких частотах ширина полосы форманты определяется, в основном, потерями на колебания стенок. Так, для однородной трубы с периметром поперечного сечения L ширина полосы есть

$$\Delta F_{\text{ст}} = \frac{R_{\text{ст}} c_0}{\pi L (R_{\text{ст}}^2 + \omega^2 M_{\text{ст}}^2)},$$

где $R_{\text{ст}} \approx 10$ г/см^{4}с, $M_{\text{ст}} \approx 0{,}02$ г/см^4. Начиная с частот $\omega = R_{\text{стр}}/M_{\text{ст}}$, т. е. с $f \approx 80$ Гц, потери на колебания стенок быстро падают.

На средних частотах основной вклад в ширину полосы формант вносят потери на вязкое трение и теплопроводность:

$$\Delta F_{\text{вт}} = \frac{k_{\text{ф}} c_0}{\pi} \sqrt{\frac{2\pi\omega\rho_0\mu}{S}},$$

где $k_{\text{ф}}$ — коэффициент формы ($k_{\text{ф}} = 1$ для кругового сечения), ρ_0 — плотность воздуха, μ — коэффициент вязкого трения, c_0 — скорость звука, S — площадь поперечного сечения трубы.

На высоких частотах ширина полосы формант определяется потерями на излучение и быстро растет с увеличением частоты:

$$\Delta F_{\text{изл}} = \frac{3\omega^2 r_l}{32 c_0}.$$

Потери на излучение зависят и от площади излучающего отверстия, причем разница на высоких частотах может быть весьма заметной (см. рис. 6.4).

Таким образом, суммарная ширина полосы формант растет обратно пропорционально квадрату частоты на низких частотах и прямо пропорционально квадрату частоты — на высоких частотах:

$$\Delta F = \frac{\alpha_1}{1 + \alpha_2 \omega^2} + \alpha_3 \sqrt{\omega} + \alpha_4 \omega^2.$$

Эта формула может служить для вычисления ширины полосы формант, если коэффициенты α_1, α_2, α_3 и α_4 подобрать в соответствии со свойствами звуков речи и характеристиками дикторов. Зависимость ширины полосы формант от их частоты для гласных звуков русской речи показана на рис. 6.5.

Так же, как и в случае изменения постоянной времени дифференцирования в характеристике излучения, при изменении площади излучающего отверстия потери на высоких частотах меняются в несколько раз. При этом наиболее существенным

может оказаться не разница в ширине полосы форманты, а скорость ее изменения. В области губной смычки эта скорость довольно велика, в результате чего меняются свойства резонансных колебаний — при совместном влиянии быстрого

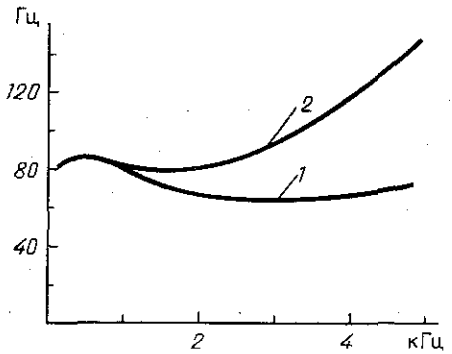


Рис. 6.4. Ширина полосы резонансов в однородной трубе: *1* — без излучения, *2* — с излучением

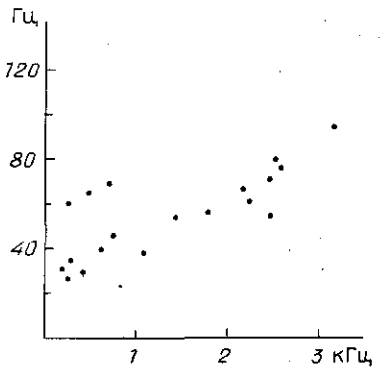


Рис. 6.5. Ширина полосы формант для гласных звуков русской речи

изменения частоты и ширины полосы резонанса его частотная характеристика сильно деформируется. Это означает, что во избежание ошибки вычисления формантных колебаний с помощью рекурсивной схемы (6.3), на сегментах с быстрым изменением g и ω необходимо существенно уменьшить интервал обновления этих параметров.

§ 6.4. Управление формантным синтезатором

В результате экспериментов по восприятию речи, синтезированной с помощью рисованных спектрограмм, в Хаскинских лабораториях пришли к понятию локуса [91]. Локус — это та частота форманты, к которой асимптотически стремится тот или иной резонанс при уменьшении площади речевого тракта в каком-либо месте до нуля (рис. 6.6). Первый импульс голосового источника после раскрытия смычки возбуждает акустические колебания на той частоте, которая соответствует определенному моменту переходного процесса форманты. Слоги «согласный — гласный», синтезированные на основе теории локусов, оказались вполне разборчивыми, тогда как фиксация значений формантных частот в начале переходных процессов приводила к изменению восприятия места артикуляции согласного в зависимости от формантных частот гласного (например, по мере повышения частоты второй форманты слышались сначала /Д/, затем /Г/, снова /Д/ и, наконец, /Б/). Теория локусов определила направление большинства работ как в области синтеза, так и в области распознавания речи.

И если распознавание речи по формантным переходам оказалось не очень надежным, то в формантном синтезе были достигнуты впечатляющие успехи.

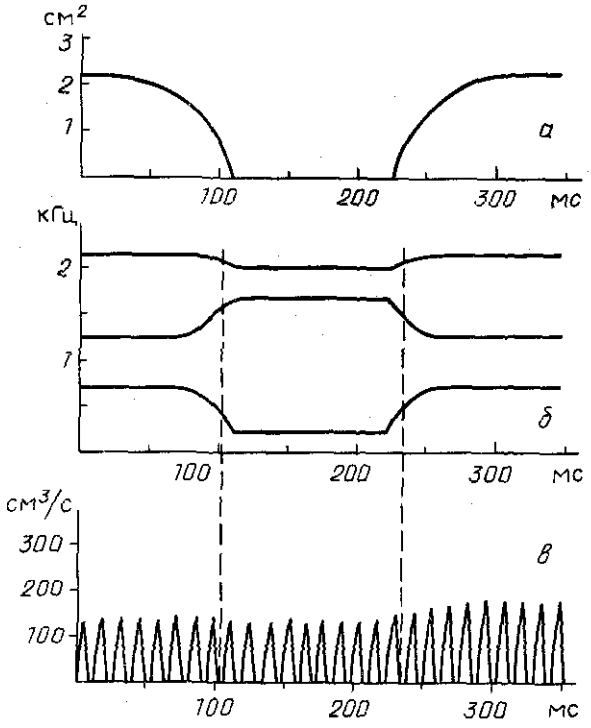


Рис. 6.6. Иллюстрация к теории локусов: *а* — площадь сужения в речевом тракте, *б* — резонансные частоты тракта, *в* — импульсы объемной скорости голосового источника. Вертикальные штриховые линии соответствуют моментам времени возбуждения акустических колебаний в начале и конце смычки

В основе теории локусов находится предположение, что какова бы ни была форма речевого тракта, его резонансы определяются местом наибольшего сужения при достаточно малой площади этого сужения. Анализ артикуляционно-акустических свойств речевого тракта немедленно показывает, что в общем случае это предположение не справедливо. Еще Рэлеем было показано, что повышение или понижение частоты того или иного резонанса акустической системы зависит от того, попадает ли место сужения в трубе на узел или пучность соответствующей собственной функции (подробнее об этом свойстве см. гл. 7). Поскольку распределение узлов и пучность в трубе зависят от ее формы, то это означает отсутствие фиксированных значений, к которым

асимптотически стремились бы частоты резонансов при смыкании стенок в каком-либо месте трубы. Следовательно, *работоспособность теории локусов зависит от свойств гласных.*

Действительно, линейная интерполяционная формула для частоты $F_{2н}$ начальной точки формантного перехода

$$F_{2н} = F_{2лок} + k(F_{2гл} - F_{2лок}) \qquad (6.5)$$

при любом фиксированном k не в состоянии дать правильные оценки. Здесь $F_{2лок}$ — значение локуса второй форманты, $F_{2гл}$ — *стационарное значение второй форманты гласного* звука, следующего за согласным. Исследуя распределение на плоскости ($F_{iгл}$, $F_{iн}$), Д. Клатт установил, что в английском языке имеется четкое разделение на два кластера — передние гласные и округленные гласные [137]. В каждом кластере начальное значение формантного перехода рассчитывается достаточно хорошо с помощью (6.5), но коэффициент k для второй *и третьей формант сильно отличаются, а для первой форманты* можно использовать один и тот же коэффициент, независимо от последующего гласного. Конечно, для согласных с разным местом артикуляции набор коэффициентов каждый раз другой. Эти правила обеспечивают 95% разборчивости слогов «согласный — гласный».

В синтезаторе русской речи, описанном в [29], не делается *различие типов гласных, но зато согласные делятся на твердые* губные /М, Б, В, П, Ф/, зубные /Н, З, Д, Т, С/, альвеолярные /Ш, Ж, Р/ и небные /К, Г, Х/, и для каждого i-го типа используется свой коэффициент α_i:

$$F_{iн} = \alpha_i F_{2гл} + p_i,$$

тогда как для всех мягких согласных $\alpha_i = 0$, и группы *губных /М, В, П, Б, Ф/, зубных и альвеолярных* /Н, З, Д, Т, С, Ш, Р, Ч/ и небных /Й, К, Г, Х/ различаются лишь величинами p_i.

Созданные на основе теории локусов формантные синтезаторы обладают довольно высокой разборчивостью, но в присутствии шумов понимание синтетической речи затрудняется в большей степени, чем естественной речи. По всем *основным показателям качество речи формантных синтезаторов* значительно уступает качеству естественной речи. Очевидно, что направление переходов формантных частот не является единственным признаком места артикуляции согласных. Более того, анализ формантных переходов для множества дикторов показывает, что направление движения частот формант для одного и того же согласного у разных дикторов может *отличаться настолько, что спектрограммы для разных согласных* становятся более похожими, чем спектрограммы для одного и того же согласного. В то же время, правила для управления формантным синтезатором обычно создаются на основе анализа поведения формант лишь для одного диктора

(чаще всего — автора разработки). При этом индивидуальные особенности произношения неизбежно абсолютизируются и качество синтеза в известной мере зависит от дикции человека, чьи спектрограммы положены в основу разработки правил для формантного синтезатора.

На рис. 6.7 показаны спектрограммы, сделанные на 48-канальном полосном спектрографе для слогов АБА, АДА,

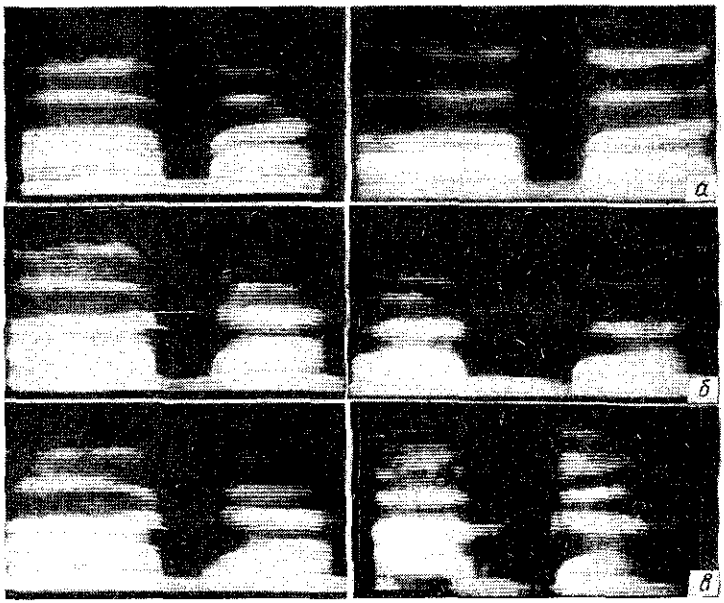


Рис. 6.7. Спектрограммы слогов: *а* — АБА, *б* — АДА, *в* — АГА

АГА, произнесенных разными дикторами. Из массива дикторов, более чем в 30 человек, отобраны реализации, демонстрирующие разнообразие формантных переходов. В левом столбце показаны спектрограммы, где переходы второй форманты соответствуют общепринятым представлениям: при подходе к смычке частота второй форманты для /Б/ падает, для /Д/ — возрастает, для /Г/ — возрастает в еще большей степени. После раскрытия смычки форманты движутся как бы с зеркальным отображением относительно середины смычки. В правом столбце помещены спектрограммы тех же слогов, произнесенных другими дикторами, но на них формантные переходы почти не отличаются по направлению, и выглядят, как реализации слога АБА.

При попытке распознавания спектрограмм слогов, включающих также согласные /М, Н/, обнаружилось, что направления формантных переходов служат далеко не самым надежным признаком места артикуляции. Наиболее информативными

признаками оказались соотношения средней энергии на участках стационарного состояния гласных и переходов первой и второй формант, а также на интервале смычки. Этот результат можно интерпретировать как следствие различия в скорости изменения минимальной площади поперечного сечения речевого тракта в месте артикуляции согласных. Действительно, скорость движения артикуляторов у заднеязычных, переднеязычных и губных согласных различна. Измерения длительности переходных процессов раскрытия для одного диктора дают следующие величины: для губных согласных — 100—120 мс, для переднеязычных — 80—120 мс, для заднеязычных — 80—180 мс.

Скорость движения артикуляторных органов различна в разных направлениях. Например, скорость опускания нижней челюсти может быть в 1,5 раза больше скорости ее подъема [59]. Следовательно, скорость переходных процессов при подходе к смычке должна быть меньше скорости при раскрытии смычки, особенно, для переднеязычных и губных, на артикуляцию которых движения нижней челюсти влияют в наибольшей степени. Изменения энергии в области той или иной форманты более чувствительны к минимальной площади в месте артикуляции, чем изменения формантных частот, если эта площадь достаточно большая. Как было показано в [59], скорость изменения частоты первой форманты значительно меньше скорости изменения минимальной площади поперечного сечения S_{min} при $S_{min} > 1$ см^2, тогда как при малых площадях ($S_{min} < 1$ см^2) скорость изменения частоты первой форманты возрастает на порядок. В результате этого вблизи смычки частота первой форманты изменяется настолько быстро, что на периоде основного тона — от одного импульса источника голосового возбуждения до другого может создаться впечатление скачкообразного движения частоты.

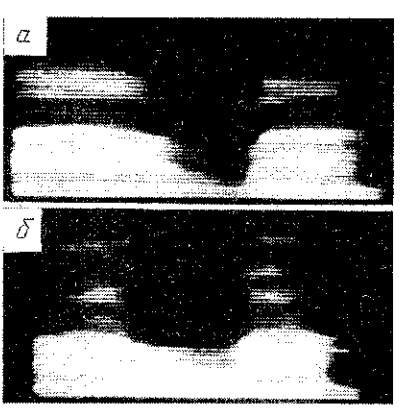


Рис. 6.8. Спектрограммы: *а* — слог АВА, *б* — слог АЛА

Несимметрия скорости переходных процессов в окрестности смычки в некоторых случаях служит отличительным признаком особого, специально формируемого способа артикуляции. К числу таких звуков относятся /В/ и /Л/ (см. рис. 6.8).

Со скачками движения первой форманты может быть связано и явление запирания, описанное в [59]. Это явление состоит в прекращении распространения акустических колеба-

ний вследствие податливости стенок, которые на низких частотах начинают колебаться в противофазе с акустической волной и тем самым препятствуют ее прохождению через сужение. Наиболее вероятно возникновение запирания на губных и заднеязычных согласных, так как сближающиеся участки тракта — губы, язык и мягкое небо — имеют импеданс инерционного типа, что служит необходимым условием возникновения запирания. Переднеязычная смычка образуется в районе твердого неба, имеющего импеданс упругого типа. Поэтому запирание волн в этом месте менее вероятно.

Артикуляция глухих согласных отличается от артикуляции звонких разведением голосовых складок в начале смычки и сведением — в конце. Как уже обсуждалось ранее, в зависимости от сдвига фаз в началах движения артикуляторных органов к размыканию и движения голосовых складок к смыканию, проходит различное время между взрывом смычки и началом фонации, т. е. голосового возбуждения. За это время формантные частоты уходят далеко от начального положения, и зачастую первый импульс голосового источника возбуждает колебания уже в конце формантных переходов, так что в направлении перехода остается мало информации о месте артикуляции согласного. Основными признаками места артикуляции служат характеристики отклика на импульсное возбуждение в момент взрыва и следующего за ним шумового сегмента. В то же время сохраняется вся информация о месте артикуляции в переходе от гласного к согласному.

В момент взрыва площадь сечения тракта в месте артикуляции еще достаточно мала, поэтому акустические характеристики лучше соответствуют месту артикуляции. При этом акустические колебания возбуждаются на частотах, соответствующих размерам и форме областей речевого тракта, находящихся между смычкой и губами. Сопоставление спектра взрыва, измеренного на интервале 10—20 мс со спектром звука после первого импульса голосового возбуждения дает возможность определить направление формантных переходов и, по мнению авторов [79], дает сведения об относительно инвариантных свойствах признака места артикуляции. Опыты по удалению взрывов, выполненные в [32], также свидетельствуют в пользу их большей информативности по сравнению с формантными переходами на участке «согласный — гласный».

Таким образом, формантные переходы несут лишь часть информации о месте артикуляции согласного, и в синтезе речи должны использоваться также и признаки, связанные со скоростью изменения энергии в различных формантных областях, и признаки взрыва.

Коартикуляция. Одна из центральных проблем в построении правил управления формантным синтезатором — это коартикуляция, т. е. совместная артикуляция звуков речи. Коартикуляция имеет две стороны. Во-первых, это неизбежная

конкуренция за каналы моторного управления, когда соседние гласный и согласный звуки в слитном потоке речи образуются путем деформации поверхности языка и изменения его положения в ротовой полости. В силу непрерывности механических движений, начальные условия и целевое положение артикуляторных органов деформируют траектории их движений, и в течение некоторого времени характеристики речевого сигнала определяются и гласным и согласным одновременно. Во-вторых, коартикуляция появляется как следствие действия критерия оптимальности в системе управления, заставляющего выполнять подготовку к формированию следующих звуков еще при реализации текущего звука. Известны примеры прогнозирующей коартикуляции, когда огубление в слове «потому» наблюдалось уже на первом гласном, т. е. за четыре звука до /У/, признак огубления для которого является фонеморазличительным [49].

Будучи неотъемлемым свойством механики артикуляции и системы управления, коартикуляция включена в модуляционно-кодовую структуру речи таким образом, что аудитор ожидает появления коартикуляционных эффектов в определенных контекстах [156]. Следовательно, коартикуляция должна быть включена в правила управления формантным синтезатором.

Первые эксперименты по определению свойств коартикуляции на акустическом уровне с помощью спектрограмм были выполнены в [179] и до сих пор остаются образцом. В этих экспериментах обнаружены коартикуляционные эффекты как типа последействия, так и типа предсказания. Иными словами, в слогах $\Gamma_1 С \Gamma_2$ для одного и того же согласного и разных гласных направление формантных переходов от согласного к последующему гласному Γ_2 может зависеть от гласного Γ_1, и, наоборот, переходы от предыдущего гласного Γ_1 к согласному зависят от последнего гласного Γ_2 (см. пример для согласного /Г/ в окружении разных гласных на рис. 6.9). Для русского языка некоторые коартикуляционные эффекты можно найти на спектрограммах, помещенных в [16].

Трудности в изучении коартикуляции заключаются не только в необходимости анализа огромного числа звукосочетаний (включая и стечения согласных), но и в зависимости этих эффектов от разговорного стиля, темпа артикуляции и индивидуальных тактик управления артикуляцией у разных дикторов. Как можно видеть из рис. 6.7, ударная или безударная позиция гласного проявляется, наряду с прочим, и в изменении стационарного положения формант. Например, для безударного /А/ в конце слогов ГСГ частота первой форманты понижается, а второй — повышается, т. е. качество гласного /А/ приближается к качеству нейтрального гласного с эквидистантно расположенными формантами. Для конечного звука в слоге нейтрализация гласного сама по себе является

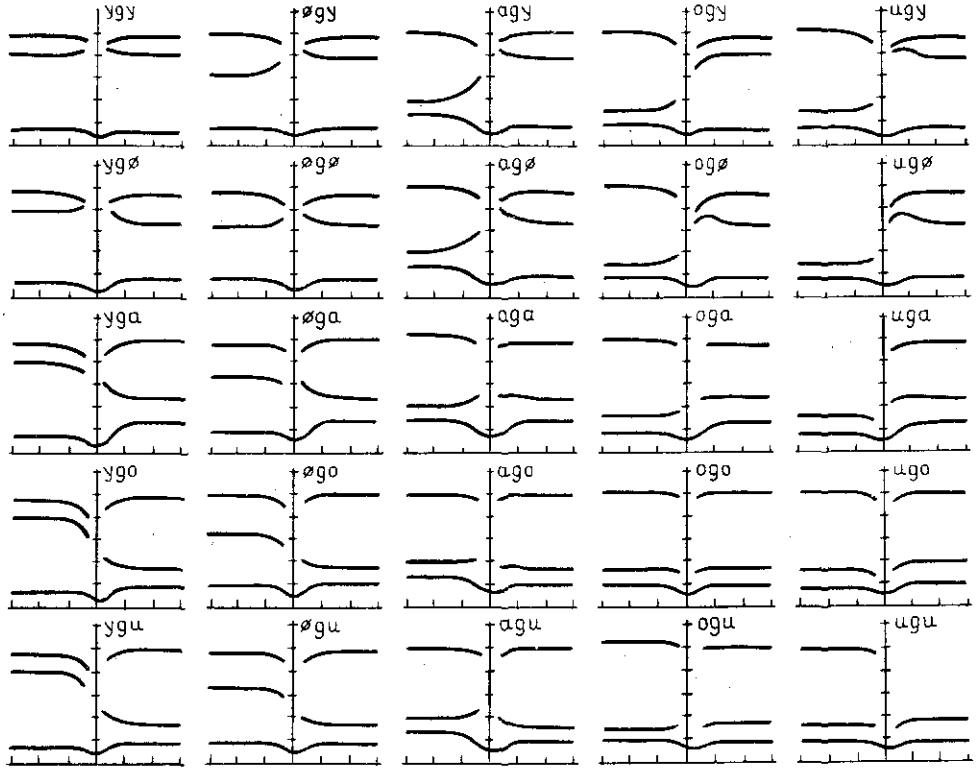


Рис. 6.9. Коартикуляция в слогах шведского языка (по [170])

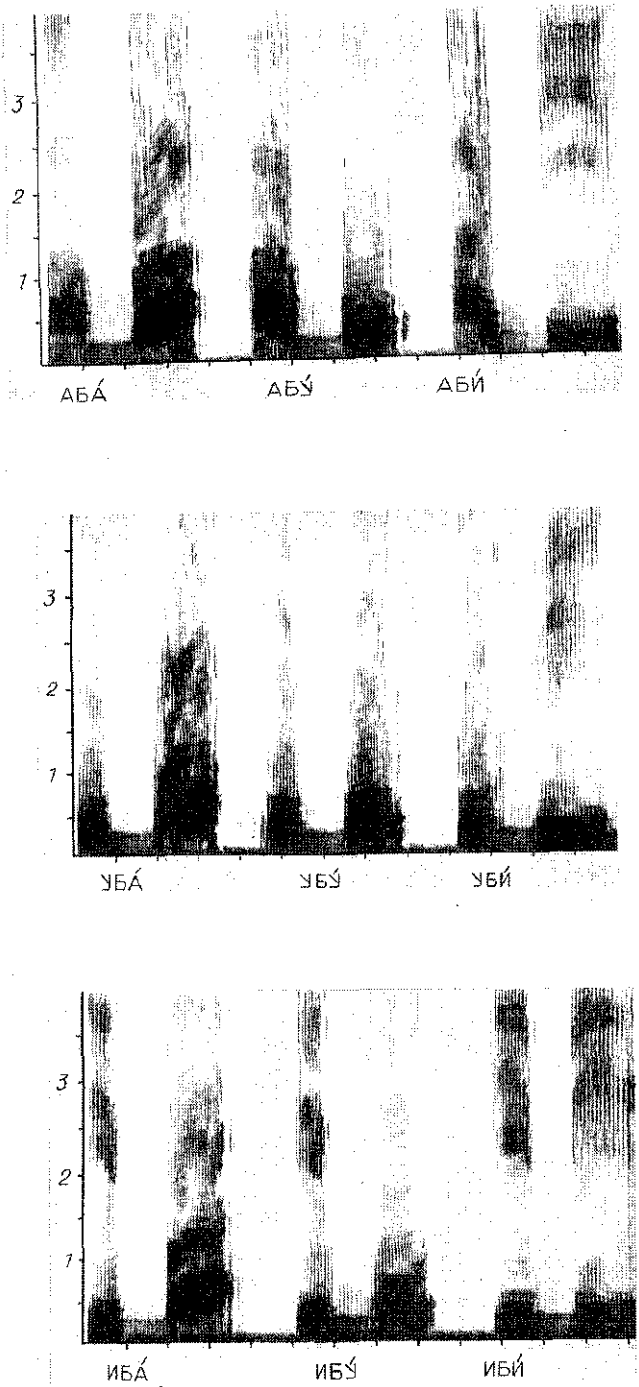


Рис. 6.10. Спектрограммы слогов русского языка

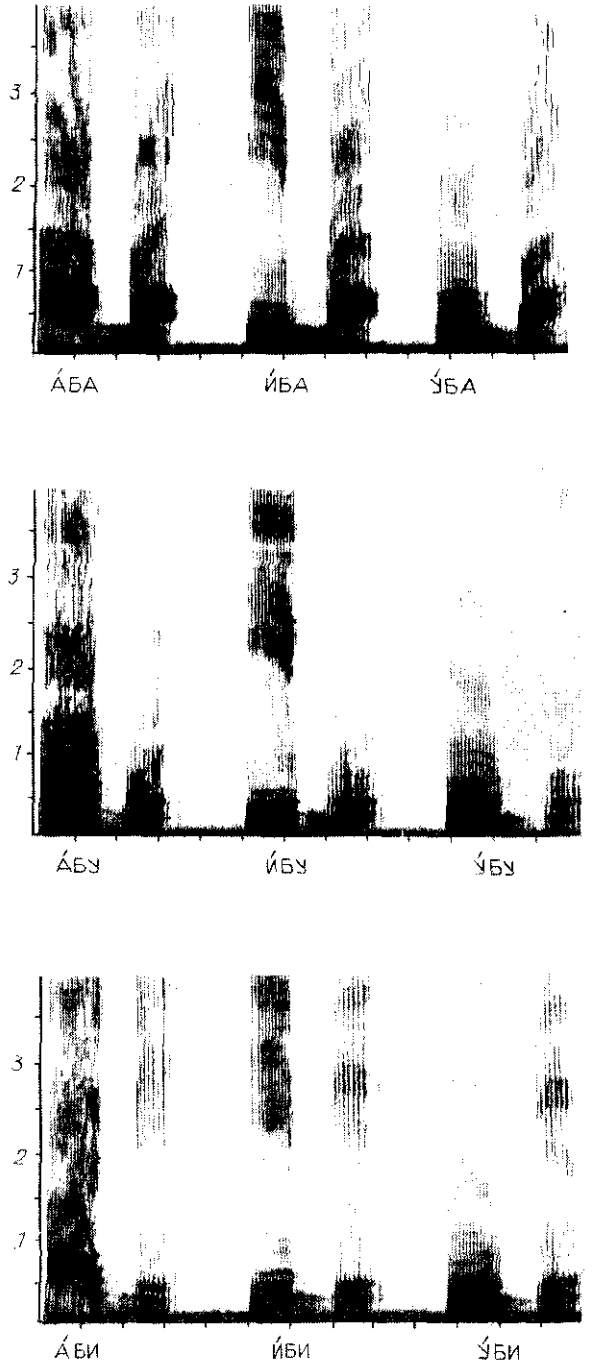
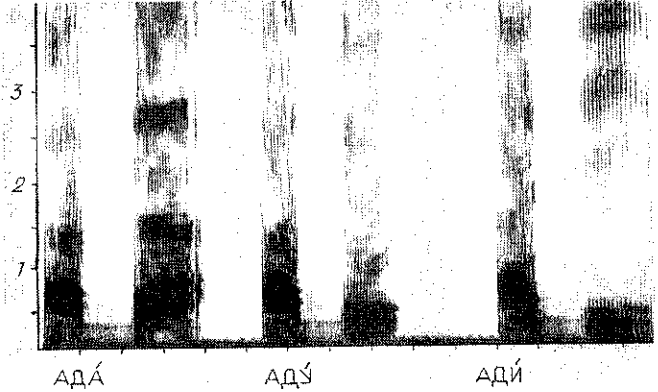
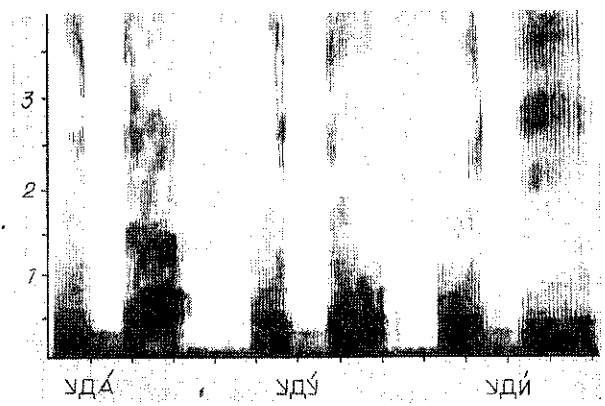


Рис. 6.10 (продолжение)





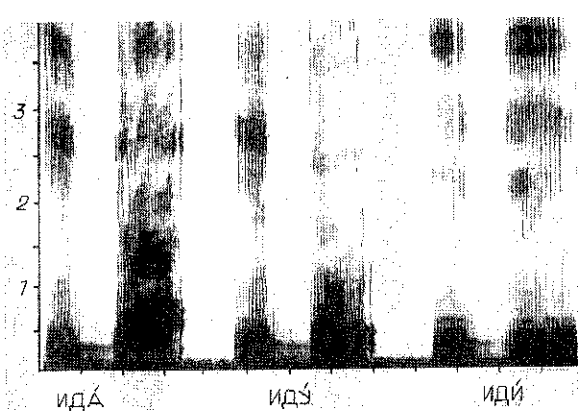
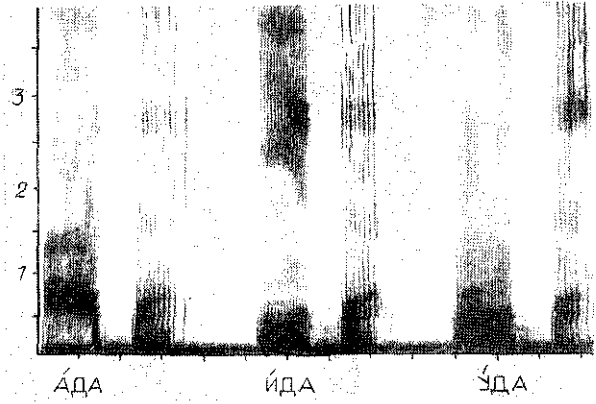
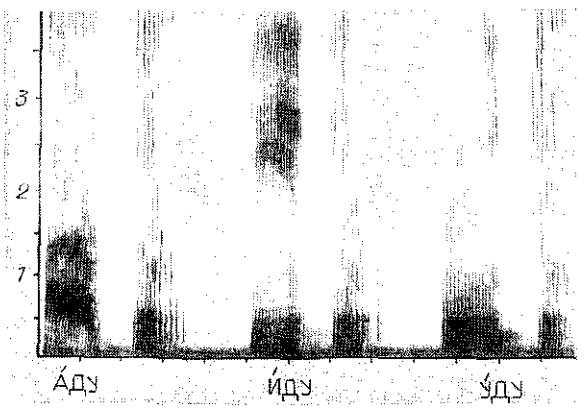


Рис. 6.10 (продолжение)





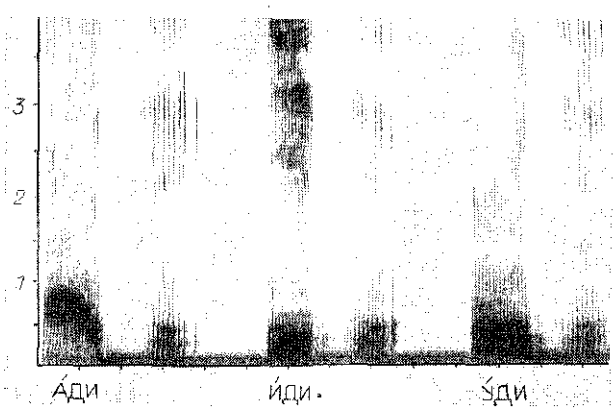
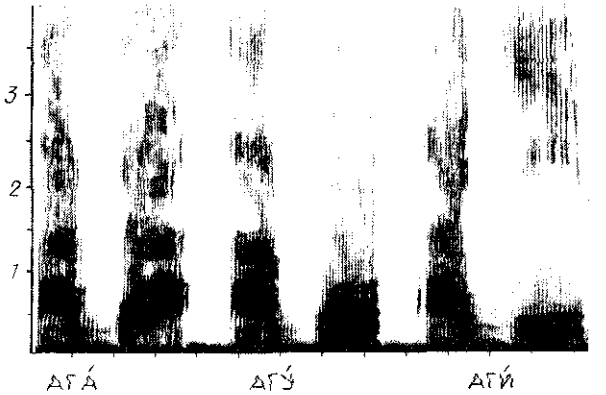
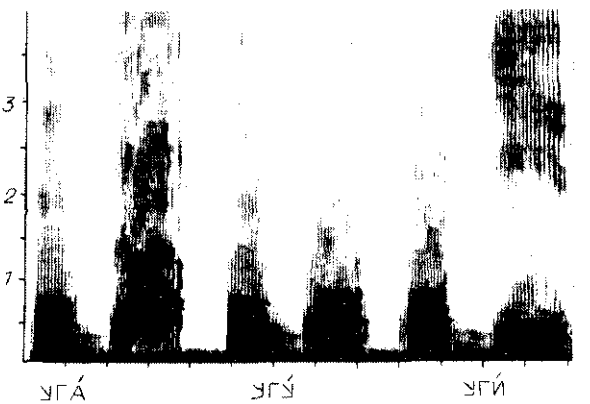


Рис. 6.10 (продолжение)





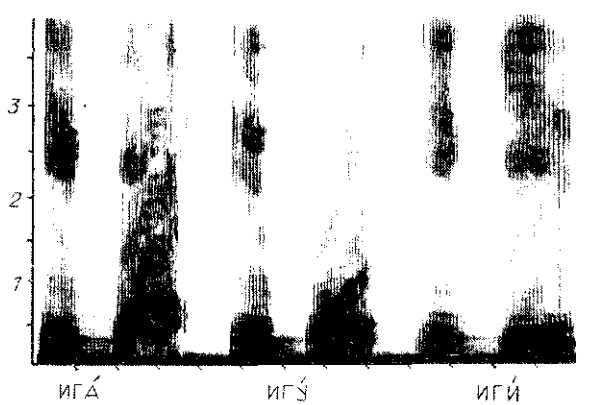
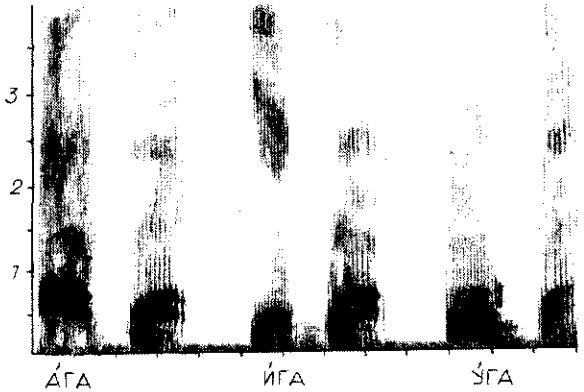
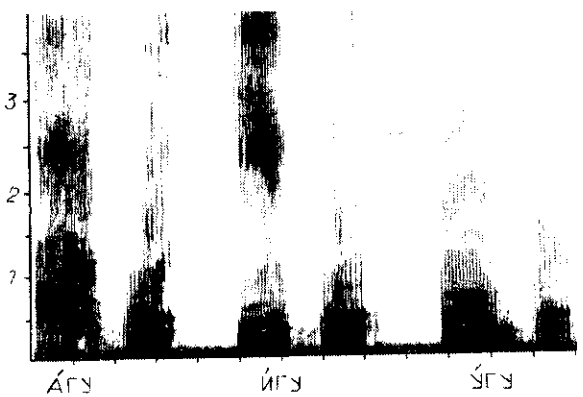


Рис. 6.10 (продолжение)





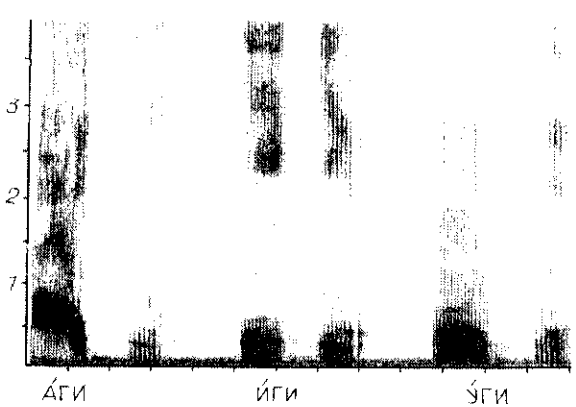
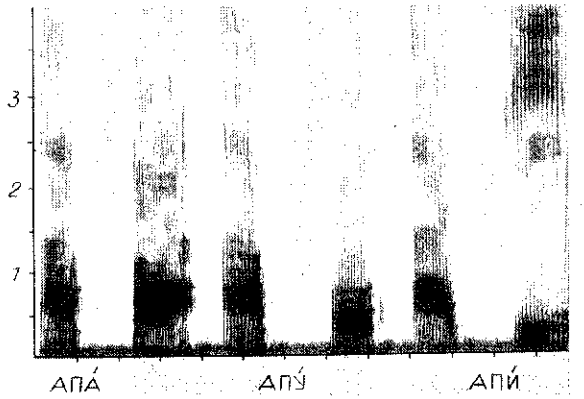
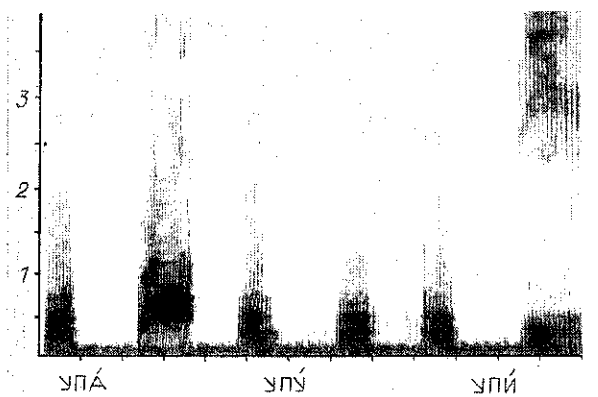


Рис. 6.10 (продолжение)





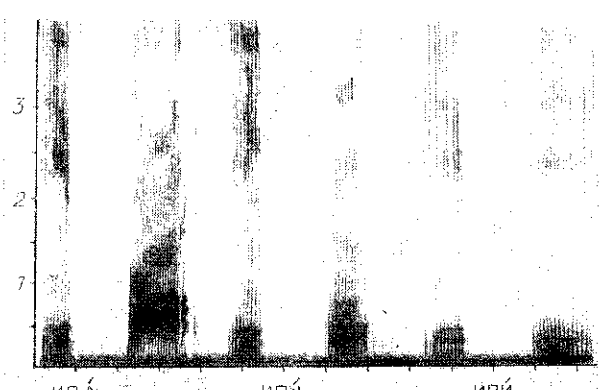
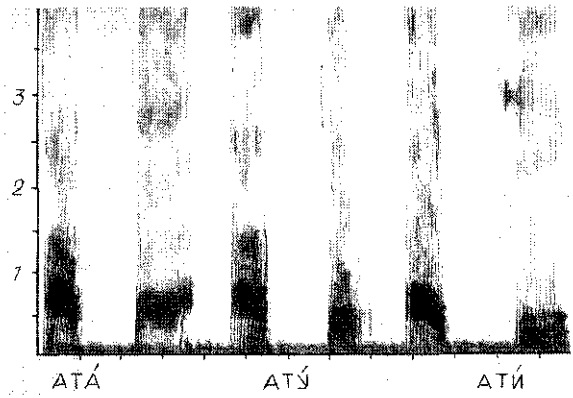
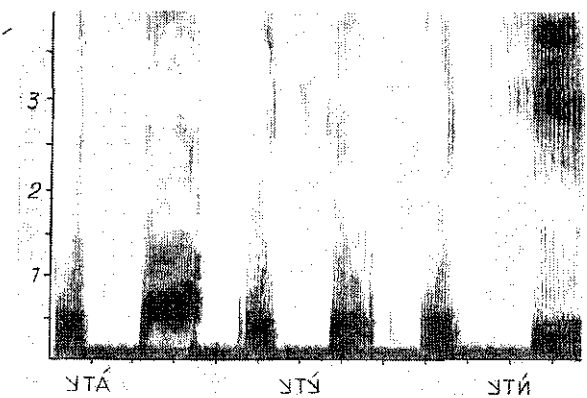


Рис. 6.10 (продолжение)





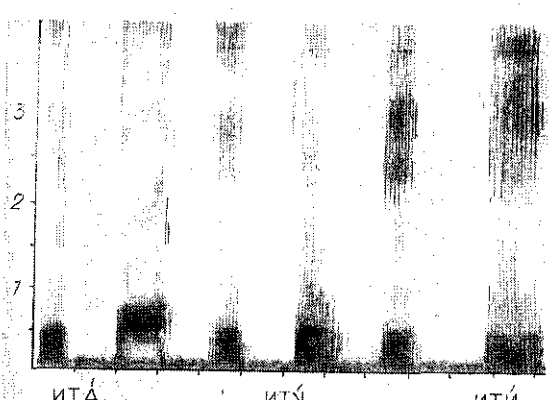
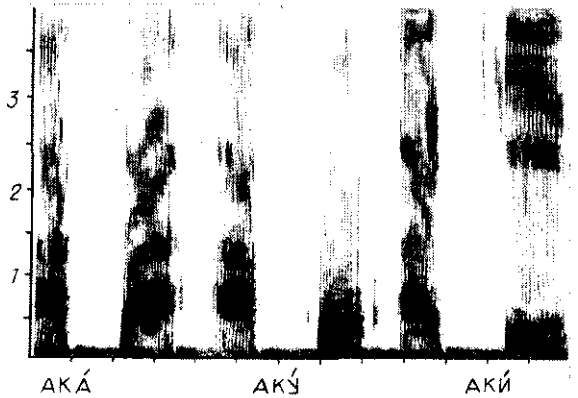
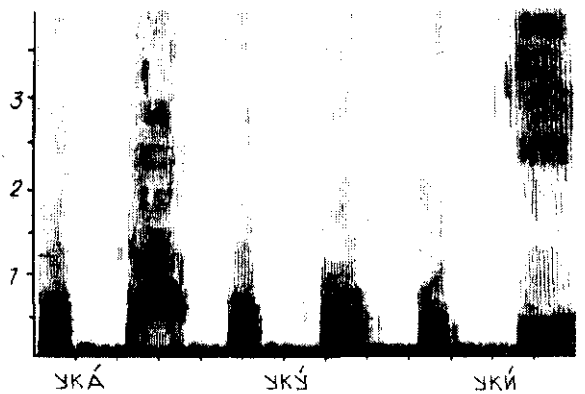


Рис. 6.10 (продолжение)





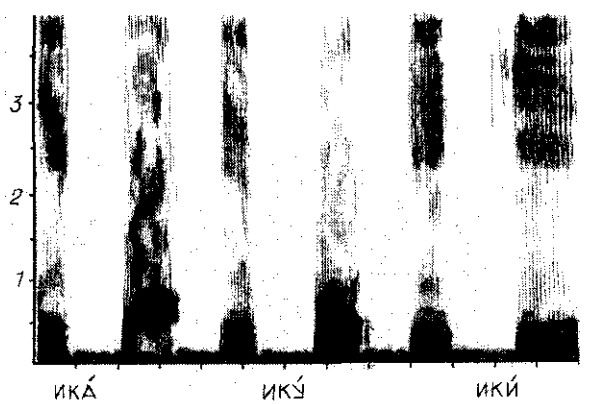


Рис. 6.10 (продолжение)

одним из эффектов коартикуляции — перехода речевого тракта к нейтральному состоянию. Следовательно, на степень коартикуляции влияет и положение гласного звука относительно ударной позиции.

При составлении звукосочетаний для исследования коартикуляции некоторое облегчение доставляет тот факт, что безударные гласные разделяются на три группы: А-образные, И-образные и У-образные. На рис. 6.10 показаны спектрограммы ГСГ слогов с согласными /Б, Д, Г, П, Т, К/ в несимметричном окружении гласных /А, У, И/ и ударением на первом или втором гласном. Из этих спектрограмм видно разнообразие эффектов коартикуляции, но видно также и то, что в ряде случаев очень трудно измерить начальные положения формант, которые так важны для количественного описания коартикуляции. Если для этой цели пользоваться только спектрограммами, то неизбежны ошибки в правилах управления синтезатором и, соответственно, снижение разборчивости синтетической речи. Поэтому анализ спектрограмм необходимо уточнять с помощью независимого анализа формантных частот речевого сигнала, что, конечно, является весьма трудной задачей. Другой путь в изучении коартикуляции состоит в использовании артикуляторно-формантного синтезатора, в котором на одном из этапов синтеза по заданной форме речевого тракта вычисляются формантные частоты. Подобрав команды управления, обеспечивающие необходимую разборчивость синтетической речи на артикуляторно-формантном синтезаторе, в дальнейшем можно проанализировать формантные переходы в любых контекстах, и найти такие правила для управления формантами, которые позволили бы отказаться от расчета собственных частот речевого тракта, т. е. ограничиться чисто формантным синтезом.

В заключении главы обсудим достоинства и недостатки формантного синтеза речи. К числу его достоинств относится сравнительная простота технической реализации, возможность формирования правил синтеза с помощью анализа легко читаемых спектрограмм, сравнительно высокая разборчивость синтетической речи. Поскольку в формантном синтезаторе все параметры поддаются явному управлению, то его можно приспосабливать к разным условиям. Например, для повышения разборчивости при использовании телефонного канала можно повысить уровень фрикативных звуков и взрывных участков глухих согласных, добавить короткий сегмент нейтрального гласного после звонкого согласного в конце слова и т. д. [137]. Формантные переходы хорошо аппроксимируются простыми линейными функциями. В некоторых пределах можно изменять индивидуальность голоса и даже имитировать женский голос. Все эти качества превращают формантный синтезатор в такое средство вывода информации в речевой форме, которое вполне соответствует запросам определенного круга потребителей.

С другой стороны, разборчивость формантного синтеза резко падает при наличии шумов в окружающей среде или канале связи, а также при необходимости выполнять какую-либо работу одновременно с прослушиванием синтезатора. Восприятие синтетической речи требует больших умственных усилий, чем восприятие естественной речи. Подробнее о перцептивных характеристиках синтетической речи будет говориться в гл. 11. Вследствие того, что правила управления синтезатором создаются на основе анализа речи одного диктора, очень трудно имитировать другие манеры артикуляции, а ограниченность используемых в синтезаторе моделей процессов речеобразования препятствует повышению разборчивости и натуральности, а также созданию различаемых индивидуальностей голоса. Составление правил управления синтезатором является скорее искусством, чем наукой, и требует исключительного фонетического чутья.

Вместе с тем, возможности усовершенствования формантного синтеза еще далеко не исчерпаны, и я надеюсь, что сведения, излагаемые в этой книге, будут способствовать дальнейшему развитию формантного синтеза.

ГЛАВА 7

АКУСТИЧЕСКИЕ ПРОЦЕССЫ В АРТИКУЛЯТОРНО-ФОРМАНТНОМ СИНТЕЗАТОРЕ

Синтез речевой волны по заданной площади поперечного сечения речевого тракта и источникам возбуждения в артикуляторном синтезаторе может быть осуществлен двумя способами: *непосредственным* решением уравнения речевого тракта и расчетом собственных частот и затуханий с последующим управлением параметрами формантного синтезатора. Каждый из этих способов имеет свои преимущества и недостатки, касающиеся как сложности вычислений, так и качества синтетической речи. Рассмотрим сначала различные способы вычисления формантных частот.

§ 7.1. Метод длинной линии

Исторически моделирование акустических процессов в речевом тракте начиналось с его представления в виде неоднородной длинной линии, которая обладает преимуществом почти полного аналога акустической системы (за исключением нелинейных аэродинамических процессов). Средствами аналоговой техники был реализован ряд электрических моделей речевого тракта [64, 93, 192], на которых исследовалась связь между его формой и акустическими характеристиками синтезированного сигнала. Однако в связи с трудностями управления параметрами длинной линии эти модели могли быть использованы лишь для синтеза гласных звуков с фиксированной площадью поперечного сечения тракта. Даже синтез фрикативных согласных в стационарном состоянии был практически невозможен, поскольку определение условий возникновения турбулентных шумов на каком-либо участке речевого тракта требует измерения тока в каждой секции длинной линии.

Поскольку вскоре было показано, что свойства речевого сигнала, в основном, характеризуются динамикой формантных частот, то уравнение длинной линии с целью поиска резонансных частот речевого тракта стали решать методами теории цепей. Такой подход позволяет использовать цифровые вычислительные машины для формирования площади поперечного

сечения тракта, а синтез речевого сигнала может быть осуществлен либо также в ЭВМ, либо путем управления формантным синтезатором.

В методе длинной линии речевой тракт разбивается на последовательность участков, на каждом из которых площадь поперечного сечения считается постоянной. В полосе частот речевого сигнала форма поперечного сечения влияет лишь на потери [59], но не на частотные характеристики. Поэтому обычно каждую секцию длинной линии рассматривают как электрический аналог отрезка цилиндрической трубы кругового сечения длиной Δl. Такое представление справедливо для всех длин волн $\lambda > 8\Delta l$ [64].

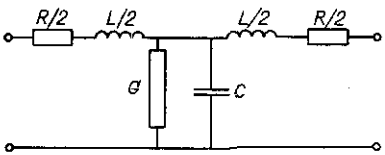


Рис. 7.1. Т-образный элемент длинной линии

Электрическая схема секции, представленная в виде Т-образного симметричного звена, показана на рис. 7.1. Здесь R_i отражает потери на вязкое трение и динамическое сопротивление в воздушном потоке, L_i и C_i — индуктивность и емкость линии:

$$L_i = \rho_0 / S_i, \quad C_i = S_i / (\rho_0 c_0^2),$$

где S_i — площадь поперечного сечения i-й секции, ρ_0 — плотность воздуха, c_0 — скорость звука. Проводимость $Y_i = 1/G_i$ равна сумме проводимостей $Y_{1i} + Y_{2i}$, где Y_{1i} — проводимость, связанная с потерями на теплопроводность

$$Y_{1i} = \mathscr{L} \frac{0{,}4}{\rho_0 c_0^2} \sqrt{\frac{k_{\text{тп}} \omega}{2\rho_0 c_0}},$$

$\mathscr{L}$ — периметр поперечного сечения, $k_{\text{тп}}$ — коэффициент теплопроводности, ω — круговая частота [65].

Импеданс стенок тракта Z_{i2} на участках с мягкими тканями носит инерционный характер, и проводимость

$$Y_{2i} = \frac{R_{2i} + j\omega M_{2i}}{R_{2i}^2 + \omega^2 M_{2i}^2},$$

где R_{2i} и M_{2i} — активная и реактивная составляющие импеданса стенок. Все перечисленные величины рассчитаны на единицу длины линии.

Для секции, показанной на рис. 7.1, известны соотношения между током I_i и напряжением E_i на ее входе и током I_{i+1} и напряжением E_{i+1} на ее выходе:

$$\begin{aligned} E_{i+1} &= E_i \operatorname{ch} \gamma_i \Delta l_i - I_i Z_{0i} \operatorname{sh} \gamma_i \Delta l_i, \\ I_{i+1} &= I_i \operatorname{ch} \gamma_i \Delta l_i - E_i \frac{1}{Z_{0i}} \operatorname{sh} \gamma_i \Delta l_i, \end{aligned} \tag{7.1}$$

где $Z_{0i} = \sqrt{L_i / C_i} = \rho_0 c_0 / S_i$ — характеристическое сопротивление

секции, а $\gamma=\sqrt{(G_i+j\omega C_i)(R_i+j\omega L_i)}$ — постоянная распространения.

Обычно считается, что речевой тракт начинается с голосовой щели, поэтому первая секция длинной линии замыкается на импеданс голосового источника, а последняя секция — на импеданс излучения.

Как видно, система (7.1) может быть решена путем последовательного исключения промежуточных переменных, после чего находится передаточная функция длинной линии как отношение, например, объемной скорости (тока I_N) на выходе линии к объемной скорости (току I_0) на ее входе:

$$H(\omega)=\frac{I_N}{I_0}=\frac{Z_{0N}}{\operatorname{sh}\gamma_N\Delta l_N}\cdot\frac{\Delta_{0N}}{\Delta},$$

где Δ — главный определитель системы (7.1), Δ_{0N} — соответствующий минор [64]. Передаточная функция $N(\omega)$ может быть представлена как

$$H(s)=\frac{1}{\prod\limits_{n=1}^{\infty}\left(1-\frac{s}{s_n}\right)\left(1-\frac{s}{s_n^*}\right)},$$

где s — комплексная частота $s=\sigma+j\omega$, s_n, s_n^* — комплексно-сопряженные пары полюсов, коэффициенты при мнимой части которых являются резонансными частотами речевого тракта. Например, для четырехтрубной модели речевого тракта в [64] получена следующая передаточная функция:

$$H(s)=\frac{1}{P(s)}=\frac{1}{\operatorname{ch}\gamma_1\Delta l_1\operatorname{ch}\gamma_2\Delta l_2\operatorname{ch}\gamma_3\Delta l_3\operatorname{ch}\gamma_4\Delta l_4(AB+CD)},$$

где

$$A=1+\frac{S_2}{S_1}\operatorname{th}\gamma_1\Delta l_1\operatorname{th}\gamma_2\Delta l_2,$$

$$B=1+\frac{S_4}{S_3}\operatorname{th}\gamma_3\Delta l_3\operatorname{th}\gamma_4\Delta l_4,$$

$$C=\frac{S_3}{S_2}\left(\operatorname{th}\gamma_3\Delta l_3+\frac{S_3}{S_4}\operatorname{th}\gamma_4\Delta l_4\right),$$

$$D=\frac{S_1}{S_2}(\operatorname{th}\gamma_1\Delta l_1+\operatorname{th}\gamma_2\Delta l_2),$$

где S_1, S_2, S_3, S_4 — площади труб.

Знаменатель передаточной функции $P(s)$ представляет собой полином, вообще говоря, бесконечной степени от переменной s, корни которого соответствуют резонансным частотам речевого тракта. Обычно осуществляют поиск только нескольких первых резонансов, а влияние высших резонансов учитывают поправочным членом в $P(s)$ (см. § 6.1).

Для нахождения резонансных частот по передаточной функции $H(s)$ используют различные приемы. Один из них состоит в сканировании частоты ω в пределах диапазона частот речевого сигнала и поиске максимумов модуля $H(\omega)$. Этот способ, однако, хорош лишь в том случае, когда резонансы сравнительно разнесены и максимумы огибающей спектра выражены достаточно четко. Если же резонансы близки, как, например, у гласного звука /О/, то поиск максимума может оказаться безуспешным. В этом случае может помочь анализ комплексной компоненты передаточной функции, поскольку известно, что на резонансной частоте вторая производная от фазы по частоте переходит через нуль [64].

Другой способ поиска резонансных частот состоит в определении нулей $P(s)$. Обычно для упрощения задачи полагают все потери равными нулю, и тогда корни $P(s)$ находятся средствами линейной алгебры. Следует, однако, отметить несомненное достоинство метода длинной линии, состоящее в том, что это метод анализа в частотной области, и он позволяет найти резонансы и их затухания даже в случае нелинейной зависимости параметров передаточной функции от частоты, а таких параметров достаточно много. К их числу относится импеданс излучения

$$Z_{\text{и}} = \frac{1}{2}\left(\frac{\omega a}{c_0}\right)^2 + j\frac{8\omega a}{3\pi c_0},$$

где a — эквивалентный радиус ротового отверстия. Вязкое трение R_l только для труб очень малого сечения не зависит от частоты, а для площадей, больших 0,5 см^2, это сопротивление зависит от корня квадратного частоты

$$R_l = \mathscr{L}_l \sqrt{\rho_0 \mu \omega / 2},$$

где μ — коэффициент вязкости воздуха, $\mathscr{L}_l$ — периметр поперечного сечения. Элементы проводимости стенок, как мы видели, также зависят от частоты ω. К тому же, при малых сечениях, когда действует закон капиллярного трения, величина вязкого сопротивления может стать настолько большой, что ею принципиально нельзя пренебречь при расчете резонансных частот. Поэтому в общем случае передаточная функция записывается как

$$H(s) = Q(s)/P(s).$$

Свойства $Q(s)$ и $P(s)$ определяются как частотными характеристиками элементов длинной линии, так и влиянием носовой полости или характеристиками турбулентного (или импульсного) источников возбуждения. Полюса $H(s)$ соответствуют нулям обратной функции $\bar{H}(s) = 1/H(s)$ и для поиска этих нулей можно применить известные процедуры итеративного

поиска корней. Пользуясь методом Ньютона [15], например, $(p+1)$-е приближение корня s_i найдем как

$$s_i^{(p+1)} = s_i^{(p)} - \left. \frac{1}{\bar{H}(s) \dfrac{d\bar{H}(s)}{ds}} \right|_{s=s_i^{(p)}}.$$

К числу недостатков метода расчета полюсов длинной линии следует отнести его вычислительную сложность. Несмотря на то, что в исследовательских целях с успехом использовалась четырехтрубная модель речевого тракта, на самом деле для достижения надлежащей точности определения полюсов в целях синтеза речи требуется 30—40 секций. Появление векторных и матричных процессоров, конечно, дает основание для пересмотра традиционных оценок сложности алгоритмов, однако, помимо высокого порядка системы, в методе длинной линии требуется производить довольно трудоемкие пересчеты физических параметров речевого тракта (площади сечения, потерь) в параметры передаточной функции $H(s)$. Поэтому с появлением новых методов решения уравнения речевого тракта, метод длинных линий используется все реже.

§ 7.2. Конечно-разностные схемы

Другой подход к поиску резонансных частот речевого тракта состоит в аппроксимации его уравнения конечно-разностными схемами и определении собственных чисел получаемой при этом системы линейных уравнений. В качестве примера рассмотрим уравнение речевого тракта относительно потенциала скорости акустических колебаний $\Phi(x, t)$ без учета потерь:

$$\frac{1}{S}\frac{\partial}{\partial x}\left(S\frac{\partial \Phi}{\partial x}\right) = \frac{1}{c_0^2}\frac{\partial^2 \Phi}{\partial t^2},$$

где

$$\Phi = -\operatorname{grad} v,$$

v — скорость колебаний частиц воздуха. К этому уравнению можно применить метод разделения переменных, положив

$$\Phi(x, t) = \varphi(x)\, q(t).$$

Это дает систему из двух уравнений, каждое из которых зависит только от одной переменной — либо от времени t, либо от пространственной координаты x:

$$\begin{aligned} q'' + \lambda^2 c_0^2 q &= 0, \\ \frac{1}{S}(S\varphi')' + \lambda^2 \varphi &= 0, \end{aligned} \tag{7.2}$$

где в первом уравнении штрих означает производную по t, а во втором уравнении — производную по x, λ — искомое

собственное число ($\lambda = \omega/c_0$), которое может принимать счетное множество значений.

Для определения λ представим (7.2) в конечно-разностной форме, воспользовавшись формулами для вычисления производных, полученными в гл. 2:

$$\varphi'(x) = \frac{\varphi_{i+1} - \varphi_{i-1}}{2\Delta x},$$

$$\varphi''(x) = \frac{\varphi_{i+1} - 2\varphi_i + \varphi_{i-1}}{\Delta x^2},$$

что дает следующую систему уравнений:

$$\varphi_{i+1} = \frac{4S_i}{S_{i+1} + 4S_i - S_{i-1}} \left[\frac{S_{i+1} - 4S_i - S_{i-1}}{4S_i} \varphi_{i-1} + \left(2 - \lambda^2 \Delta x^2\right) \varphi_i \right],$$

$$i = 2, \ldots, N-1. \qquad (7.3)$$

Эта система должна быть дополнена граничными условиями в точках $x = 0$ и $x = l$. Напомним, что акустическое давление связано с потенциалом скорости как

$$P = \rho_0 \frac{\partial \Phi}{\partial t},$$

а *скорость акустических колебаний*

$$v = -\partial \Phi / \partial x.$$

Воспользовавшись определением импеданса как отношением давления к скорости $Z = P/v$, имеем

$$\rho_0 \frac{\partial \Phi}{\partial t} = -Z \frac{\partial \Phi}{\partial x}.$$

Поскольку $\Phi(x, t) = \varphi(x)\, q(t)$, то

$$\partial \Phi / \partial x = \varphi'(x)\, q(t),$$

$$\partial \Phi / \partial t = \varphi(x)\, q'(t),$$

где штрих означает производную по той переменной, от которой зависит функция φ или q. Взятие производной по времени, как известно, для гармонических процессов соответствует умножению на $j\omega$, где j — мнимая единица. Поэтому граничные условия можно представить как

$$\rho_0 \varphi - \frac{Z}{j\omega} \varphi' = 0.$$

При закрытой голосовой щели импеданс сомкнутых голосовых складок весьма велик [59], и в это время граничное условие есть $\varphi_0' = 0$, или, в конечно-разностной форме

$$(\varphi_1 - \varphi_0)/\Delta x = 0.$$

Равенство $\varphi'=0$, или, что то же самое, $\partial\Phi/\partial x=0$, означает равенство нулю скорости акустических колебаний на абсолютно жесткой стенке, к свойствам которой близки свойства тканей голосовых складок в сомкнутом состоянии. При открытой голосовой щели на граничные условия влияет импеданс Z_0 голосовой щели, причем этот импеданс зависит от площади голосовой щели, и как мы видели в гл. 5, формантные частоты в это время подвергаются изменениям.

В конечной точке $i=N$, т. е. на губах, действует импеданс излучения $Z_{\text{и}}$, поэтому, воспользовавшись определением левой производной, имеем

$$\varphi_N-\frac{Z_{\text{и}}}{j\omega\rho_0}\frac{\varphi_N-\varphi_{N-1}}{\Delta x}=0.$$

Таким образом, располагая сведениями об импедансах голосовой щели Z_0 и излучении $Z_{\text{и}}$, мы имеем возможность решить систему (7.3), найдя ее нули, т. е. определить собственные числа λ_k и резонансные частоты $\omega_k=\lambda_k c_0$ для речевого тракта с заданной формой $S(x)$.

Число точек в конечно-разностной форме такое же, как и в методе длинной линии, так как дискретное представление $\varphi(x)$ в конечном числе точек соответствует аппроксимации $S(x)$ последовательности цилиндрических секций. Для обеспечения надлежащей точности описания артикуляционных процессов число точек N должно быть не меньше 30—40, но сложность решения системы (7.3) ниже, чем сложность решения системы (7.1), поскольку для (7.3) не требуется пересчитывать площадь поперечного сечения в параметры длинной линии. В обоих методах имеется возможность расчета собственных частот и для разветвленной акустической системы, образующей при артикуляции звуков с опущенной небной занавеской и распространением колебаний через носовую полость. Число секций или точек отсчета при этом увеличивается примерно в полтора раза. Соответственно увеличивается и число искомых собственных частот.

§ 7.3. Метод Галеркина

Метод поиска собственных чисел волнового уравнения, основанный на разложении Ритца—Галеркина, позволяет уменьшить порядок решаемой системы уравнения в несколько раз. Метод Галеркина состоит в следующем [22]. Если имеется некоторое дифференциальное уравнение $L(\omega)=0$ с однородными граничными условиями, то его приближенное решение можно найти в виде

$$\tilde{U}(x)=\sum_{i=1}^{n} c_i\psi_i^{(0)}(x),$$

где $\{\psi_i^{(0)}\}$ — полная система линейно-независимых функций,

удовлетворяющих тем же граничным условиям. Для того чтобы $\tilde{U}(x)$ составляло решение данного дифференциального уравнения, необходимо, чтобы $L(\tilde{U})=0$, а это означает требование ортогональности $L(\tilde{U})$ ко всем функциям $\psi_i^{(0)}(x)$. Это условие приводит к системе линейных, относительно коэффициентов c_i, уравнений:

$$\int_D L\Big[\sum_{j=1}^{n} c_j \psi_j^{(0)}(x)\Big]\,\psi_i^{(0)}(x)\,dx=0.$$

Решая эту систему для первых n функций, получаем коэффициенты c_i.

В применении к анализу уравнения речевого тракта метод Галеркина описан в [59]. Рассмотрим уравнение речевого тракта относительно давления в предположении, что площадь поперечного сечения S меняется во времени настолько медленно, что производной $\partial S/\partial t$ можно пренебречь. Тогда имеем

$$\frac{1}{S}\frac{\partial}{\partial x}\left(S\frac{\partial P}{\partial x}\right)=\frac{1}{c_0^2}\frac{\partial^2 P}{\partial t^2},$$

и, применяя метод разделения переменных $P(x, t)=p(x)\,q(t)$, получим систему

$$q''+\lambda^2 c_0^2 q=0,$$
$$(Sp')'+\lambda^2 Sp=0.$$

Последнее уравнение этой системы является обыкновенным дифференциальным уравнением. Необходимо найти собственные числа λ этого уравнения и его собственные функции $\psi(x)$. Согласно методу Галеркина, представим искомые собственные функции в виде

$$\psi(x)=\sum_{i=1}^{n} c_i \psi_i^{(0)}(x),$$

и условия ортогональности запишем как

$$\int_0^l \left[(S\psi_k')'+\lambda_k^2 s\psi_k\right]\psi_j^{(0)}(x)\,dx=0,$$

где l — длина речевого тракта. Для неназализованных звуков это уравнение полностью определяет собственные числа и коэффициенты через решение системы

$$\sum_{j=1}^{n}\left(A_{ij}+\lambda^2 B_{ij}\right)c_j=0,\quad j=1, 2, \ldots, n, \tag{7.4}$$

где

$$A_{ij}=S\psi_i^{(0)}\psi_j^{(0)\prime}\,|_0^l-\int_0^l S\psi_i^{(0)\prime}\psi_j^{(0)\prime}\,dx=0,$$

$$B_{ij}=\int_0^l S\psi_i^{(0)}\psi_j^{(0)}\,dx.$$

Используя условие нормированности собственных функций

$$\int_0^l S\psi_j^2(x)\,dx = 1 \tag{7.5}$$

получаем $2n$ уравнений, необходимых для определения $2n$ неизвестных — коэффициентов c_i и собственных чисел λ_i. Как известно, система (7.4), (7.5) разрешима только в том случае, когда ее определитель равен нулю, что и дает алгоритм вычисления.

При использовании специальных приемов решения задачи поиска собственных чисел и коэффициентов при собственных функциях число требуемых операций пропорционально n^3, где n — порядок приближения. Одним из таких приемов является схема Гаусса с выбором главных элементов. В этом методе сначала переставляются столбцы и строки решаемой системы так, чтобы на диагонали матрицы оказались элементы с наибольшими по модулю значениями. Затем осуществляется последовательное исключение неизвестных, в результате чего матрица превращается в верхнюю треугольную. Такая матрица обладает тем свойством, что ее определитель равен произведению диагональных элементов. Приравнивая нулю это произведение, получаем характеристический многочлен n-й степени, корни которого являются искомыми собственными числами λ_i. После нахождения собственных чисел определение коэффициентов c_i не представляет труда.

Существует ряд других методов поиска собственных чисел λ_i и коэффициентов c_i, в том числе итерационных. Среди них отметим метод, напоминающий метод Релея. Приведем матрицу (7.4) к виду, в котором наибольшие по модулю элементы располагаются на ее диагонали, на первом шаге итеративного процесса оценим собственные числа как

$$\lambda_j^2 = \int_0^l \left(S\psi_i^{(0)\prime}\right)\psi_j^{(0)}\,dx \Big/ \int_0^l S\psi_i^{(0)}\psi_j^{(0)}\,dx.$$

Затем найдем коэффициенты c_j, с их помощью уточним значения λ_j и т. д. Этот итеративный процесс может быть организован разными способами, в частности, положив на первом шаге коэффициенты при недиагональных элементах равными нулю.

В методе Галеркина порядок приближения ограничен снизу числом искомых собственных чисел. Если, например, для неназализованных звуков нужно найти 5 первых собственных чисел (соответствующих пяти первым резонансам), то и порядок приближения $n=5$. Как мы увидим ниже, для назальных звуков число резонансов почти удваивается по сравнению с неназализованными звуками, и порядок приближения оказывается $n=10$. Значит, для решения системы (7.4) потребуется около 10^3 операций. Фактически оказывается, что число

операций определяется не столько алгоритмом поиска собственных чисел λ_i и коэффициентов c_i при собственных функциях, сколько затратами на вычисление интегралов для A_{ij} и B_{ij}. Здесь полезны рекурсивные схемы вычисления интегралов, рассмотренные нами в гл. 2.

Как уже говорилось, система функций нулевого приближения $\{\psi_i^{(0)}\}$ должна удовлетворять тем же граничным условиям, что и исходное дифференциальное уравнение. Эту систему удобно определить, как решение волнового уравнения для трубы (т. е. с постоянной площадью поперечного сечения $S=\text{const}$). Покажем схему определения $\psi_i^{(0)}(x)$ и поиска решения методом Галеркина для разветвленного речевого тракта в случае артикуляции назализованных звуков. Эта же схема дает решение и для неназализованных звуков, как частный случай.

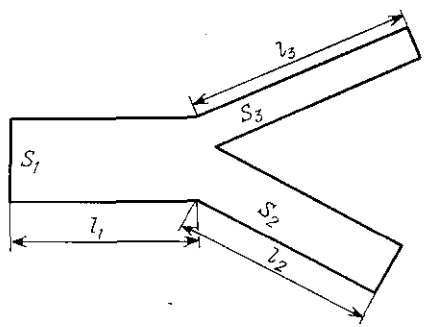


Рис. 7.2. Разветвленная акустическая система

Представим модель речевого тракта нулевого приближения в виде трех однородных труб длиной l_1, l_2 и l_3 и площадью S_1, S_2 и S_3 (рис. 7.2). Как было показано в [59], для такой модели система собственных функций нулевого приближения $\{\psi_i^{(0)}(x)\}$ есть

$$\psi_1^{(0)}(x_1)=c_k\cos\lambda_k^{(0)}x_1,$$

$$\psi_2^{(0)}(x_2)=c_k\cos\lambda_k^{(0)}l_1\left(\cos\lambda_k^{(0)}x_2+\frac{\lambda_k^{(0)}\cos\lambda_k^{(0)}l_2-G_2}{G_2\cos\lambda_k^{(0)}l_2+\lambda_k^{(0)}}\sin\lambda_k^{(0)}x_2\right),$$

$$\psi_3^{(0)}(x_3)=c_k\cos\lambda_k^{(0)}l_1\left(\cos\lambda_k^{(0)}x_3+\frac{\lambda_k^{(0)}\cos\lambda_k^{(0)}l_3-G_3}{G_3\cos\lambda_k^{(0)}l_3+\lambda_k^{(0)}}\sin\lambda_k^{(0)}x_3\right),$$

где x_1, x_2, x_3 — координаты вдоль каждой из труб. Коэффициенты G_2 и G_3 определяются из условий излучения для второй и третьей трубы:

$$G_2=\frac{3\pi^2}{8S_2}\sqrt{\frac{S_2(l_2)}{\pi}},\qquad G_3=\frac{3\pi^2}{8S_3}\sqrt{\frac{S_3(l_3)}{\pi}},$$

причем подразумевается, что площади излучающего отверстия $S_2(l_2)$, а также $S_3(l_3)$, могут быть меньше площади трубы S_2 или S_3. Собственные числа нулевого приближения такой системы находятся из

$$S_1\operatorname{tg}\lambda_k^{(0)}l_1+S_2\frac{\lambda_k^{(0)}\operatorname{tg}\lambda_k^{(0)}l_2-G_2}{G_2\operatorname{tg}\lambda_k^{(0)}l_2+\lambda_k^{(0)}}+S_3\frac{\lambda_k^{(0)}\operatorname{tg}\lambda_k^{(0)}l_3-G_3}{G_3\operatorname{tg}\lambda_k^{(0)}l_2+\lambda_k^{(0)}}=0.$$

Коэффициенты нормировки c_k найдем из условия

$$\int_0^{l_1} \psi_{1k}^{(0)2}\, dx_1 + \int_0^{l_2} \psi_{2k}^{(0)2}\, dx_2 + \int_0^{l_3} \psi_{3k}^{(0)2}\, dx_3 = 1.$$

Поскольку здесь для более точного отражения условий в речевом тракте предусматривается отличие площадей труб S_1, S_2 и S_3 и их окончаний, то нужно найти способ их определения из реальной конфигурации речевого тракта. В [59] используется средняя площадь

$$S_i = \frac{1}{l_i} \int_0^{l_i} S^{(i)}(x_i)\, dx_i.$$

Согласно гл. 2, это оптимальное значение при аппроксимации $S(x)$ полиномом нулевого порядка. Возможно, что более точным будем определение S_i не как среднеарифметической величины, а как среднегеометрической, учитывая тот факт, что на самом деле площадь поперечного сечения входит в уравнение речевого тракта в виде логарифма:

$$\frac{\partial^2 P}{\partial x^2} + \frac{\partial (\ln S)}{\partial x} \frac{\partial P}{\partial x} = \frac{1}{c_0^2} \frac{\partial^2 P}{\partial t^2}.$$

Известно также, что наибольшее влияние на резонансные частоты оказывают участки с наименьшей площадью поперечного сечения. Поэтому площади труб нулевого приближения можно определить как

$$\ln S_i = \frac{1}{l_i} \int_0^{l_i} \ln S^{(i)}(x_i)\, dx_i.$$

Подставляя функции нулевого приближения $\psi_k^{(0)}$ в уравнения для каждого участка речевого тракта и учитывая условие неразрывности, в конечном счете получаем систему уравнений

$$\sum_{j=1}^{n} \left[A_{kj}^{(0)} + A_{kj}^{(1)} + A_{kj}^{(2)} + \lambda^2 \left(B_{kj}^{(0)} + B_{kj}^{(1)} + B_{kj}^{(2)} \right) \right] c_j = 0,$$

где

$$A_{kj}^{(i)} = \int_0^{l_i} \left[S^{(i)} \psi_{ik}^{(0)\prime} \right]' \psi_{ij}^{(0)}\, dx_i,$$

$$B_{kj}^{(i)} = \int_0^{l_i} S^{(i)} \psi_{ik}^{(0)} \psi_{ij}^{(0)}\, dx_i.$$

Решая эту систему, находим требуемое количество собственных чисел λ и соответствующих им собственных функций ψ.

Эксперименты по синтезу назальных звуков показывают, что приемлемая точность решения достигается при порядке

приближения $n=10$ *и интервале дискретизации функции* $S(x)$, равном $\Delta x=0{,}5$ см. При длине тракта в 17,5 см это соответствует 36 отсчетам, так что по сравнению с методом длинных линий или конечно-разностных уравнений порядок решаемой системы ниже в 3,6 раза, а количество вычислительных операций почти в 50 раз меньше, если принять их кубическую зависимость от порядка системы.

До сих пор решение уравнения речевого тракта производилось в предположении отсутствия потерь. Зная, однако, резонансные частоты и собственных функции, можно найти и коэффициенты затухания (или ширину полосы формант). При этом так же, как и в методе длинной линии, не составляет никакой трудности тот факт, что ряд видов потерь зависит от частоты, иногда даже и нелинейно. В [59] было показано, что коэффициент затухания колебаний на k-й резонансной частоте можно найти как

$$\delta_k=\int_0^l QS\psi_k\,dx/(2\rho_0\int_0^l S\psi_k^2\,dx),$$

где ψ_k — k-я собственная функция, $Q(x)$ — распределенные потери в речевом тракте. Эти потери складываются из потерь на вязкое трение вблизи стенок тракта, теплопроводности и податливости стенок, потерь на излучение в атмосферу и в легкие. Все эти виды потерь подробно рассмотрены в [59]. Сопротивление вязкого трения есть

$$Q_{\text{отр}}=\frac{k_\Phi}{\sqrt{S}}\sqrt{2\pi\rho_0\mu\omega},$$

где k_Φ — коэффициент формы, $k_\Phi=\mathscr{L}/2\sqrt{\pi S}$, $\mathscr{L}$ — периметр сечения. Для кругового сечения $k_\Phi=1$. Потери на теплопроводность составляют примерно половину от потерь на вязкое трение и их обычно учитывают кажущимся увеличением вязкости. Потери на колебания стенок представляют как

$$Q_{\text{ст}}=\frac{R_{\text{ст}}}{\mathscr{L}(R_{\text{ст}}^2+\omega^2M_{\text{ст}}^2)},$$

где $R_{\text{ст}}\approx 10$ г/(см^{4}с), $M_{\text{ст}}\approx 0{,}02$ г/см^4. Потери на излучение есть

$$Q_{\text{и}}=\frac{S_{\text{и}}\omega^2}{2\pi c_0^2},$$

где $S_{\text{и}}$ — площадь излучающего отверстия. Потери в голосовой щели складываются из потерь на динамическое трение и потерь на капиллярное вязкое трение:

$$Q_{\text{г}}=\frac{\rho_0 c_x v}{2S_{\text{г}}}+\frac{12\mu hd}{S_{\text{г}}^3},$$

где $c_x=0{,}7$, v — скорость воздушного потока через голосовую

щель, h — длина голосовой щели, d — длина голосовых складок. Это собственные потери в голосовой щели. К ним следует добавить потери в подсвязочной области, которые мы здесь не будем рассматривать, отсылая к [59].

Наконец, при возникновении турбулентных процессов в речевом тракте, необходимо учесть потери на турбулизацию

$$c_{\text{тб}} = \left(\frac{S_2}{S_1} - 1\right)^2 \frac{v}{c_0},$$

где S_1 — площадь сечения, в котором возникла турбулентность, S_2 — площадь расширяющегося участка за сужением, v — скорость воздушного потока в сужении. Коэффициент $c_{\text{тб}}$ используется для расчета динамических потерь как

$$Q_{\text{тб}} = \rho_0 c_{\text{тб}} v/(2S_1).$$

Расчет ширины полос формант по перечисленным потерям хорошо соответствует измерениям на реальных звуках речи.

Все потери зависят от площади поперечного сечения речевого тракта $S(x, t)$ и изменяются вместе с ней. В голосовой щели эти изменения происходят довольно быстро и в большом диапазоне, в результате чего затухания резонансных колебаний сильно отличаются на интервалах открытой и закрытой голосовой щели. Иногда можно наблюдать почти полное прекращение колебаний на частоте первого резонанса при открытой голосовой щели.

Так же, как и источник голосового возбуждения, распределенными вдоль речевого тракта оказываются и источники турбулентного и шумового возбуждения. Хотя эти последние источники имеют небольшое пространственное протяжение, они могут находиться в различных местах речевого тракта. Отвлекаясь в данном случае от того, что для турбулентного источника волновое уравнение недействительно и нужно пользоваться уравнениями Навье — Стокса, будем считать турбулентные источники внешними источниками, и укажем единый для *всех видов источников способ определения амплитуд возбуждения* для каждого резонанса в речевом тракте.

Пусть, в общем виде, в речевом тракте имеется распределенный источник $F(x, t)$. Тогда каждый k-й резонанс возбуждается функцией $F_k(t)$, зависящей только от времени, и

$$F_k(t) = \frac{2\int_0^l F(x, t)\, S(x, t)\, \psi_k(x)\, dx}{\rho_0 l \int_0^l S(x, t)\, \psi_k^2(x)\, dx}.$$

Это означает, что относительные амплитуды возбуждения резонансных колебаний зависят не только от формы возбуждающейся функции, но и от формы собственных функций, *соответствующих каждому резонансу*, т. е. *в конечном счете* и от формы речевого тракта.

Свойства $F_k(t)$ уже рассматривались в § 5.7 с использованием дельта-функции. К такому виду можно привести, например, турбулентный и импульсный источники, считая, что они сосредоточены в одной точке с координатой x_0. В этом случае для $F_k(t)$ было получено

$$F_k(t)=\frac{2f(t)\,S(x_0,\,t)\,\psi_k(x_0)}{\rho_0 l\int\limits_0^l S(x,\,t)\,\psi_k^2(x)\,dx}.$$

Если x_0 совпадает с одним из нулей собственной функции ψ_k, то и колебания на резонансной частоте ω_k не возникнут, поскольку амплитуда возбуждения F_k на этой частоте равна нулю. Это объясняет полосовой характер спектра фрикативных согласных, хотя спектр турбулентного источника непрерывен.

§ 7.4. Метод прогонки

Имеется ряд методов поиска собственных частот с меньшими затратами, чем в рассмотренных выше методах, однако обычно они применимы только к расчету характеристик неразветвленного речевого тракта. Пусть, например, уравнение речевого тракта записано относительно давления

$$\frac{1}{S}\frac{\partial}{\partial x}\left(S\frac{\partial P}{\partial x}\right)=\frac{1}{c_0^2}\frac{\partial^2 P}{\partial t^2}.$$

Тогда, применяя метод разделения переменных, получим уравнение для пространственной компоненты:

$$(Sp')'+\lambda^2 Sp=0. \tag{7.6}$$

Взяв левые производные в конечно-разностной схеме, имеем

$$p_{i+1}=(1-\lambda^2\Delta x^2)\,p_i+\frac{S_{i-1}}{S_i}(p_i-p_{i-1}).$$

Собственные числа этого уравнения можно искать в предположении, что на губах давление в точности равно нулю, т. е. пренебрегая сопротивлением излучения. Пробуя разные значения λ, нужно добиться того, чтобы на конце тракта было $p_n=0$ и для k-го резонанса оказалось бы точно k нулей. Предположение о нулевом импедансе излучения, однако, применимо только к гласным звукам и то с большой погрешностью. Поэтому учет импеданса излучения обязателен. Это можно сделать с помощью метода прогонки [11]. Применяя к (7.6) конечно-разностную схему симметричной производной, получим

$$a_i p_{i-1}+b_i p_i+c_i p_{i+1}=0,$$

где

$$a_i=(S_{i-1}+4S_i-S_{i+1})/S_i,$$

$$b_i=4(\lambda^2\Delta x^2-2),$$

$$c_i=(S_{i+1}+4S_i-S_{i-1})/S_i,$$

а граничные условия есть

$$\rho_0 p_0-\frac{Z_0}{j\lambda c_0}\frac{p_1-p_0}{\Delta x}=0, \qquad \rho_0 p_n-\frac{Z_n}{j\lambda c_0}\frac{p_n-p_{n-1}}{\Delta x}=0.$$

Метод прогонки состоит в том, что сначала для возрастающих индексов i от 0 до n находят коэффициенты

$$L_{i+1/2}=-\frac{c_i}{b_i+a_iL_{i-1/2}}, \qquad K_{i+1/2}=-\frac{a_iK_{i-1/2}}{b_i+a_iL_{i-1/2}},$$

где $L_{1/2}=0$, $K_{1/2}=p_0$. Затем для убывающих индексов i от $n-1$ до 1 определяют искомую переменную p как

$$p_i=L_{i+1/2}p_{i+1}+K_{i+1/2}.$$

Если параметр λ выбран неправильно, то при обратной прогонке произойдет несовпадение с граничными условиями при $i=0$. Оценивая это несовпадение, можно скорректировать λ таким образом, чтобы удовлетворить граничным условиям на обоих концах речевого тракта. Таким образом, будут найдены *собственные числа и соответствующие им собственные функции*. Отметим, что в методе прогонки для решения системы из n уравнений требуется число операций, пропорциональное n, а не n^3, которое приходится затрачивать, например, в методе исключения. Поэтому теоретически метод прогонки является эффективным средством решения краевых задач.

В задаче расчета резонансных частот речевого тракта имеются особенности, затрудняющие создание устойчивой вычислительной схемы. Как будет более подробно обсуждаться в § 7.6, условие устойчивости для метода прогонки есть

$$|b_i|\geqslant|a_i|+|c_i|+\delta,$$

где $\delta>0$, $1\leqslant i\leqslant n$. Для уравнения (7.6) это условие не выполняется, и, таким образом, прогонка может оказаться неустойчивой. В действительности же удается добиться устойчивого решения за счет уменьшения шага Δx до величин порядка 0,01 см. Анализ поведения вычислительного процесса показывает, что при больших Δx неустойчивость начинает развиваться в точке резкого изменения площади поперечного сечения $S(x)$, т. е. в точке аномально большой производной $\partial S/\partial x$.

Эта точка обычно находится в области соединения фарингиальной полости с ротовой полостью у корня языка. В ряде случаев наблюдается даже разрыв функции $S(x)$, т. е. бесконечно большое значение первой производной по x.

Учитывая эту особенность, целесообразно разбить речевой тракт на два участка, на каждом из которых функция $S(x)$ достаточно гладкая, и сшивать решение в точке разрыва. Это приводит к схеме встречной прогонки, где начальной является точка разрыва, а коэффициенты a_i, b_i и c_i и значения собственной функции вычисляются сначала по расходящимся, а затем — по сходящимся путям. Условиями сшивания решения в точке разрыва x_0 служат непрерывность давления и объемной скорости:

$$p^-(x_0) = p^+(x_0),$$

$$S^-(x_0)\frac{\partial p^-(x)}{\partial x}\bigg|_{x=x_0} = S^+(x_0)\frac{\partial p^+(x)}{\partial x}\bigg|_{x=x_0},$$

где p^- и p^+ — давление в точке x_0 при интегрировании слева и справа от x_0, $\partial p^-/\partial x$ и $\partial p^+/\partial x$ — производные в точке x_0 при подходе слева и справа, $S^-(x_0)$ и $S^+(x_0)$ — значения площади слева и справа от x_0. Такая схема устойчива при весьма крупном шаге $\Delta x = 0{,}5$ см, и средняя погрешность для первых пяти резонансных частот однородной трубы не превышает 5%. Число итераций, требующихся для достижения минимума невязки, обычно близко к 4—5. По скорости вычислений метод встречной прогонки примерно в 25 раз быстрее метода Галеркина. В силу простоты вычислительной схемы объем программного обеспечения для метода встречной прогонки значительно меньше, чем для метода Галеркина.

Метод встречной прогонки пригоден и для расчета резонансов разветвленной акустической системы, как в случае назализованных звуков. Скорость вычислений при этом падает всего лишь в два-три раза по сравнению с неразветвленным речевым трактом.

Еще один вычислительно простой метод состоит в решении (7.6) не как краевой задачи, а как задачи с начальными условиями. Выбирая одно из начальных условий на каком-то конце речевого тракта произвольным, можно добиться удовлетворения начального условия на другом конце тракта итеративными уменьшениями невязки. По окончании процесса вычислений собственные функции нормируются, так что произвольный выбор одного из начальных условий не имеет значения. Иногда этот метод называют методом пристрелки. Для обеспечения устойчивости вычислительной схемы функция давления $P(x)$ рассчитывается как решение обыкновенного дифференциального уравнения по § 2.5. Метод пристрелки быстрее обычного метода прогонки, но в два раза медленнее метода встречной прогонки. Средняя погрешность для первых пяти резонансов однородной трубы составляет около 10%, и ошибки быстро возрастают при понижении частоты первого резонанса до частот около 200 Гц.

Методы встречной прогонки или пристрелки решают задачу вычисления резонансов и собственных функций речевого тракта по его площади поперечного сечения в реальном масштабе времени на процессорах умеренной мощности.

§ 7.5. Артикуляторно-формантный синтезатор

Из приведенного анализа складывается следующая вычислительная схема артикуляторно-формантного синтезатора. Выходными сигналами для синтезатора служат площадь поперечного сечения речевого тракта $S(x, t)$ и сигналы источников возбуждения: голосового (вместе с поршневым), турбулентного и импульсного. Функция $S(x, t)$ используется для вычисления собственных частот ω_k и собственных функций ψ_k речевого тракта. По S и ψ_k рассчитываются коэффициенты затухания δ_k и возбуждающие сигналы F_k, которые вместе с частотами ω_k управляют процессом генерирования речевого сигнала в формантном синтезаторе. Как видно, по сравнению с традиционной схемой формантного синтезатора, здесь полностью автоматизирован переход от площади поперечного сечения к частотным характеристикам речевого тракта и, к тому же, автоматически учитывается форма тракта в возбуждении резонансов, что никак не принимается во внимание в формантных синтезаторах. Поскольку в артикуляторно-формантном синтезаторе используется метод разделения переменных, то амплитуда каждого резонансного колебания зависит от значения соответствующей собственной функции на конце речевого тракта

$$P_k(t) = \psi_k(l) e^{-\delta_k t} \cos \psi_k t,$$

что более точно отражает влияние формы тракта на соотношение спектральных компонент речевого сигнала.

Частота связанного с податливостью стенок первого радиального резонанса $\omega_{\text{рад}}$ в артикуляторно-формантном синтезаторе, в *отличие от формантного синтезатора, не остается* фиксированной для всех звуков, а зависит от объема V речевого тракта или объема участка, находящегося перед смычкой:

$$\omega_{\text{рад}} = 1/\sqrt{L_{\text{ст}} C_{\text{рт}}}, \quad C_{\text{рт}} = V/(\rho_0 c_0^2),$$

$$V(t) = \int_0^{l_{\text{см}}} S(x, t)\, dx,$$

где $l_{\text{см}}$ — *расстояние от голосовой щели до смычки*, $L_{\text{ст}} \approx 0{,}01 - 0{,}015$ г/см^4. Это дает возможность определить поправки на резонансные частоты как

$$\omega_k^* = \sqrt{\omega_k^2 + \omega_{\text{рад}}^2},$$

где ω_k — частоты, вычисленные в предположении абсолютно

жестких стенок, как это делается в описанных выше методах. Во время звонкой смычки частота первого резонанса равна не нулю, а $\omega_{\text{рад}}$, и она различная для согласных с разным местом артикуляции вследствие различия объема речевого тракта перед смычкой.

Для оценки требуемой вычислительной мощности в артикуляторно-формантном синтезаторе существенное значение имеет количество обращений к процедуре нахождения собственных чисел в единицу времени. Это количество зависит от скорости изменения площади сечения S во времени, отношения скорости изменения резонансных частот ω_k к скорости изменения S и способа интерполяции частот ω_k между отсчетами. Ясно, что при $\partial S/\partial t = 0$ не нужно производить никаких вычислений, а достаточно лишь использовать последние значения ω_k и δ_k в момент наступления стационарного состояния. При переходных процессах во время образования смычки или щели скорость изменения, например, частоты первого резонанса сильно возрастает. Поэтому во время этих переходных процессов частота вызовов процедуры поиска собственных чисел должна быть максимальной. Если же в речевом тракте нет явно выраженного сужения, т. е. минимальная площадь в подвижной области тракта больше 0,5—1,0 $см^2$, то скорость изменения резонансных частот ω_k невелика, и здесь частота вызова процедуры поиска собственных чисел может быть небольшой. Таким образом, момент следующего обращения к вычислению собственных чисел и функций зависит от максимальной скорости изменения частот $d\omega_k/dt$ по всем вычисляемым резонансам.

Пусть, например, $\omega_k(t)$ и $d\omega_k/dt$ — значения k-й частоты и ее скорости в момент времени t. Ограничиваясь только первыми двумя членами разложения $\omega_k(t)$ в ряд Тейлора, экстраполируем значение частоты $\omega_k(t+\Delta t)$ через интервал времени Δt:

$$\omega_k(t+\Delta t) = \omega_k(t) + \Delta t \frac{d\omega_k}{dt}.$$

Если задать некоторый порог приращения частоты $\Delta_{\omega_k} = \omega_k(t+\Delta t) - \omega_k(t)$, то момент коррекции оценок частот найдем как

$$\Delta t = \frac{1}{\Delta_{\omega_k}} \frac{d\omega_k}{dt}.$$

К числу достоинств артикуляторно-формантного синтезатора относится возможность перераспределения вычислительной мощности во времени, заранее вычисляя и запоминая значения ω_k и ψ_k на тех интервалах, где скорость изменения резонансных частот мала. Поэтому средняя требуемая производительность процессора ниже максимальной. Экстраполяция значений ω_k и δ_k удобна еще и тем, что в процессе синтеза

синтезатором можно пользоваться той или иной интерполяционной формулой. Например, при линейной интерполяции

$$\omega_k(t)=\omega_k(t_i)+\frac{\omega_k(t_{i+1})-\omega_k(t_i)}{t_{i+1}-t_i}(t-t_i),$$

где t — текущее время, $t_i<t$ — момент предыдущего вычисления ω_k, а $t_{i+1}>t_i$ — еще не наступивший момент обращения к процедуре вычисления собственных чисел, которая, тем не менее, уже отработала на интервале медленного изменения ω_k и создала запас значений ω_k на некоторый интервал вперед.

Для реализации процедуры поиска собственных чисел методом Галеркина совершенно естественно использовать матричный процессор, и можно ожидать, что при обращении к этой процедуре примерно каждые 5 мс будет обеспечен реальный масштаб времени. Еще меньше вычислительных мощностей требует метод прогонки. Остальные блоки артикуляторно-формантного синтезатора могут быть реализованы вычислительными средствами с параллельной архитектурой, поскольку работа этих блоков относительно независима и они редко обмениваются данными. Таким образом, с технической точки зрения артикуляторно-формантный синтезатор вполне реализуем уже имеющимися средствами электронной технологии.

Артикуляторно-формантный синтезатор, однако, обладает и недостатками, главный из которых заключается в том, что все используемые в нем методы расчета акустических характеристик речевого тракта, строго говоря, справедливы только для неизменной формы речевого тракта. Поэтому обширный класс нестационарных явлений либо вообще остается нереализованным в этом синтезаторе, либо для их моделирования приходится прибегать к искусственным и довольно сложным приемам. В то же время известно, что система восприятия речи человеком во многом ориентируется на нестационарные процессы, причем для достижения требуемого качества речи требуется сохранение естественных связей между различными процессами. В качестве примера можно привести колебания стенок тракта, которые не только вносят потери и создают радиальный резонанс, но и развертываются во времени и вдоль речевого тракта. Другой пример относится к необходимости учета изменений частот и затуханий резонансных колебаний на интервале открытой голосовой щели. Методы расчета таких явлений также основываются на стационарном представлении, как бы давая закончиться всем переходным процессам для каждого значения площади голосовой щели, тогда как динамика этого процесса может оказать существенное влияние на восприятие синтетической речи.

Объем программного обеспечения артикуляторно-формантного синтезатора с использованием метода Галеркина

значительно превышает объем для артикуляторного синтезатора с непосредственным расчетом акустической волны в речевом тракте, но метод прогонки очень экономен.

§ 7.6. Устойчивость и точность решения

Конечной целью решения уравнения речевого тракта является синтез речевого сигнала, поэтому было бы естественно попытаться исключить промежуточные операции в виде поиска собственных частот и функций. В такой постановке задача синтеза речи является частным случаем задачи решения уравнений в частных производных. Для решения таких задач разработан ряд методов, и некоторые из них применимы и к синтезу речи. Прежде чем приступить к описанию этих методов, необходимо обсудить ряд проблем, возникающих при решении уравнений в частных производных, а именно — точности и устойчивости. Проблема объема вычислений и связанных с ним затрат процессорного времени является специфической для синтеза речи, где требуется обеспечить масштаб времени решения, близкий к реальному, для избежания недопустимых задержек в ответах синтезатора.

Наиболее распространенным методом решения уравнений в частных производных является метод сеток, в котором пространственная и временная координаты x и t разбиваются на множество отсчетов с шагами Δx и Δt (необязательно постоянными) и решение уравнения отыскивается в узлах этой сетки. Существует много способов задания этих узлов, но далеко не все они дают устойчивое решение, причем различают устойчивость по отношению к возбуждению, т. е. правой части уравнения, устойчивость к граничным условиям и погрешности округления в узлах сетки. Дополнительная трудность, не всегда учитываемая в работах по синтезу речи, состоит во влиянии погрешностей округления, связанных с конечной длиной машинного слова. При уменьшении шагов Δx и Δt, которое может потребоваться для обеспечения точности решения, необходимо увеличивать и число значащих разрядов. Если этого не сделать, то даже в устойчивой вычислительной схеме погрешности округления быстро накапливаются. Вообще, по мнению [51], в отличие от задач с обыкновенными дифференциальными уравнениями, конечно-разностные методы решения уравнений в частных производных обеспечивают «более, чем скромную точность».

Для устойчивого решения уравнений в частных производных нельзя брать произвольные Δx и Δt. Для волнового уравнения необходимым условием устойчивости является

$$c_0 \frac{\Delta t}{\Delta x} \leqslant 1,$$

где c_0 — скорость звука. Однако при $c_0 \Delta t = \Delta x$ происходит хотя и медленный, но неограниченный рост ошибок и после достаточно

большого числа шагов решение теряет смысл. К тому же скорость звука c_0 может быть переменной (например, в системах с податливыми стенками) и в любом случае c_0 известна только с конечным числом десятичных звуков. Поэтому в качестве практического условия устойчивости используют строгое неравенство $c_0\Delta t<\Delta x$.

Если ошибки в задании граничных условий, правой части и погрешности округления мало влияют на решение, то такое разностное уравнение называется хорошо обусловленным. Пусть мы имеем разностное уравнение

$$a_i u_{i-1}+b_i u_i+c_i u_{i+1}=f_i, \quad i=1, \ldots, N-1,$$

где коэффициенты a_i, b_i и c_i определяются по исходному дифференциальному уравнению. Тогда условие хорошей обусловленности задачи есть

$$\frac{|b_i|-|a_i+c_i|}{|b_i|+|a_i|+|c_i|}\geqslant\theta>0,$$

$$\max\{|a_i|,\ |b_i|,\ |c_i|\}\geqslant B>0,$$

где θ и B — некоторые постоянные, не зависящие от i и N [11]. Это условие, в частности, требует гладкости коэффициентов a_i, b_i и c_i, например, в виде

$$|a(x)-a(x')|\leqslant D|x-x'|^{\omega}.$$

Площадь поперечного сечения речевого тракта может претерпевать разрывы при отсчетах вдоль координаты x. Поэтому, если не предпринимать специальных приемов, то конечно-разностная задача для уравнения речевого тракта оказывается плохо обусловленной.

Примеры устойчивых и неустойчивых схем для волнового уравнения исследуются в [51]. Воспользовавшись результатами этих исследований, рассмотрим волновое уравнение относительно смещения y для однородной трубы:

$$\frac{\partial^2 y}{\partial t^2}=c_0^2\frac{\partial^2 y}{\partial x^2}.$$

Обозначив $v=\partial y/\partial t$ и $w=c_0\partial y/\partial x$, имеем

$$\frac{\partial v}{\partial t}=c_0\frac{\partial w}{\partial x},$$
$$\frac{\partial w}{\partial t}=c_0\frac{\partial v}{\partial x}.$$

Очевидная разностная схема для этой системы есть

$$\frac{v_j^{n+1}-v_j^n}{\Delta t}=c_0\frac{w_{j+1}^n-w_{j-1}^n}{2\Delta x},$$
$$\frac{w_j^{n+1}-w_j^n}{\Delta t}=c_0\frac{v_{j+1}^n-v_{j-1}^n}{2\Delta x},$$

где j — индекс по координате x, n — индекс по координате t.

Эта схема оказывается неустойчивой, если $\Delta t/\Delta x^2$ не остается ограниченным при $\Delta t,\ \Delta x \to 0$. С другой стороны, система

$$\frac{v_j^{n+1}-v_j^n}{\Delta t}=c_0\frac{w_{j+1/2}^n-w_{j-1/2}^n+w_{j+1/2}^{n+1}-w_{j-1/2}^{n+1}}{2\Delta x},$$

$$\frac{w_{j-1/2}^{n+1}-w_{j-1/2}^n}{\Delta t}=c_0\frac{v_j^{n+1}-v_{j-1}^{n+1}+v_j^n-v_{j-1}^n}{2\Delta x},$$

безусловно устойчива. Эта система называется неявной, потому что и в правой, и в левой частях присутствуют индексы $n+1$. Известно, что неявные схемы обеспечивают лучшую точность, чем явные, поскольку они используют интерполяцию назад, а явные — интерполяцию вперед, причем часто явные схемы оказываются неустойчивыми, как предыдущая схема для волнового уравнения. В гл. 2 мы уже встречались с явлением разной погрешности при разных способах вычисления производных в конечно-разностных схемах.

Если по какой-либо координате, например, по t шаг Δt устремить к нулю, то возникающий предельный случай метода сеток называется методом прямых [1]. В этом случае для уравнения речевого тракта относительно потенциала скорости

$$\frac{1}{c_0^2}\frac{\partial^2\Phi}{\partial t^2}=\frac{1}{S}\frac{\partial}{\partial x}\left(S\frac{\partial\Phi}{\partial x}\right)-Q(x)\Phi+F(x,\ t),$$

получаем систему обыкновенных уравнений

$$\frac{S_k}{c_0}\Phi_k''(t)=\frac{S_k[\Phi_{k-1}(t)-\Phi_k(t)]-S_{k-1}[\Phi_k(t)-\Phi_{k-1}(t)]}{\Delta x^2}-S_kQ_k\Phi_k(t)+$$

$$+F_k(t),\quad k=0,\ 1,\ \ldots,\ n+1;\ (n+1)\Delta x=l,$$

где $S_k=S(x_k)$, $\Phi_k(t)=\Phi(x_k,\ t)$, $Q_k=Q(x_k)$, $F_k(t)=F(x_k,\ t)$. Известно, что эта схема имеет сходящееся решение. В ней имеется $2n$ произвольных постоянных, для отыскания которых используются граничные условия, что дает систему из $2n$ алгебраических уравнений. Точность решения пропорциональна $(\Delta x)^2$, т. е. невысока. Однако метод прямых открывает возможность для повышения точности путем использования таких хорошо известных способов решения обыкновенных дифференциальных уравнений, как метод Рунге — Кутта и Адамса [1, 11].

Пусть имеется обыкновенное уравнение первого порядка $y'=f(x,\ y)$. Тогда в схеме Рунге — Кутта r-го порядка значение функции y в точке $n+1$ описывается как

$$y_{n+1}=y_n+\Delta x(p_1k_1+\ldots+p_rk_r), \tag{7.7}$$

где

$$k_1 = f(x_n, y_n),$$

$$k_2 = f(x_n + \alpha\Delta x,\ y_n + \alpha\Delta x k_1),$$

$$k_3 = f(x_n + \beta\Delta x,\ y_n + \beta\Delta x k_2),$$

. . . . . .

$$k_r = f(x_n + \gamma\Delta x,\ y_n + \gamma\Delta x k_{r-1}).$$

Коэффициенты α, β, ..., γ и p_1, p_2, ..., p_r путем сопоставления с членами разложения (7.7) в ряд Тейлора подбираются так, чтобы при заданном r получить аппроксимацию возможно более высокого порядка. Ошибка аппроксимации равна

$$\varepsilon_r = \frac{\Delta x^{r+1}\varphi^{r+1}(\xi)}{(r+1)!},$$

где $\varphi = y_{n+1} - y_n - \Delta x(p_1 k_1 + \ldots + p_r k_r)$.

Относительно схем Рунге—Кутта известно, что схема пятого порядка, $r=5$, имеет примерно такую же точность, что и схема четвертого порядка, $r=4$, из-за свойств φ, а схема шестого порядка весьма громоздка. Поэтому практически распространение получила схема четвертого порядка:

$$y_{n+1} = y_n + \frac{\Delta x^2}{6}(k_1 + 2k_2 + 2k_3 + k_4),$$

где

$$k_1 = f(x_n, y_n),$$

$$k_2 = f\left(x_n + \frac{\Delta x}{2},\ y_n + \frac{k_1\Delta x}{2}\right),$$

$$k_3 = f\left(x_n + \frac{\Delta x}{2},\ y_n + \frac{k_2\Delta x}{2}\right),$$

$$k_4 = f(x_n + \Delta x,\ y_n + k_3\Delta x).$$

Погрешность этой схемы

$$\varepsilon_4 = \frac{\Delta x^5}{120}\varphi^5(\xi).$$

Как видно, ни одно из вычисленных значений f на n-м шаге в этой схеме не может быть использовано на следующих шагах. Это обстоятельство относят к числу недостатков схемы

Рунге—Кутта, поскольку на каждом шаге f приходится вычислять r раз, и если эта функция сложна, то требуется много времени. В применении к решению уравнения речевого тракта, однако, это обстоятельство оказывается не столь существенным.

Вычислительная схема Адамса получается путем интерполяции $f(x, y)$ в n точках алгебраическим многочленом $L_{n,r}$ порядка r. Это приводит к выражению

$$y_{n+1}=y_n+\int_{x_{n-j}}^{x_{n+1}} L_{n,r}(x)\,dx=y_{n-j}+\Delta x\sum_{i=0}^{r}\beta_i f_{n-i},$$

где $\beta_i=\int_{-j}^{1} Q_j(z)\,dz$, а $Q_j(z)$ получается путем замены $x-x_n=z\Delta x$, т. е. является многочленом, не зависящим ни от Δx, ни от n. Также не зависят от Δx, n и f постоянные коэффициенты β_i. Как видно, в схеме Адамса для каждого шага требуется вычислять лишь одно новое значение f, поскольку остальные значения используются от предыдущего шага. В зависимости от вида полинома $L_{n,r}$ схема Адамса может быть либо экстраполяционной, либо интерполяционной. Как обычно, точность интерполяционных формул выше, чем точность экстраполяционных. Так, для $r=4$, погрешность интерполяционной формулы $\varepsilon_{4и}=0{,}0105\,\Delta x^5$, а экстраполяционной — $\varepsilon_{4э}=0{,}0523\,\Delta x^5$, т. е. интерполяционная формула имеет погрешность в 5 раз меньшую, чем экстраполяционная.

Схема Адамса, *однако, имеет тот недостаток, что для* аппроксимации r-го порядка в начале вычислений нужно знать $r-1$ значение функции y. Кроме того, в отличие от метода Рунге—Кутта в нем невозможно без существенного усложнения изменить шаг Δx, а такое изменение может потребоваться *для повышения как точности, так и скорости вычислений.*

Представим уравнение речевого тракта в виде системы

$$\begin{aligned}\frac{\partial P}{\partial x}+\rho_0\frac{\partial}{\partial t}\left(\frac{U}{S}\right)&=0,\\ \frac{\partial U}{\partial x}+\frac{1}{\rho_0 c_0^2}\frac{\partial}{\partial t}(SP)+\frac{\partial S}{\partial t}&=0,\end{aligned}\qquad(7.8)$$

где P — акустическое давление, U — объемная скорость, S — площадь поперечного сечения,

$$S(x, t)=S_0(x, t)+\delta S(x, t),$$

S_0 — функция, определяемая движением артикуляторных органов, δS — изменение площади сечения за счет податливости стенок, причем $\delta S=Ly$, где L — периметр сечения, y — линейное

смещение стенок, и

$$m\frac{\partial^2 y}{\partial t^2}+b\frac{\partial y}{\partial t}+ky=P, \tag{7.9}$$

где m, b, k рассчитаны на единицу площади стенок. Обозначим объемную скорость, создаваемую движением стенок, как

$$U_{\text{ст}}(t)=\Delta x L\frac{\partial y}{\partial t}.$$

Тогда (7.9) можно представить, как

$$P(t)=\frac{m}{L\Delta x}\frac{\partial U_{\text{ст}}}{\partial t}+\frac{b}{L\Delta x}U_{\text{ст}}+\frac{k}{L\Delta x}\int_{-\infty}^{t}U_{\text{ст}}\,d\tau.$$

Воспользуемся методом прямых и, взяв левые производные по пространственной координате, получим следующую систему:

$$\frac{dP_i}{dt}=\frac{\rho_0 c_0^2}{S_i}\left[U_i(t)-U_{i+1}(t)-U_{\text{ст}\,i}(t)\right],$$

$$\frac{dU_i}{dt}=\frac{S_i}{\rho_0}\left[P_{i-1}(t)-P_i(t)\right],$$

$$\frac{dU_{\text{ст}\,i}}{dt}=\frac{1}{L_{\text{ст}}}\left[P_i(t)-V_{\text{ст}\,i}(t)-R_{\text{ст}\,i}U_{\text{ст}\,i}(t)\right],$$

$$\frac{dV_{\text{ст}\,i}}{dt}=\frac{U_{\text{ст}\,i}(t)}{C_{\text{ст}\,i}},$$

где $L_{\text{ст}}=\dfrac{m}{L(x)\Delta x}$, $R_{\text{ст}}=\dfrac{b}{L(x)\Delta x}$, $C_{\text{ст}}=\dfrac{L(x)\Delta x}{k}$. Здесь предполагается, что площадь поперечного сечения S достаточно медленно изменяется во времени, так что ее можно считать постоянной. При решении этой системы с переменным шагом Δx оказывается, что метод Рунге—Кутта в 2—3 раза быстрее метода Адамса, и качество синтетической речи довольно хорошее. Правда, при реализации на мини-ЭВМ с производительностью $\approx 10^5$ операций в секунду на одну секунду синтезированной речи требуется более 5 часов процессорного времени [81].

Решение системы (7.8) без предположения о постоянстве площади поперечного сечения S во времени было осуществлено в [154]. Голосовая щель была также включена в систему (7.8) и предусматривалось разветвление речевого тракта при синтезе назализованных гласных. Кроме того, в первое уравнение был добавлен член с потерями на вязкое трение rU/S, где $r=8\pi\mu/S^2$ — капиллярное трение, μ — коэффициент вязкости. При формировании конечно-разностной системы сначала использовался метод прямых с дискретизацией по пространственной координате, а затем производные по времени также были представлены в конечно-разностной форме. Полученная система алгебраических уравнений разрешается относительно

объемной скорости и давления методами подстановок и исключения. Анализ ошибок этого алгоритма показывает, что при $\Delta x = 0{,}5$ см и частоте отсчетов $F_{\text{отс}} = 40$ кГц в полосе частот до 4 кГц искажения незаметны, а при $\Delta x = 1$ см и $F_{\text{отс}} = 20$ кГц наблюдаются искажения на третьей форманте, однако перцептивно оба варианта почти эквивалентны.

Конечно-разностные методы определения акустической волны в речевом тракте вычислительно даже более сложны, чем методы поиска собственных чисел, поскольку здесь добавляются расчеты изменений переменных во времени. Поэтому при вычислениях на обычных ЭВМ последовательного действия реальный масштаб времени вряд ли достижим. Дальнейшее сокращение времени вычислений возможно лишь при переходе от универсальных методов решения дифференциальных уравнений в частных производных к методам, учитывающим специфику процессов речеобразования. Такие методы будут рассмотрены в гл. 10.

ГЛАВА 8

ФОРМА И ПЛОЩАДЬ СЕЧЕНИЯ РЕЧЕВОГО ТРАКТА

§ 8.1. Системы координат

Длина и форма речевого тракта изменяются в результате движений артикуляторных органов: гортани, языка, губ, нижней челюсти и небной занавески. Эти движения совершаются под воздействием сокращений мышц, причем вследствие кинематических связей движения артикуляторных органов оказываются зависимыми между собой. Так, поворот или горизонтальное смещение нижней челюсти меняют положение губ и языка, смещения корня языка приводят к смещению гортани и к изменению ширины надгортанной области. Кроме того, вертикальные смещения корня языка меняют высоту небной занавески, поскольку между ними имеется связь через мышцу, опускающую небную занавеску, которая в виде петли проходит через ткани небной занавески и крепится к подъязычной косточке.

Основные движения происходят в среднесагиттальной плоскости, т. е. плоскости симметрии лица, проходящей через позвоночник. Исследования движений языка с помощью палатограмм (степени контакта на искусственном нёбе) показывают, что они симметричны относительно этой плоскости [116]. Движения губ также симметричны относительно среднесагиттальной плоскости. Поэтому различные системы координат, в которых описываются движения артикуляторных органов, располагаются только в среднесагиттальной плоскости. Выбор системы координат зависит как от кинематики артикуляторных органов, так и от способа вычисления длины и площади речевого тракта. Основными координатными системами являются одна полярная система (для языка) и две прямоугольные — неподвижная (для всего тракта) и подвижная (для нижней челюсти). Эти системы показаны на рис. 8.1.

Неподвижная система координат XOY предназначена для вычисления длины и площади поперечного сечения речевого тракта. В этой системе ось Y совпадает со слегка идеализированной линией задней вертикальной поверхности речевого тракта на участке от фарингиальной области до небной

занавески. Горизонтальная ось X проходит через верхнюю поверхность голосовых складок, когда гортань находится в самом нижнем положении.

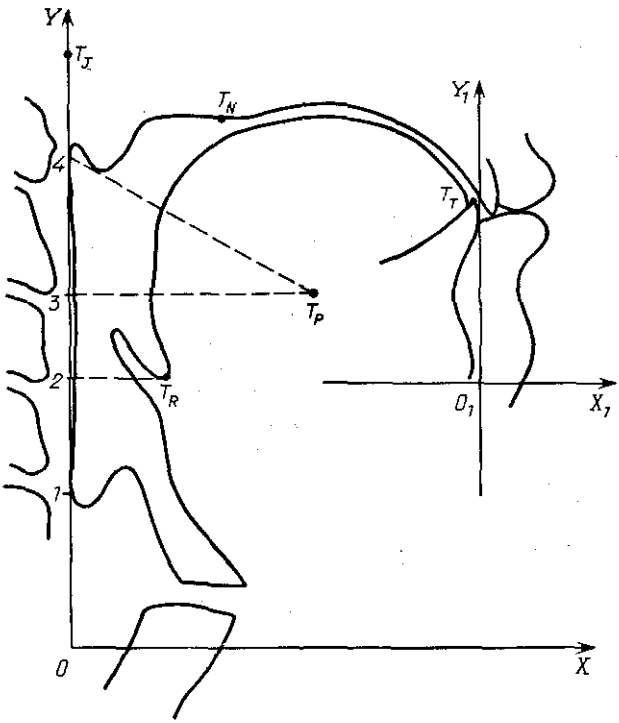


Рис. 8.1. Система координат речевого тракта

Подвижная система координат $X_1O_1Y_1$ связана с нижней челюстью. Вертикальная ось Y_1 проходит через наружную поверхность нижнего зуба, а горизонтальная ось X_1 проходит через точку T_R корня языка, находящегося в верхнем нейтральном положении. *При верхнем нейтральном положении* рот закрыт, зубы сомкнуты и ось X_1 параллельна оси X, а ось Y_1 — параллельна оси Y. Система координат $X_1O_1Y_1$ поворачивается относительно точки вращения нижней челюсти T_J. Точка T_J в нейтральном положении находится на оси несколько выше небной занавески. Эта точка может смещаться в горизонтальном направлении вперед, в результате чего вся система координат $X_1O_1Y_1$ также может смещаться относительно системы XOY.

Форма языка описывается в полярной системе координат с центром в точке T_P, которая находится на прямой линии, соединяющей точку T_R корня языка с вершиной нижнего зуба. При повороте и смещениях нижней челюсти изменяется и положение точки T_P, поэтому ее координаты задаются

в подвижной системе $X_1O_1Y_1$. Для нейтрального состояния форма поверхности языка очень близка к дуге полуокружности, поэтому точка T_P располагается посередине отрезка, соединяющего точку T_R и вершину нижнего зуба.

Движения небной занавески сопровождаются существенными изменениями ее формы, но приблизительно можно считать, что она просто поворачивается относительно точки T_N, координаты которой задаются в системе XOY.

Форма речевого тракта в плоскости XOY описывается в соответствии с измерениями, выполненными на кинорентгенограммах. При этом форма языка определяется как суперпозиция собственных функций с переменными во времени коэффициентами. Участок от голосовой щели до корня языка и участок небной занавески также подвергаются изменениям во времени. Остальные поверхности не меняют свою форму.

Фарингиальная область, а также форма мягкого и твердого нёба вместе с верхними зубами и верхней губой измеряются в системе координат XOY, а подъязычная поверхность, открывающаяся при сдвиге языка назад — в подвижной системе $X_1O_1Y_1$. Форма нижней губы также описывается в системе $X_1O_1Y_1$.

§ 8.2. Кинематика речевого тракта

Форма речевого тракта в фарингиальной области — участка, находящегося непосредственно над голосовой щелью, в нейтральном положении измеряется на рентгенограммах. В связи с тем, что хрящи гортани чувствительны к рентгеновским лучам, обычно эта область не захватывается при съемке. Имеется ограниченное число съемок в статике для гласных звуков, доступных для измерения. Для описания этой области тракта используем рентгеновские снимки, сделанные для гласных в лаборатории передачи речи Королевского технологического института в Стокгольме. Об изменении формы этой области и высоте гортани приходится судить по косвенным данным, в частности, по высоте корня языка, подъязычной косточки и входа в пищевод (в тех случаях, когда он виден на кинорентгенограммах).

Положение гортани и корня языка может быть изменено с помощью разных групп мышц, поэтому их смещения относительно зависимы, причем при опускании гортани и подъеме корня языка эта зависимость больше, чем при подъеме гортани и опускании корня языка. Упрощая, можно положить, что любые вертикальные смещения корня языка полностью передаются на положение гортани:

$$y'_{\text{гщ}} = y_{\text{гщ}} + \Delta y_{\text{кор}},$$

где $y'_{\text{гщ}}$ — новое значение вертикальной координаты голосовой щели, $y_{\text{гщ}}$ — ее старое значение, $\Delta y_{\text{кор}}$ — изменение вертикальной координаты корня языка. Такое смещение испытывают и все

точки фаринкса: задняя поверхность до точки 1 (на рис. 8.1). и передняя поверхность — до уровня точки 2. Величина смещения гортани может достигать 30 мм и более. В [103] с помощью стереоэндоскопа было установлено, что при переходе от нейтрального состояния к фонации гортань поднимается на 10 мм. По данным [152], разница в высоте гортани может быть до 30 мм.

При самостоятельном вертикальном смещении гортани будем считать, что его влияние распространяется до корня языка линейно убывающему закону:

$$y' = y + \delta y,$$

где

$$\delta y = \delta y_{\text{гщ}} \frac{y_{\text{гщ max}} - y}{y_{\text{гщ max}} - y_{\text{гщ min}}},$$

$\delta y_{\text{гщ}}$ — вертикальное смещение гортани, δy — вертикальное смещение точек фаринкса, $y_{\text{гщ max}}$ и $y_{\text{гщ min}}$ — наименьшая и наибольшая вертикальные координаты фаринкса. Видно, что для последней, самой высокой точки фаринкса, совпадающей с корнем языка, вертикальное смещение становится равным нулю. Вследствие горизонтальных смещений корня языка изменяются и горизонтальные координаты поверхностей фаринкса. При этом будем считать, что сама гортань не сдвигается в горизонтальном направлении, т. е. что влияние горизонтальных смещений корня языка убывает по линейному закону:

$$x' = x + (y - y_{\text{гщ}}) \frac{x_{\text{кор}} - x_{\text{гщ}}}{y_{\text{кор}} - y_{\text{гщ}}},$$

где x — старая координата поверхности фаринкса, x' — новая координата, $x_{\text{гщ}}$, $y_{\text{гщ}}$ — координаты голосовой щели до смещения, $x_{\text{кор}}$, $y_{\text{кор}}$ — координаты корня языка.

Поскольку измерения поверхности фаринкса на рентгенограмме выполняются лишь один раз для нейтрального состояния речевого тракта, то могут возникать разрывы в координатах корня языка и верхней точки фаринкса. Для того чтобы избежать разрывов, сначала вычисляется фактическая разница в этих координатах, а затем определяются смещения всех точек фарингиальной поверхности.

Такие преобразования координат точек выполняются и для передней и для задней поверхности фаринкса в плоскости XOY. Координаты вертикальной линии, изображающей заднюю поверхность речевого тракта выше точки 1, не нуждаются в преобразованиях, поскольку эта линия не смещается в горизонтальном направлении (совпадает с осью координат Y), и задается не своими отсчетами, а уравнением $x = \text{const}$.

Форма твердого нёба и верхних зубов также измеряется на рентгенограммах. Она неизменная для всех звуков и не подвергается никаким преобразованиям.

Форма мягкого нёба, измеренная для неносовых звуков (при поднятой небной занавеске), также может считаться постоянной во всех случаях, кроме артикуляции звуков /М, Н/. В последнем случае происходит поворот участка мягкого нёба относительно некоторой точки с координатами (x_N, y_N).

Описание формы языка с помощью небольшого числа параметров в целях синтеза речи было предложено в [98, 190]. В [145] форма языка описывается через ряд Фурье. Путем факторного

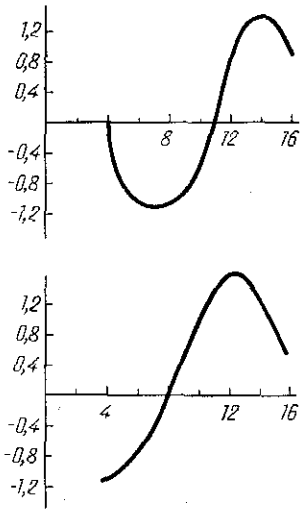


Рис. 8.2. Две главные компоненты формы языка по [140]

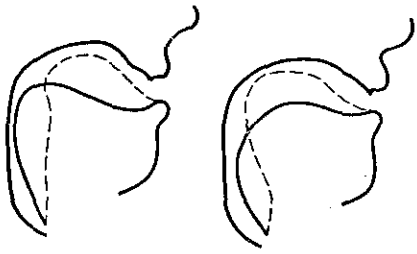


Рис. 8.3. Формы языка при максимальном значении каждой из главных компонент (по [140])

анализа рентгенограмм в [117] были найдены две компоненты, *отвечающие за сдвиг передней части языка вверх со смещением* корня вперед и подъем языка с одновременным сдвигом назад (рис. 8.2 и 8.3). Статистический анализ форм языка на рентгенограммах, выполненный в [189], привел к определению четырех компонент (рис. 8.4). Все эти системы, однако, описывают форму языка в статике, не сообщая ничего о траекториях перехода от одной формы к другой. Расчет изменения формы языка в трех измерениях под воздействием сокращения мышц дает возможность описания динамики [129], однако требует большого времени для вычислений. Для артикуляторного синтезатора представляется более предпочтительным способ, объединяющий физическую адекватность описания с экспериментально найденными параметрами модели. Этот способ состоит в анализе упругих деформаций языка, представленного как изогнутая пластинка. Из теоретического анализа получаем форму собственных функций $\psi_k(\varphi)$, а аппроксимация формы языка Ф, измеренной на рентгенограммах, дает значения коэффициентов c_k при собственных функциях:

$$\Phi(\varphi, t) = \sum_k c_k \psi_k(\varphi) f_k(t) + R_0, \qquad (8.1)$$

где φ — угол в полярной системе координат языка, t — время,

R_0 — радиус поверхности языка в нейтральном положении, f_k — затухающие гармонические колебания для каждой моды, являющиеся решением уравнений

$$f_k'' + 2gf_k' + \omega_k^2 f_k = F_k(t).$$

В [59] экспериментально показано, что 5 собственных функций обеспечивают достаточно малую ошибку (6—7%) в равномерной метрике, т. е. при оценке максимальной

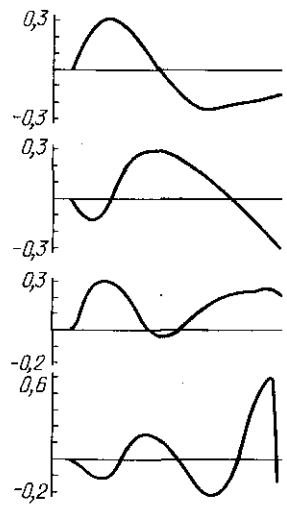


Рис. 8.4. Четыре главные компоненты формы языка (по [190])

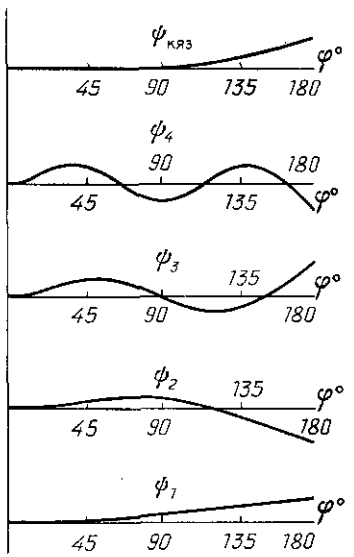


Рис. 8.5. Пять собственных функций языка (по [59])

разности между истинной и вычисленной формами языка. Вид собственных функций показан на рис. 8.5. Четыре из этих собственных функций описывают форму языка в целом и одна — форму кончика языка. Этот способ удобен тем, что одновременно с формой дает и динамику языка. Кроме того, сами собственные функции рассчитываются лишь один раз — для выбранных размеров языка, сохраняющихся постоянными в процессе артикуляции, так что при синтезе формы языка нужно лишь вызвать из памяти собственные функции. Число отсчетов этих функций, равное 50, вполне удовлетворяет требованиям гладкости описания формы языка.

Как уже говорилось во Введении, собственные функции тела языка вычисляются как

$$\psi_k(\varphi) = \operatorname{sh} p_k\varphi + \frac{p_k^2}{q_k^2}\,\frac{\operatorname{sh} p_k\varphi + p_k \operatorname{ch} p_k\varphi}{\sin \pi q_k + q_k \cos \pi q_k} \sin q_k\varphi,$$

а для кончика языка — как

$$\psi_5(x) = a_5(\operatorname{ch} p_5 x - \cos p_5 x) + b_5(\operatorname{sh} p_5 x - \sin p_5 x).$$

Аргументами для первых четырех функций является угол в полярной системе координат языка, а для пятой собственной функции — расстояние от середины языка до кончика.

Корни характеристических уравнений для собственных функций языка p_k определяются при условии, что в нейтральном положении язык представляет собой дугу полуокружности, т. е. $\varphi_{max} = \pi$. Их значения приведены во Введении. Заметим, что корни p_k не зависят от размеров языка. Собственные частоты находим как

$$\omega_k = \sqrt{\frac{EJ_z p_k^2 q_k^2}{\rho_т R_0^2} + \frac{c}{\rho_т}},$$

где $\rho_т$ — погонная плотность тканей языка, $\rho_т = \rho_{т0} S$, $\rho_{т0} \approx \approx 1{,}05$ г/см^3, S — площадь поперечного сечения языка, $S = 3$ см^2; c — погонная упругость тканей, находящихся под языком, $c = 25$ г/(см · с^2); J_z — момент инерции языка относительно оси Z, перпендикулярной плоскости XOY, $J_z = 0{,}4$ см^4, E — модуль упругости языка, $E = 2 \cdot 10^3$ Па; коэффициент вязких потерь $g = 5 - 10$ с$^{-1}$. Для кончика языка $p_5 = 2g_5/l$, где l — длина языка вдоль криволинейной оси, $q_5 = 0{,}597\pi$, а коэффициенты $a_5 = 0{,}707\sqrt{4/l}$, $b_5 = -0{,}518\sqrt{4/l}$. Собственная частота $\omega_5 = 2p_k^2\sqrt{J_z E/\rho_т}$.

Координаты центра полярной системы языка в системе нижней челюсти $X_1O_1Y_1$ находятся как

$$x_{1p} = x_{1\,кор} + R_0 \cos\theta_0 = 0{,}5(x_{1\,кор} + x_{1\,нз}),$$
$$y_{1p} = y_{1\,кор} + R_0 \sin\theta_0 = 0{,}5(y_{1\,кор} + y_{1\,нз}),$$

где $x_{1\,кор}$, $y_{1\,кор}$ — координаты корня языка, а θ_0 угол наклона базовой линии языка

$$\theta_0 = \operatorname{arctg}\frac{y_{1\,нз} - y_{1\,кор}}{x_{1\,нз} - x_{1\,кор}},$$

где $x_{1\,нз}$, $y_{1\,нз}$ — координаты точки на нижних зубах, в которой находится кончик языка в нейтральном состоянии. Координаты $x_{1\,нз}$, $y_{1\,нз}$ остаются неизменными при любых артикуляциях, а координаты корня $x_{1\,кор}$, $y_{1\,кор}$ изменяются в процессе артикуляции. В неподвижной системе координат XOY координаты центра полярной системы x_p, y_p языка зависят от угла поворота нижней челюсти и смещения точки поворота нижней челюсти в горизонтальном направлении:

$$x_p = x_{1\,p}\cos\alpha + (y_{1\,p} - y_J)\sin\alpha + x_J,$$
$$y_p = (y_{1\,p} - y_J)\cos\alpha - x_{1\,p}\sin\alpha + y_J,$$

где α — угол поворота нижней челюсти относительно состояния

с сомкнутыми зубами, x_J — горизонтальная координата точки вращения, y_J — вертикальная координата точки вращения. Горизонтальные движения нижней челюсти наблюдаются на кинорентгенограммах при артикуляции переднеязычных звуков /С, Ш, Т, Д/ [59]. Несмотря на то, что величина смещения x_J сравнительно невелика — 2—3 мм, она играет важную роль, особенно при быстром темпе артикуляции. Координаты точек поверхности языка в системе XOY находится как

$$x = x_p - \Phi(\varphi, t)\cos\gamma,$$
$$y = y_p + \Phi(\varphi, t)\sin\gamma,$$

где $\Phi(\varphi, t)$ определяется по (8.1) как деформация полуокружности с радиусом R_0, а $\gamma = \gamma^* - \theta$ — угол в полярной системе координат, где $\theta = \theta_0 - \alpha$ — изменение угла наклона базовой линии языка из-за поворота нижней челюсти. Угол γ^* изменяется от нуля до $\gamma^*_{\max}$, где $\gamma^*_{\max} = \pi - \operatorname{arctg}\frac{y_{1\,\text{кон}} - y_{1\,\text{кор}}}{x_{1\,\text{кон}} - x_{1\,\text{кор}}} + \theta_0$ — угол в полярной системе координат языка для данного положения кончика языка $x_{1\,\text{кон}}$, $y_{1\,\text{кон}}$ в подвижной системе $X_1O_1Y_1$. В случае, когда кончик языка находится на нижних зубах (в нейтральном положении), $\gamma^*_{\max} = \pi$. Поскольку число отсчетов N для собственных функций фиксировано, то при изменении максимального угла $\gamma^*_{\max}$ изменяются угловые приращения между этими отсчетами:

$$\Delta\gamma^* = \gamma^*_{\max}/N.$$

От нижних зубов по направлению к позвоночнику под язык уходит связанная с нижней челюстью поверхность $\Phi_{\text{пп}}$, которая обнажается при сдвиге языка назад, как, например, во время артикуляции гласного звука /Ы/. Эта поверхность в подвижной системе $X_1O_1Y_1$ с координатами $O_1(x_{01}, y_{01})$ имеет текущие координаты $(x_{1\text{пп}}, y_{1\text{пп}})$, а в неподвижной системе — координаты $(x_{\text{пп}}, y_{\text{пп}})$:

$$x_{\text{пп}} = (x_{1\text{пп}} + x_{01})\cos\alpha + (y_{1\text{пп}} + y_{01} - y_J)\sin\alpha + x_J,$$
$$y_{\text{пп}} = (y_{1\text{пп}} + y_{01} - y_J)\cos\alpha - (x_{1\text{пп}} + x_{01})\sin\alpha + y_J.$$

Форма губ определяется тремя параметрами: длиной, прогибом и выпячиванием. В вертикальном направлении смещается, в основном, нижняя губа, тогда как выпячивание при артикуляции гласных (Y, O) осуществляется обеими губами. Форма нижней губы в плоскости XOY измеряется на рентгенограммах, причем в качестве дополнительных параметров выступают координаты $(x_{1\text{нз}}, y_{1\text{нз}})$ точки соприкосновения с нижними зубами, отсчитываемые в подвижной системе $X_1O_1Y_1$. При повороте и сдвиге нижней челюсти координаты нижней губы в системе XOY есть

$$x_{\text{г}} = (x_{1\text{г}} + x_{01})\cos\alpha + (y_{1\text{г}} + y_{01} - \Delta y_{\text{г}} - y_J)\sin\alpha + x_J,$$
$$y_{\text{г}} = (y_{1\text{г}} + y_{01} - \Delta y_{\text{г}} - y_J)\cos\alpha - (x_{1\text{г}} + x_{01})\sin\alpha + y_J,$$

где $y_{\text{г}}$ — прогиб нижней губы в средней точке.

В [59] было показано, что собственные функции упругих деформаций губ могут быть представлены как $\psi_{\text{г}}(z)=\sin p_{\text{г}k} z$, где $p_{\text{г}k}=k\pi/l_{\text{г}}$, $l_{\text{г}}$ — длина губы, причем достаточно хорошее описание формы и длины и динамики губ достигается при использовании лишь одной — первой собственной функции. Поэтому вертикальное смещение губы равно $\Delta y_{\text{г}}=c_{1\text{г}}(t)\sin\frac{\pi}{2}$, целевому значению для данного звука. При выпячивании длина губ изменяется на $\Delta l_{\text{г}}$ и точка с максимальной высотой на поверхности губы смещается по закону $c_{2\text{г}}\sin\frac{\pi z}{l_{\text{г}}\Delta l_{\text{г}}}$. Рассматривая выпячивание в плоскости XOZ (рис. 8.6), приблизительно оценим величину выпячивания $\Delta x_{\text{г}}=bd$ из условия сохранения длины губы (несжимаемости):

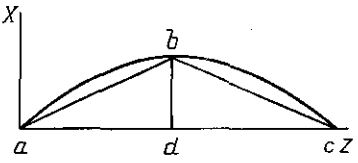


Рис. 8.6. Форма губы в горизонтальной плоскости XOZ

$$\Delta x_{\text{г}}=\sqrt{\Delta l_{\text{г}}\left(2l_{\text{г}}-\Delta l_{\text{г}}\right)/2}.$$

При выпячивании образуется пространство между губами и зубами, которое можно заполнить различными способами. Простейший из них — это поверхность с постоянным значением вертикальной координаты, равной координате точки прикосновения губы и зубов.

Итак, имеется 16 управляемых параметров и координат в речевом тракте: высота голосовой щели $\delta y_{\text{гщ}}$, координаты корня языка ($x_{1\text{кор}}$, $y_{1\text{кор}}$) в подвижной системе $X_1O_1Z_1$, координаты кончика языка ($x_{1\text{кон}}$, $y_{1\text{кон}}$), угол поворота нижней челюсти α_J и горизонтальное смещение точки ее вращения x_J, угол поворота нёбной занавески α_N, длина губ $l_{\text{г}}$, изменение длины $\Delta l_{\text{г}}$ и величины прогиба нижней губы $\Delta y_{\text{г}}$, а также пять коэффициентов при собственных функциях языка. Как будет показано ниже, к этим параметрам добавляются еще два управляемых параметра длины голосовых складок и расстояние между передними концами складок.

§ 8.3. *Средняя линия и длина речевого тракта*

Определим длину речевого тракта как расстояние от голосовой щели до губ, измеренное вдоль криволинейной средней линии, а среднюю линию определим как геометрическое место точек в плоскости XOY таких, что они находятся посередине кратчайшей прямой линии, соединяющей подвижные и неподвижные участки речевого тракта. При этом оказывается, что длина тракта зависит от его формы вследствие изгиба. Для иллюстрации этого свойства рассмотрим цилиндрическую

трубу, изогнутую по дуге окружности (рис. 8.7). Радиус изгиба средней линии $R_c=(R_н+R_в)/2$, а длина средней линии $L_c=\int_0^{\theta} R_c d\varphi$, где φ — угол в полярной системе координат. Положив $\theta=\pi/2$, получим $L_c=\pi(R_н+R_в)/4$, т. е. при постоянном наруж-

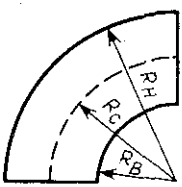


Рис. 8.7. Средняя линия изогнутой трубы

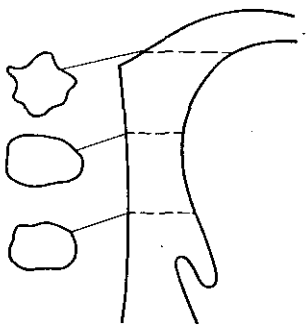


Рис. 8.8. Формы поперечного сечения речевого тракта (по [130])

ном радиусе $R_н$ длина средней линии пропорциональна величине внутреннего радиуса $R_в$ — при его увеличении увеличивается и L_c. Поэтому подъем или опускание языка (при постоянной высоте гортани) также приводят к увеличению или уменьшению длины речевого тракта. Координаты средней линии в фарингиальной области находятся как полусумма координат по оси абсцисс задней и передней поверхностей $x_m=(x_1+x_2)/2$ при заданной вертикальной координате y, отсчеты которой одинаковы у обеих поверхностей. В области входа в пищевод задняя поверхность имеет прогиб (см. рис. 8.1), поэтому возникает неопределенность в установлении положения средней линии и вычислении площади поперечного сечения. Если не учитывать дополнительную площадь под входом в пищевод, то способ расчета координат средней линии остается тем же, что и раньше. С учетом этой площади средняя линия смещается назад.

От точки 1 до точки 2 (корня языка) (см. рис. 8.1) горизонтальная координата задней поверхности равна нулю вследствие выбора системы координат XOY. Поэтому горизонтальная координата средней линии $x_m=x_2/2$. От точки 2 до точки 3, вертикальная координата которой равна вертикальной координате центра полярной системы языка y_p, координаты средней линии находятся как

$$x_m=x_t/2, \quad y_m=y_t,$$

где (x_t, y_t) — координаты поверхности языка в системе XOY.

Здесь также возникает проблема, связанная с наличием надгортанника — положение средней линии и площади сечения тракта зависят от того, учитывается или нет область, занятая надгортанником. Одно из решений состоит в том, чтобы положение средней линии рассчитывать без учета надгортанника, а площадь — с учетом.

Вплоть до высоты y_p нет проблем с определением координат средней линии, так как задняя поверхность задана уравнением $x_1 = 0$, а вертикальные координаты средней линии просто равняются координатам поверхности языка y_t. Однако выше точки 3 дискретное представление поверхности языка требует иного способа, поскольку кратчайшее расстояние между языком и задней поверхностью уже не лежит на горизонтальных линиях, а зависит от угла φ полярной системы языка. Расстояние от данной точки языка с координатами x_t, y_t до задней поверхности вычисляется как

$$R_в = \sqrt{x_t^2 + y_в y_t}, \tag{8.2}$$

где $y_в = y_p + x_p \operatorname{tg} \varphi$. Координаты средней линии есть

$$x_m = x_p - (\Phi + 0{,}5R_п)\cos\varphi, \quad y_m = y_p + (\Phi + 0{,}5R_в)\sin\varphi,$$

где Φ определено по (8.1).

Начиная от точки 4, расчеты осложняются тем, что мягкое нёбо и неподвижная поверхность речевого тракта заданы не непрерывно, а дискретными отсчетами. Поэтому кратчайшее расстояние в направлении вектора, заданного углом α в полярной системе языка до противоположной поверхности приходится искать через координаты пересечения двух прямых линий упомянутого вектора и линии, соединяющей два отсчета координат небной занавески и твердого нёба:

$$x_{пер} = \frac{B_1C_2 - B_2C_1}{B_2C_1 - A_2B_1}, \quad y_{пер} = \frac{C_1A_2 - C_2A_1}{B_2A_1 - A_2B_1}, \tag{8.3}$$

где

$$A_2 = y_p - y_t, \quad B_2 = x - x_p, \quad C_2 = -x_tA_1 - y_tB_1,$$

$$A_1 = x_{вi} - x_{вi+1}, \quad B_1 = y_{вi+1} - y_{вi}, \quad C_1 = -x_{вi}A_1 - y_{вi}B_1,$$

$(x_{вi}, y_{вi})$, $(x_{вi+1}, y_{вi+1})$ — координаты соседних точек в отсчетах мягкого и твердого нёба. Если горизонтальная координата пересечения $x_{пер} > x_{вi+1}$, то берется следующий отсчет поверхности нёба с большей горизонтальной координатой. Этот процесс повторяется до тех пор, пока решение не будет удовлетворять условиям $x_{вi} \leqslant x_{пер} \leqslant x_{вi+1}$.

Расстояние от центра полярной системы языка до нёба есть

$$R_в = \sqrt{(x_{пер} - x_p)^2 + (y_{пер} - y_p)^2},$$

а до поверхности языка

$$R_t=\sqrt{(x_t-x_p)^2+(y_t-y_p)^2}.$$

Отсюда координаты средней линии найдем как

$$x_m=x_p-[\Phi+0{,}5(R_в-R_t)]\cos\varphi,$$
$$y_m=y_p+[\Phi+0{,}5(R_в-R_t)]\sin\varphi.$$

Если произошло соприкосновение языка и нёба, то координаты средней линии языка совпадают с координатами языка на участке соприкосновения.

Координаты средней линии вычисляются подобным образом до тех пор, пока не закончится поверхность языка, т. е. радиус-вектор не дойдет до кончика языка, либо радиус-вектор пересечется с наружной поверхностью верхних зубов или верхней губы. В первом случае необходимо проверить, не обнажилась ли подъязычная поверхность в результате сдвига языка назад. Если это произошло, то точка пересечения отыскивается по (8.3), но параметры A_2, B_2 и C_2 определяются по отсчетам $(x_{ппi}, y_{ппi})$ и $(x_{ппi+1}, y_{ппi+1})$ подъязычной поверхности:

$$A_2=y_{ппi+1}-y_{ппi},\quad B_2=x_{ппi}-x_{ппi+1},\quad C_2=A_2x_{ппi}-B_2y_{ппi},$$

а параметры другой линии — как

$$A_1=1,\quad B_1=-\operatorname{tg}(\alpha_J+\theta_0)/2,\quad C_1=-x_{вi}B_1-y_{вi},$$

где $x_{вi}$, $y_{вi}$ — отсчеты поверхности твердого нёба.

Изменение направления линии, определяемой теперь углом $(\alpha+\theta_0)/2$, связано с необходимостью сгладить переход от полярной системы к системе нижней челюсти, задающей направление линий кратчайшего расстояния между губами как $\pi-\alpha_J$, т. е. почти вертикальными линиями.

На губном участке

$$A_1=1,\quad B_1=-\operatorname{tg}\alpha_J,\quad C_1=-B_1x_{вi}-y_{вi}.$$

Координаты средней линии на участке твердого нёба и губ есть

$$x_m=(x_{вi}+x_{пер})/2,$$
$$y_m=(y_{вi}+y_{пер})/2.$$

Расстояние между отсчетами средней линии находим как

$$l_{mi}=\sqrt{(x_{mi}-x_{mi+1})^2+(y_{mi}-y_{mi+1})^2},$$

а длину речевого тракта — как сумму расстояний между отсчетами

$$l=\sum_{i=1}^{N}l_{mi}.$$

§ 8.4. Площадь поперечного сечения речевого тракта

Площадь поперечного сечения голосовой щели имеет две компоненты: постоянную и переменную. Переменная компонента возникает при автоколебаниях голосовых складок, а постоянная зависит от расстояния $d_{\text{гщ}}$ между задними концами складок (рис. 8.9). Если это расстояние больше некоторой величины, то автоколебания не поддерживаются. В дыхательном положении складки далеко разведены, в положении фонации они полностью сведены. Во время артикуляции глухих взрывных и фрикативных звуков складки расходятся на расстояние, достаточное для срыва автоколебаний и голосовая щель имеет сопротивление потоку воздуха, сравнимое с сопротивлением речевого тракта. Если взрывной глухой звук создается с аспиративным участком, то складки сходятся на такое расстояние, при котором автоколебания еще не возникают, а турбулентные шумы имеют достаточную интенсивность. Так, на рисунке 8.10 показаны с интервалами в 33 мс фазы сближения и расхождения голосовых складок во фронтальной плоскости при артикуляции придыхательного глухого взрыва /$P_{\text{г}}^{h}$/. На рис. 8.11 показано изменение площади сечения голосовых щелей во время артикуляции глухих взрывных /П, Т, К/ и аффрикаты /Ц/ [184]. Таким образом, постоянная компонента площади голосовой щели заметно изменяется в процессах артикуляции и ее величина играет важную роль в создании акустических характеристик звуков речи.

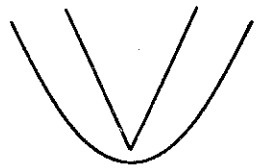


Рис. 8.9. Форма голосовой щели в состоянии дыхания

Поскольку голосовая щель часто имеет почти строго треугольную форму, то ее площадь легко рассчитывается по расстоянию между задними концами голосовых складок:

$$S_{\text{гщ}} = d_{\text{гщ}} l_{\text{гщ}} / 2,$$

где $l_{\text{гщ}}$ — длина голосовых складок. Определение же площади поперечного сечения почти всех остальных участков речевого тракта сталкивается с известными трудностями. Обычно стараются использовать какую-либо достаточно простую зависимость между расстоянием R от одной до другой поверхности речевого тракта в плоскости XOY и площадью поперечного сечения, например, в виде $S = aR^{b}$. В [146] эти коэффициенты различны для нижнего и верхнего фаринкса и ротовой области: ($a = 1{,}1$; $b = 2{,}21$); ($a = 0{,}67$; $b = 1{,}9$); ($a = 2{,}2$; $b = 1{,}38$). В [189] эти коэффициенты выбираются постоянными величинами, кроме губной области: $a = 1{,}8$; $b = 1{,}6$. Ясно, однако, что эта формула довольно грубо описывает реальное многообразие формы поперечного сечения речевого тракта.

Как видно из рис. 8.10, непосредственно над голосовой щелью находится расширение, называемое морганиевым желудочком. Его ширина мало меняется при расхождении голосовых складок, а высота меняется несколько больше.

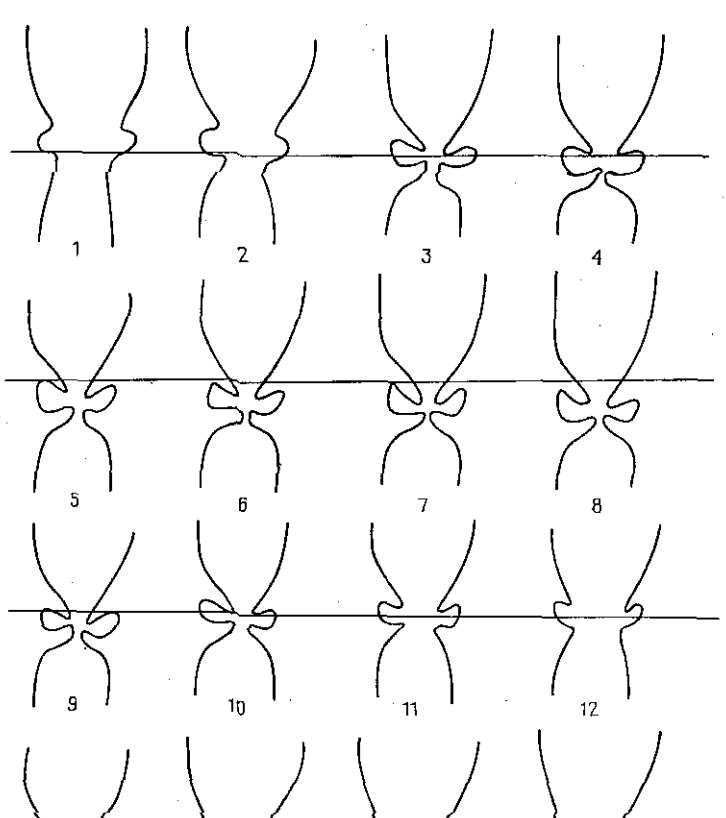


Рис. 8.10. Фазы состояния голосовой щели при артикуляции придыхательного взрывного $/P_{\text{в}}^{\text{h}}/$. Кадры следуют через 33 мс

В этой области форма поперечного сечения представляет собой деформированный круг, и, соответственно, площадь может быть вычислена либо как площадь круга $S=\pi R_1^2/4$, либо как площадь эллипса $S=\pi R_1 R_2/4$, где R_1 — расстояние между стенками фаринкса в плоскости XOY, R_2 — расстояние в плоскости ZOY. Последнее выражение более предпочтительно и может использоваться и выше морганиевого желудочка — до точки 1 на рис. 8.1. Это связано с необходимостью учета

независимого управления площадью фаринкса в плоскостях XOY и ZOY.

Близкая к эллиптической форме поперечного сечения фарингиальной области вплоть до нёбной зависимости получена томографическими исследованиями [130] (рис. 8.8). Непосредственные измерения, выполненные в [64], дают несколько иную форму. Прогиб языка по средней линии (желобок), создающий вогнутую форму сечения, иногда сменяется плоской или даже выпуклой линией — когда язык упирается в верхние зубы или твердое нёбо. Поэтому представляется целесообразным описать поперечное сечение речевого тракта в виде полуокружности, находящейся на прямоугольном основании. Криволинейная часть представляет собой либо заднюю стенку фаринкса, либо свод нёба, а горизонтальная линия — поверхность языка (рис. 8.12). Если язык отодвинут на расстояние $R > r_a$, где r_a — радиус свода, то площадь сечения

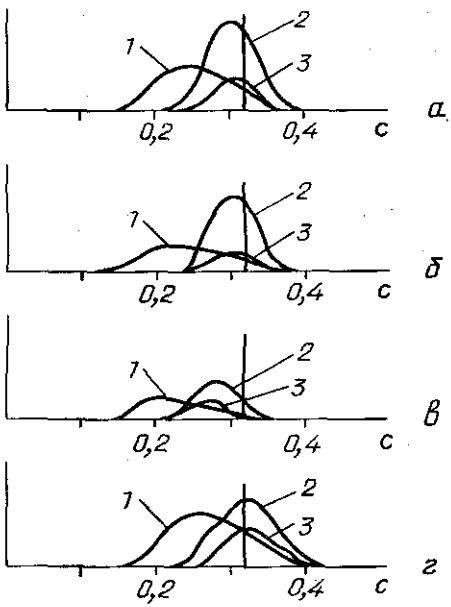


Рис. 8.11. Площадь голосовой щели для глухих взрывных /К/ (*а*), /Т/ (*б*), /П/ (*в*) и аффрикаты /Ц/ (*г*). *1* — удвоенные согласные, *2* — начальные, *3* — средние

$$S = \frac{\pi r_a^2}{2} + 2(R - r_a) r_a.$$

Если $R \leqslant r_a$, то $S = r_a^2 \arccos(1 - R/r_a)$.

В области корневой части языка расстояние R от языка до задней стенки равно просто горизонтальной координате поверхности языка (до высоты $y_t \leqslant y_p$, где y_p — вертикальная координата центра полярной системы языка). В области от $y_t > y_p$ до нёбной занавески расстояние от точки языка с координатами (x_t, y_t) до задней стенки измеряется вдоль линии, соединяющей центр полярной системы с этой точкой. Оно находится по (8.2). В области мягкого и твердого нёба

$$R = \sqrt{(x_{\text{пер}} - x_t)^2 + (y_{\text{пер}} - y_t)^2},$$

где $x_{\text{пер}}$, $y_{\text{пер}}$ — координаты точки на поверхности нёба, находящиеся на ближайшем расстоянии от точки на поверхности языка с координатами (x_t, y_t).

На фарингиальном участке радиус свода задней стенки r_a примерно равен 2 см. Начиная от задних зубов, ширина речевого тракта уменьшается почти по параболическому закону (рис. 8.13), а поверхность нёба представляет собой часть

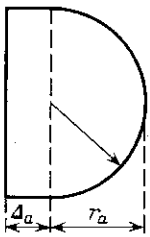


Рис. 8.12. Аппроксимация формы поперечного сечения речевого тракта

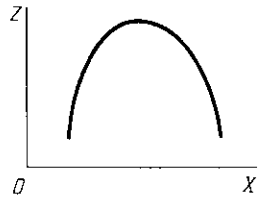


Рис. 8.13. Форма передней части речевого тракта в плоскости XOZ

поверхности параболоида. Радиус свода уменьшается по мере приближения к передним зубам как $r_a = \sqrt{x_{зп} - x_{пер}}$, где $x_{зп}$ — горизонтальная координата начала задних зубов. Величина r_a вычисляется не до передних зубов, где r_a равнялась бы нулю, а немного не доходя — примерно на 0,5 см по горизонтальной координате. От этой координаты до губ площадь поперечного сечения тракта вычисляется как площадь прямоугольника $S = Rr_a^*$, где R — расстояние от подъязычной поверхности до твердого нёба, r_a^* — ширина прохода на уровне зубов.

Площадь поперечного сечения на губах находим как

$$S = R_п \int_0^{l_г} \sin \frac{\pi z}{l_г} dz = \frac{2R_г l_г}{\pi},$$

где $l_г$ — длина губ, $R_г$ — расстояние между губами. Эта площадь вычисляется в предположении, что прогиб имеется лишь на нижней губе, а верхняя губа имеет постоянную координату y вдоль оси Z.

Изгибом губ в плоскости ZOY пренебрегаем.

Площадь поперечного сечения носовой полости остается постоянной, независимо от типа артикулированного звука. Обычно полагают, что разделенные носовой перегородкой полости симметричны, и поэтому в синтезаторах используют удвоенную площадь одной из полостей. Синтезированные таким способом носовые звуки, однако, обладают недостаточной степенью назализации. Для исправления этого недостатка в [153] предлагается учитывать дополнительные полости, в том числе гайморову пазуху. Это дополнение, однако, встречает возражение на том основании, что при определенных болезнях эти пазухи оказываются заполненными гноем, но это никак не сказывается на звучании носовых звуков. Возможно, что

все-таки необходимо использовать не одну, а две полости, поскольку из-за неизбежных отличий в геометрии каждой полости должны возникать дополнительные отражения акустических волн. Располагая формой и размерами речевого тракта, измеренными для одного диктора, можно получить многообразие речевых трактов путем деформации различных участков исходного тракта.

Размер прохода в носовую полость при опускании нёбной занавески оценивается как расстояние R_N по горизонтали от задней поверхности тракта до внутренней поверхности нёбной занавески. Форма этого прохода иногда довольно сложна, поэтому ее площадь рассчитывают как $S_N = 10R_N$ [146].

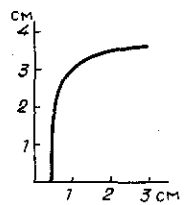


Рис. 8.14. Зависимость ширины губного прохода от высоты нижней челюсти

Вследствие смыкания губ из-за поверхностного натяжения и отклонения их поверхности от прямой линии в направлении оси Z площадь сечения губного отверстия оказывается несколько меньше. В [138] было показано, что ширина губного прохода меньше длины губ для высокого положения нижней челюсти (рис. 8.14). Поэтому для вычисления площади поперечного сечения нужно корректировать длину l_L в зависимости от угла поворота нижней челюсти.

Функция площади поперечного сечения речевого тракта оказывается представленной своими отсчетами, находящимися на неравном расстоянии l_{mi} друг от друга. Для перехода к равномерным отсчетам необходимо применить тот или иной вид интерполяции. Если величины l_{mi} малы, т. е. находятся в диапазоне 1—3 мм, то линейная интерполяция дает достаточно хорошую точность. В этом случае

$$S(j\Delta r) = S^*(r_i) + k_i\left[S^*(r_{i+1}) - S^*(r_i)\right], \qquad (8.4)$$

где $S(j\Delta r)$ — площадь поперечного сечения с равномерными отсчетами через Δr вдоль средней линии, $j = 1, \ldots, n$, $S^*(r_i)$ — площадь сечения с неравномерными отсчетами, $i = 1, \ldots, N$,

$$k_i = \frac{j\Delta r - r_{i+1}}{r_{i+1} - r_i}.$$

При этом проверяется условие $r_i \leqslant j\Delta r \leqslant r_{i+1}$ и при его невыполнении увеличивается номер отсчета i в таблице функций S^*.

Величина шага дискретизации площади поперечного сечения определяется следующими факторами. Она должна позволять различать место артикуляции переднеязычных фрикативных согласных /С/ и /Щ/. Не должны пропадать мелкие детали

в областях наибольшего сужения тракта, например, весьма важна форма зубов при артикуляции переднеязычных и губных согласных. Наконец, величина Δr равна удвоенной ошибке в оценке длины речевого тракта. Все эти факторы требуют уменьшения шага дискретизации, однако при этом возрастает число операций. В частности, при вычислении бегущих акустических волн число операций пропорционально квадрату числа цилиндрических секций, т. е. обратно пропорционально квадрату величины Δr. Таким образом, необходим компромисс, который достигается путем выбора $\Delta r = 3-5$ мм при аппроксимации формы речевого тракта цилиндрическими секциями равной длины.

Более эффективное представление функции $S(r)$ достигается путем использования отсчетов, находящихся на разных расстояниях, но кратных некоторой весьма малой величине Δr. При этом достигается необходимая точность описания формы и длины речевого тракта, а число операций в акустическом процессе не только не возрастает, но даже падает при надлежащем выборе способа интерполяции.

§ 8.5. Динамика площади поперечного сечения

В [59] рассматривались динамические характеристики голосовых складок, нижней челюсти, нёбной занавески, губ, языка. В дополнение к ним необходимо рассмотреть динамику смещения гортани, корня и кончика языка для того, чтобы получить полное описание динамики площади поперечного сечения речевого тракта. Масса гортани невелика (20—30 г), поэтому скорость переходных процессов для нее еще в большей степени, чем для нижней челюсти, определяется вязко-упругими свойствами связанных с нею мышц. Поэтому в первом приближении уравнение для изменения высоты гортани можно записать как

$$\Delta y''_{\text{гщ}} + 2g_{\text{гщ}}\Delta y'_{\text{гщ}} + \omega^2_{\text{гщ}}\Delta y_{\text{гщ}} = \Delta y^{(0)}_{\text{гщ}}(t),$$

где $\Delta y^{(0)}_{\text{гщ}}$ — управление положением гортани, $g_{\text{гщ}}$ — коэффициент вязких потерь ($8-20\ \text{с}^{-1}$), $\omega_{\text{гщ}}$ — частота, которую можно найти, исходя из свойств сокращения мышцы, нагруженной на массу

$$\omega_{\text{гщ}} = \frac{1}{l_{\text{м}}}\sqrt{\frac{E}{\rho_{\text{т}}}\left(\frac{m_{\text{м}}}{M_{\text{м}}} + 0{,}333\right)},$$

где $l_{\text{м}}$ — длина мышцы, E — модуль упругости, $\rho_{\text{т}}$ — погонная плотность тканей, $m_{\text{м}}$ — масса мышцы, $M_{\text{м}}$ — масса гортани.

К гортани присоединено несколько мышц, поэтому можно дать лишь приближенную оценку частоты $\omega_{\text{гщ}}$. Примем длину мышцы $l_{\text{м}} = 4$ см, $\rho_{\text{т}} = 0{,}8$ г/см (считая площадь сечения мышцы 0,8 см^2); массу мышцы $m_{\text{м}} = 4$ г, модуль упругости

$E=2\times10^3$ г/(см·с^2), отсюда частота $\omega_{\text{гщ}}=19$, т. е. $f_{\text{гщ}}=\omega_{\text{гщ}}/2\pi\approx$ $\approx$3 Гц. Если учесть упругость мышц, присоединенных с другого конца гортани, то собственная частота должна быть несколько увеличена, скажем до 4—5 Гц. Степень этого увеличения зависит от напряженности мышцы-антагониста.

Движения корня в большей степени определяются массой языка, и здесь основную роль играют вязко-упругие свойства мышц-антагонистов. Поэтому смещение корня языка описывается двумя дифференциальными уравнениями, одно из которых отражает динамику тела языка, а другое — динамику мышечного сокращения:

$$x''_{\text{кор}}+2g_{1\,\text{кор}}x'_{\text{кор}}+\omega^2_{1\,\text{кор}}x_{\text{кор}}=f_{x\,\text{кор}}(t),$$
$$y''_{\text{кор}}+2g_{1\,\text{кор}}y'_{\text{кор}}+\omega^2_{1\,\text{кор}}y_{\text{кор}}=f_{y\,\text{кор}}(t),$$
$$f''_{x\,\text{кор}}+2g_{2\,\text{кор}}f'_{x\,\text{кор}}+\omega^2_{2\,\text{кор}}f_{x\,\text{кор}}=F_{x\,\text{кор}}(t),$$
$$f''_{y\,\text{кор}}+2g_{2\,\text{кор}}f'_{y\,\text{кор}}+\omega^2_{2\,\text{кор}}f_{y\,\text{кор}}=F_{y\,\text{кор}}(t),$$

где $F_{x\,\text{кор}}$, $F_{y\,\text{кор}}$ — целевое положение координат корня языка $x_{\text{кор}}$ и $y_{\text{кор}}$. Смещение корня языка вверх и назад достигается сокращением одной мышцы — styloglossus, а опускание и сдвиг вперед — нижними волокнами подъязычно-подбородочной мышцы [59]. Поэтому коэффициенты в уравнениях для вертикальной и горизонтальной координат — одни и те же. Частота $\omega_{1\,\text{кор}}$ может быть оценена как близкая к частоте первой моды упругих деформаций языка, т. е. около 15 рад/с, или немного меньше 3 Гц. Собственная частота мышцы несколько выше — 3—4 Гц.

Координаты кончика языка, которые используются для вычисления максимального значения угла в полярной системе языка, вообще говоря, являются функцией от степени деформаций языка. Эти координаты могут быть вычислены, исходя из условия несжимаемости языка — сохранения его длины. *Практически, однако, удобнее задавать траектории перехода* от одного положения кончика языка к другому в виде решения дифференциальных уравнений с одними и теми же коэффициентами

$$x''_{\text{кону}}+2g_{\text{кону}}x'_{\text{кону}}+\omega^2_{\text{т}}x_{\text{кону}}=x^*_{\text{кону}},$$
$$y''_{\text{кону}}+2g_{\text{кону}}y'_{\text{кону}}+\omega^2_{\text{т}}y_{\text{кону}}=y^*_{\text{кону}},$$

где $x^*_{\text{кону}}$, $y^*_{\text{кону}}$ — целевые положения координат кончика языка. Если смещение преимущественно вертикальное, то $\omega_{\text{т}}$ близко к собственной частоте кончика языка, а при доминировании горизонтальных смещений $\omega_{\text{т}}$ близко к первой собственной частоте языка в целом.

Использование дифференциальных уравнений для описания движений артикуляторных органов не только удовлетворяет требованию сохранения физической адекватности динамической модели артикуляции. Решения дифференциальных уравнений

гарантируют непрерывность движений и разрешают проблему согласования целевых и текущих значений координат, поскольку начальные условия естественным образом и абсолютно автоматически изменяют целевую команду и обеспечивают достижение целевого положения из любого начального положения. В сложных случаях коартикуляции, когда возникают конфликты в системе уравнения, применяются методы, описанные в гл. 9.

Итак, координаты любой точки подвижной поверхности речевого тракта движутся по траекториям, определенным решениями дифференциальных уравнений второго порядка. Однако отсчеты функции площади изменяются уже как нелинейно преобразованные координаты подвижной поверхности. Стационарное состояние не достигается в течение значительного интервала времени, определяемого длительностью наиболее медленного переходного процесса. Эти свойства являются необходимыми компонентами натуральности звучания речи, генерируемой артикуляторным и формантно-артикуляторным синтезаторами.

ГЛАВА 9

УПРАВЛЕНИЕ АРТИКУЛЯТОРНЫМ СИНТЕЗАТОРОМ

§ 9.1. Внутренняя модель артикуляции — случай линейного программирования

Система управления синтезатором является существенной частью, обеспечивающей необходимое качество синтетической речи. В артикуляторном синтезаторе система управления не служит чем-то внешним по отношению к синтезатору, а использует те же механизмы, которые лежат в основе синтеза речи на различных уровнях. Как и всякая система управления, система управления синтезатором действует в соответствии с заданными целями речевого общения. Наиболее важная и часто используемая цель — это создание необходимой в данных условиях разборчивости речевого сообщения. Весьма важна также цель передачи личной оценки к высказыванию. В отношении синтезатора использование понятия «личной оценки» может показаться ненужным антропоморфизмом. Однако поскольку люди пользуются различными стилями разговора, обеспечивая разную разборчивость в зависимости от обстоятельств, темы разговора и межличностных отношений, то невольно и синтетическая речь будет оцениваться с точки зрения именно этих критериев. Поэтому важное сообщение, произнесенное скороговоркой, может не достичь своей цели, а чересчур отчетливо выговариваемая тривиальная фраза вызовет раздражение в связи с непроизвольным повышением внимания слушателя. Точно так же, речевое сообщение, генерируемое в присутствии помех, должно отличаться, и не только громкостью, от сообщения, предназначенного для восприятия в хороших акустических условиях. Следовательно, даже на уровне глобальной цели — передачи речевого сообщения с необходимой разборчивостью, количественное значение обеспечиваемой разборчивости может быть разным, и, как мы увидим ниже, разными оказываются и формируемые команды управления.

Степень разборчивости, однако, понятие в значительной мере субъективное, и система управления достигает требуемой

в данных условиях разборчивости только путем формирования определенных акустических параметров речевого сигнала. Эти параметры в свою очередь зависят от формы речевого тракта, а форма тракта — от артикуляторных параметров (поворота нижней челюсти, смещения губ, языка, небной занавески и т. д.). Система управления должна обеспечить возможность создания индивидуальных голосов без потери разборчивости. Наконец, свойства системы управления играют определяющую роль в оптимизации синтезатора, т. е. в поиске таких команд управления, которые обеспечивают максимальную натуральность и разборчивость речи.

Таким образом, в системе управления одновременно действует иерархия целей, связанных друг с другом физическими свойствами процессов речеобразования на различных уровнях. В таких условиях система управления может успешно решать задачи, лишь располагая математическими моделями всех управляемых процессов, т. е. в системе управления должна находиться так называемая внутренняя модель.

Понятие внутренней модели не специфично для системы управления речеобразованием, оно присутствует во многих задачах, связанных с управлением движениями человека. Рассмотрим, например, задачу управления положением кончика указательного пальца правой руки, потребовав перенести его из одной точки пространства в другую. Мы выполняем это движение, контролируя его с помощью зрения и механорецепторов. С большей или меньшей точностью можно осуществить это движение и с закрытыми глазами. Когда невропатологи просят коснуться пальцем кончика носа, закрыв при этом глаза, они проверяют таким образом состояние системы управления движениями.

Получая информацию о текущем положении кончика пальца и сравнивая его с заданным положением, система управления должна выдать команды на многочисленные мышцы, управляющие положением руки, а в некоторых случаях и тела (если целевое положение находится достаточно далеко). Ясно, что выработать такие команды система управления в состоянии лишь в том случае, если она заранее знает, к какому перемещению кончика пальца приведет сокращение тех или иных мышц, т. е. располагает кинематической моделью своего тела. В задачах на выполнение заданной траектории потребуется также и динамическая модель. Если движение выполняется с закрытыми глазами, то оно возможно лишь в том случае, если имеется внутренняя модель окружающего пространства. Необходимо обратить внимание и на несоответствие размерностей в разных пространствах управления — если мерой выполнения задачи является лишь перемещение кончика пальца в заданную точку пространства, то в этом случае задача управления одномерна, так как ее целью является сведение к нулю расстояния между кончиком пальца и заданной точкой.

Однако в пространстве мышечных усилий размерность может исчисляться десятками координат.

Другой пример, более близкий к управлению речеобразованием — повторение голосом ранее не слышанной мелодии. В этом случае цель задается в пространстве акустических *ощущений — высоты и длительности тональных компонент* сигнала. Основываясь на этой информации, система управления должна сформировать такую последовательность команд на сокращения мышц гортани, чтобы частота основного тона соответствовала заданной мелодии. Снова ясно, что переход из пространства одной физической природы (акустических параметров) и одной степени свободы (частоты) к пространству другой физической природы и размерности (пространству мышц гортани) не может произойти иначе как путем использования внутренней модели, отображающей точки одного *пространства на другое.*

Внутренняя модель в системе управления движениями человека, таким образом, решает сразу две задачи: планирование управления и пересчета сигналов рецепторов в команды управления, т. е. замыкания обратной связи. Представления о роли внутренней модели в целенаправленных движениях обсуждалось в работах [14, 28, 67, 128, 182, 187, 188], хотя не всегда использовался этот термин. В применении к процессам речеобразования понятие внутренней модели было введено в [59, 191]. Основная трудность, препятствовавшая до сих пор разработке алгоритмов для использования внутренней модели в управлении сложными движениями состоит в различии числа целевых функций и числа управляемых параметров. Эта трудность преодолевается, если допустить, что решение своей основной задачи — достижение заданного положения в пространстве целей системы управления — осуществляется путем оптимизации некоторого критерия. Тогда задача управления *движениями, в том числе и артикуляцией,* сводится к вариационной задаче.

Существенным элементом такой вариационной задачи является наличие ограничений на области допустимых значений управляемых параметров. В результате этого управляющие команды и совершаемое движение зависят не только от цели движения, но и от критерия оптимальности и ограничений, действующих на данном интервале времени. Эта зависимость не только не сужает возможности системы управления, но и значительно расширяет их, позволяя достичь заданной цели *в различных обстоятельствах, в том числе и в случаях* поражения отдельных участков двигательной системы.

Известны случаи, когда при параличе мышц нижней челюсти артикуляция губных согласных /Б, П, М, В, Ф/ осуществлялась почти без изменения акустических характеристик речи за счет увеличения диапазона движений губ. Аналогичная компенсация искусственно наложенных

ограничений на подвижность нижней челюсти наблюдалась в экспериментах, описанных в [99, 109].

Прежде чем перейти к описанию алгоритмов управления артикуляцией, рассмотрим несколько вариантов частного случая управления совместными движениями нижней челюсти и губ, что поможет лучше проиллюстрировать возникающие при этом проблемы с помощью весьма простого математического аппарата.

Одним из параметров, определяющих фонетическое качество гласных и губных согласных, служит расстояние между губами $h_г$. Для взрывных согласных на интервале смычки это расстояние должно быть равно нулю, для фрикативных /В/ и /Ф/ оно должно быть достаточно мало для обеспечения возникновения турбулентного шума, а открытые и закрытые гласные различаются также и по величине $h_г$.

Уравнение связи для положения нижней челюсти и губ есть

$$h_г = y_{вг} - (y_{нч} + \Delta y_{нг}),$$

где $y_{вг}$ — вертикальная координата верхней губы в некоторой неподвижной системе координат XOY, $y_{нч}$ — координата некоторой точки на нижней челюсти, $\Delta y_{нг}$ — смещение нижней губы относительно этой точки. Нижняя челюсть и нижняя губа обычно движутся в одном направлении, а верхняя губа — в противоположном с ними направлении. Таким образом, имеем одномерное пространство целей, содержащее расстояние $h_г$, и трехмерное пространство управляемых параметров $y_{вг}$, $y_{нг}$ и $y_{нч}$. Для наибольшего упрощения задачи рассмотрим порознь движения нижней губы и нижней челюсти, и движения губ.

Пусть целью системы управления служит некоторая высота нижней губы, $h_{нг} = y_{нч} + \Delta y_{нг}$. Как видно, одна и та же величина $h_{нг}$ может быть достигнута путем использования различных сочетаний движений нижней челюсти и нижней губы. Следовательно, для решения этой задачи необходимо ее доопределение путем использования дополнительной информации об условиях функционирования системы управления артикуляцией. Допустим, что система управления действует под влиянием критерия оптимальности, состоящего в минимизации суммы мышечных усилий, т. е. $\min(F_{нч} + F_{нг})$. Тогда, рассматривая лишь стационарные состояния и учитывая только упругое сопротивление, можно записать

$$c_{нч} y_{нч} + c_{нг} \Delta y_{нг} = \min, \qquad (9.1)$$

где $c_{нч}$ и $c_{нг}$ — коэффициенты жесткости для нижней челюсти и нижней губы.

Если выбрать начало системы координат XOY так, чтобы при максимальном опускании нижней челюсти $\Delta y_{нг}$ равнялось нулю, то имеем следующие ограничения на смещения ар-

тикуляторных органов:

$$0 \leqslant y_{\text{нч}} \leqslant a_1,$$
$$0 \leqslant y_{\text{нг}} \leqslant a_2.$$

Начальные условия примем в виде $y_{\text{нч}}=0$, $\Delta y_{\text{нг}}=0$. Поскольку величина $h_{\text{нг}}$ известна (она задана системе управления типом артикулируемого звука), то можно выразить одну неизвестную через другую: $y_{\text{нч}}=h_{\text{нг}}-\Delta y_{\text{нг}}$. Тогда критерий оптимальности (9.1) изменится на (9.2):

$$(c_{\text{нч}}-c_{\text{нг}})\, y_{\text{нч}}+c_{\text{нг}}h_{\text{нг}}=\min, \tag{9.2}$$

а неизвестная $y_{\text{нч}}$ приобретает новые ограничения

$$\max(0,\ h_{\text{нг}}-a_{\text{нг}}) \leqslant y_{\text{нч}} \leqslant \min(a_{\text{нч}},\ h_{\text{нг}}). \tag{9.3}$$

Как видно, левая часть (9.2) представляет собой уравнение прямой линии, наклон которой зависит от разности жесткостей $c_{\text{нч}}$ и $c_{\text{нг}}$. Ясно, что минимум достигается на одной из границ диапазона допустимых значений $y_{\text{нч}}$. При $c_{\text{нч}}>c_{\text{нг}}$ это происходит для $y_{\text{нч}}=0$ и $\Delta y_{\text{нг}}=h_{\text{нг}}$, если $h_{\text{нг}} \leqslant a_{\text{нг}}$, или для $y_{\text{нч}}=h_{\text{нг}}-a_{\text{нг}}$, $\Delta y_{\text{нг}}=h_{\text{нг}}$, если $h_{\text{нг}} \leqslant a_{\text{нг}}$. Когда присоединенная к нижней челюсти жесткость $c_{\text{нч}}$ меньше жесткости губы $c_{\text{нг}}$, то минимум усилий достигается при $y_{\text{нч}}=a_1$, $\Delta y_{\text{нг}}=h_{\text{нг}}-a_1$, если $h_{\text{нг}}>a_1$, или при $y_{\text{нч}}=h_{\text{нг}}$, $\Delta y_{\text{нг}}=0$, если $h_{\text{нг}} \leqslant a_1$. Иными словами, на максимальное расстояние смещается тот артикуляторный орган, присоединенная жесткость которого наименьшая. Если при этом достигается поставленная цель, то задача решена. В противном случае требуемая высота $h_{\text{нг}}$ обеспечивается дополнительным смещением другого артикуляторного органа на необходимую величину.

Рассмотрим теперь встречное движение губ в случае реализации губной смычки. При этом условие смычки есть $y_{\text{вг}}-y_{\text{нг}}=0$, а ограничения

$$a_{\text{нг}} \leqslant y_{\text{нг}} \leqslant b_{\text{нг}},$$
$$a_{\text{вг}} \leqslant y_{\text{вг}} \leqslant b_{\text{вг}},$$

где высота $y_{\text{нг}}$ — высота нижней губы с учетом положения нижней челюсти. Начальные условия примем как $y_{\text{нг}}=a_{\text{нг}}$, $y_{\text{вг}}=b_{\text{вг}}$. Силы упругого сопротивления есть $F_{\text{вг}}=c_{\text{вг}}(b_{\text{вг}}-y_{\text{вг}})$, $F_{\text{нг}}=c_{\text{нг}}(y_{\text{нг}}-a_{\text{нг}})$, так что критерий минимума суммарных усилий есть

$$(c_{\text{нг}}-c_{\text{вг}})\, y_{\text{нг}}+c_{\text{вг}}b_{\text{вг}}-c_{\text{нг}}a_{\text{нг}}=\min, \tag{9.4}$$

а переменная $y_{\text{нг}}$ имеет новые ограничения

$$\max(a_{\text{нг}},\ a_{\text{вг}}) \leqslant y_{\text{нг}} \leqslant \min(b_{\text{нг}},\ b_{\text{вг}}). \tag{9.5}$$

Опять (9.4) представляет собой линейную функцию, и минимум достигается на одной из границ допустимого диапазона смещений верхней и нижней губ. Поскольку нижняя губа

более подвижна в вертикальном направлении, чем верхняя, то на ее смещение требуется меньшее усиление. Поэтому можно принять $c_{нг} < c_{вг}$, и минимум усилий достигается при $y_{нг} = \min(b_{нг}, b_{вг})$, т. е. при $y_{нг} = b_{нг}$, так как $b_{вг} > b_{нг}$. Таким образом, нижняя губа смещается на большее расстояние, чем верхняя, и, если смычка при этом достигается, то задача решена. В противном случае верхняя губа опускается на расстояние, оставшееся до смыкания губ.

С математической точки зрения оба рассмотренных примера относятся к области линейного программирования, поскольку и критерий оптимальности, и ограничения представляют собой линейные формы. Основной результат теории линейного программирования состоит в том, что минимум оптимизируемой функции всегда достигается в одной из вершин многоугольника, образованного пересечением гиперплоскостей ограничений в пространстве искомых неизвестных. Это означает, что на максимальное расстояние смещаются те артикуляторные органы, «стоимость» движения которых с точки зрения принятого критерия оптимальности наименьшая. Этот результат мы и получили в только что рассмотренных примерах.

Критерий минимума усилий, однако, не является единственным критерием, которым может руководствоваться система управления артикуляцией. Посмотрим, что произойдет, если потребуется минимизировать не усилия, а работу $A = F\Delta x$, где Δx — смещение артикуляторного органа. Тогда в задаче совместного движения нижней челюсти и нижней губы необходимо обеспечить

$$(c_{нг} + c_{нч})\, y_{нч}^2 - 2c_{нг} h_{нг} y_{нч} + c_{нг} h_{нг}^2 = \min. \qquad (9.6)$$

Это уже квадратичная форма и минимум не обязательно достигается на одной из границ диапазона смещений артикуляторов. Приравняв нулю производную от левой части (9.6), получим, что минимум может иметь место в точке

$$y_{нч}^* = c_{нг} h_{нг} / (c_{нг} + c_{нч}),$$

если $y_{нч}^*$ удовлетворяет условиям (9.3). Если же $y_{нч}^* < h_{нг} - a_{нг}$, при $h_{нг} > a_{нг}$, то $y_{нч} = h_{нч} - a_{нг}$, $\Delta y_{нг} = a_{нг}$; а при $h_{нг} < a_{нг}$ и $y_{нч}^* < 0$ имеем $\Delta y_{нг} = 0$, $\Delta y_{нг} = h_{нг}$. В случае, когда $y_{нч}^* > \min(a_{нч}, h_{нг})$, при $a_{нч} > h_{нг}$ имеем $y_{нч} = a_{нч}$, $\Delta y_{нг} = h_{нг} - a_{нч}$, а при $a_{нч} < h_{нг}$ имеем $y_{нч} = h_{нг}$, $\Delta y_{нг} = 0$.

Для встречного движения губ с критерием минимума работы имеем

$$(c_{нг} - c_{вг})\, y_{нг}^2 + (c_{вг} b_{вг} - c_{нг} a_{нг})\, y_{нг} = \min. \qquad (9.7)$$

Поскольку по условию $b_{вг} > a_{нг}$ и $c_{вг} > c_{нг}$, то левая часть (9.7) представляет собой параболу, обращенную выпуклостью вверх, и минимум достигается на одной из границ смещения артикуляторов, точнее, либо при $y_{нг} = a_{вг}$, либо при $y_{нг} = b_{нг}$. Для объяснения этого выбора приведен рис. 9.1, на котором

показана схема допустимых движений губ. Тогда, если точка максимума

$$y^* = \frac{c_{вг} a_{нг} - c_{вг} b_{вг}}{2(c_{нг} - c_{вг})}$$

удовлетворяет условиям (9.5), то $y_{нг} = a_{вг}$, если $|y^* - a_{вг}| > |y^* - b_{вг}|$, и $y_{нг} = b_{вг}$, если $|y^* - a_{вг}| < |y^* - b_{нг}|$. Если $y^* < a_{вг}$, то $y_{нг} = b_{нг}$, а при $y^* > b_{нг}$ имеем $y_{нг} = a_{вг}$. Если бы жесткость верхней губы $c_{вг}$ была меньше жесткости нижней губы $c_{нг}$, то, как и в случае совместного движения нижней челюсти и губы, оптимум мог бы достигаться и где-то внутри диапазонов допустимых смещений артикуляторов.

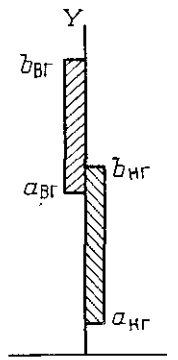


Рис. 9.1. Схема вертикальных движений губ

Мы видим, таким образом, что при другом критерии оптимальности положение точки оптимума может зависеть не только от механических параметров артикуляторов (в данном случае, присоединенной жесткости), но и от величины требуемого смещения h. Это означает, что относительный вклад смещений каждого артикулятора не постоянен, а определяется артикулируемым звуком. Представляется, что критерий минимума работы более адекватен наблюдаемым свойствам артикуляции.

При изменении темпа артикуляции меняются скорости и ускорения движения артикуляторных органов и, соответственно, меняются и потребные усилия для мышечных сокращений. В результате этого положение оптимума изменяется, и одна и та же цель в пространстве конфигураций речевого тракта достигается с помощью других сочетаний сокращающихся мышц и движений артикуляторов. Примеры такой перестройки *действий артикуляторного аппарата при взаимодействии языка* и нижней челюсти приводятся в [59]. Таким образом, сумма экспериментальных данных о свойствах системы управления артикуляцией и теоретический анализ дают основания для предположения, что система управления артикуляцией решает вариационную задачу по оптимизации некоторого критерия с заданными ограничениями и использование этого аппарата для управления артикуляторным синтезатором.

§ 9.2. Внутренняя модель артикуляции — нелинейное программирование

Рассмотрим теперь возможный алгоритм для системы управления на уровне связи между артикуляторными параметрами и формой речевого тракта. В предыдущей главе мы показали, что площадь поперечного сечения однозначно, хотя

и нелинейно, связана с расстояниями между подвижными и неподвижными поверхностями речевого тракта в среднесагиттальной плоскости. Поэтому в дальнейшем будем заниматься, в основном, описанием именно расстояния r, как функции от координаты z вдоль средней линии речевого тракта.

Как и всякая функция, $r(z)$ может быть с любой степенью точности представлена своими отсчетами в конечном числе точек. Для этого можно было бы использовать приемы теории интерполяции функций, указывающей число и расположение узлов интерполяции при заданных требованиях на точность аппроксимации и известных свойствах аппроксимируемой функции $r(z)$. Пользуясь результатами предыдущей главы, однако, мы имеем гораздо больше информации о $r(z)$, чем просто сведения о ее гладкости, поскольку знаем условия порождения этой функции. Поэтому и задача интерполяции $r(z)$ решается более содержательными средствами, чем в теории интерполяции абстрактных функций.

Расположение узлов интерполяции в нашем случае диктуется артикуляторными свойствами речи. Для русского языка можно указать семь координат z, расстояния r в которых важны для создания всех фонетически различимых звуков. Это расстояние между губами, между кончиком языка и зубами и двумя передними участками твердого неба, расстояние до мягкого неба и до задней поверхности фаринкса (см. рис. 9.2). Расстояния в сечении *1* важны для артикуляции губных звуков, в сечении *2* — при артикуляции /Т, Д/, в сечении *3* — при артикуляции /С, З/, в сечении *4* — при артикуляции /Ш, Ж/. Расстояние в сечении *5* определяет степень мягкости согласных звуков, важно при артикуляции гласного /И/ и полугласного /Й/. Расстояние в сечении *6* характеризует артикуляцию заднеязычных согласных /К, Г, Х/, а расстояние в сечении *7* влияет на частоту первой форманты гласных звуков.

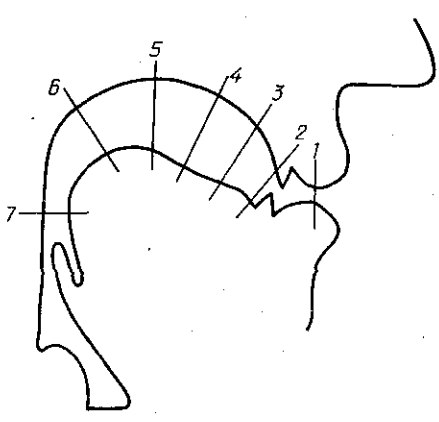


Рис. 9.2. Контрольные сечения

Концентрация контролируемых сечений в передней части речевого тракта объясняется, в основном, анатомическим строением артикуляторного аппарата и его возможностями точного управления положением передней части языка.

Конкретные координаты контролируемых сечений различны для разных артикуляторных наборов (т. е. для разных языков), но они дают полное описание формы речевого тракта,

необходимое для различения всех фонетически значимых артикуляторных состояний для любого языка. Набор контролируемых сечений физически более нагляден и доступен измерениям, поэтому, если бы удалось создать алгоритм вычисления артикуляторных параметров по расстояниям в контролируемых сечениях, это намного упростило бы процесс создания и оптимизации артикуляторного синтезатора для любого языка. Как показали рассмотренные выше примеры, существует принципиальная возможность построения такого алгоритма, если задачу отображения в пространстве расстояний в контролируемых сечениях на пространство артикуляторных параметров рассматривать как вариационную задачу. Необходимо, однако, конкретизировать условия этой задачи.

Пусть нам известно множество расстояний $\{r_i^{(0)}\}$ и соответствующее им множество значений артикуляторных параметров $\{p_j^{(0)}\}$ в некоторый момент времени $t=0$. Требуется найти такие артикуляторные параметры $\{p_j^{(T)}\}$, которые в момент времени $t=T$ обеспечивают достижение заранее заданных расстояний $\{r_i^{(T)}\}$ при условии минимума некоторого критерия, например, суммарной работы, выполняемой при движении артикуляторов. В начальный момент времени артикуляторные параметры могут обладать ненулевой скоростью и ускорением. Поэтому задачу оптимизации, вообще говоря, нужно решать во времени, например, средствами, предлагаемыми принципом максимума [43]. Но при решении этой задачи необходимо принимать во внимание свойства системы управления движениями человека. Так, известно, что при движениях рук и даже глаз смена одной программы управления на другую не может произойти быстрее, чем за 150 мс, а коррекция текущей программы — не ранее, чем через 30—50 мс [188]. Эти последние цифры включают время реакции механорецепторов мышц (веретен), время обработки сигналов обратной связи в центральной нервной системе и время распространения по нервным каналам. При управлении артикуляцией наименьшее время реакции на сигналы обратной связи лишь несколько меньше — 20—50 мс [59]. Отсюда следует, что при управлении движениями артикуляторов, чьи переходные процессы длятся 50—100 мс, система управления может корректировать цели движения лишь несколько раз за время переходного процесса. Поэтому на первых порах можно принять квазистатический вариант вариационной задачи управления, в котором цель и управляющие воздействия вычисляются лишь один раз для данного фонетического сегмента. Для того чтобы математически задача была корректной, допустим, что всякое новое управляющее воздействие совершается лишь после затухания переходных процессов от предыдущего воздействия, так что скорости dp_j/dt и ускорения d^2p_j/dt^2 артикуляторов равны нулю.

Поскольку мы приняли, что находимся в статическом режиме, то работу W на перемещение каждого артикуляторного органа можно определить как произведение силы, приложенной для преодоления упругого сопротивления $F_i = c_i \Delta\xi_i$ на смещение артикуляторного органа $\Delta\xi_i$,

$$W = \sum_{i=1}^{f} c_i \Delta\xi_i^2. \tag{9.8}$$

Выпишем смещения в явном виде для всех управляемых координат артикуляторного аппарата. Для корня языка имеем

$$\Delta\xi_R^2 = \left(x_{1\,\text{кор}}^{(0)} - x_{1\,\text{кор}}^{(T)}\right)^2 + \left(y_{1\,\text{кор}}^{(0)} - y_{1\,\text{кор}}^{(T)}\right)^2.$$

Индекс 1 означает, что координаты измеряются в системе $X_1 O_1 Y_1$, связанной с нижней челюстью (см. предыдущий раздел). Для кончика языка

$$\Delta\xi_T^2 = \left(x_{1\,\text{кон}}^{(0)} - x_{1\,\text{кон}}^{(T)}\right)^2 + \left(y_{1\,\text{кон}}^{(0)} - y_{1\,\text{кон}}^{(T)}\right).$$

Полагаем также, что поворот нижней челюсти на угол α сопровождается упругим сопротивлением, пропорциональным α:

$$\Delta\xi_\alpha = \alpha^{(0)} - \alpha^{(T)}.$$

Вертикальные смещения губ описываются приращениями их координат по оси Y:

$$\Delta\xi_{\text{вг}} = y_{\text{вг}}^{(0)} - y_{\text{вг}}^{(T)},$$
$$\Delta\xi_{\text{нг}} = y_{1\,\text{нг}}^{(0)} - y_{1\,\text{нг}}^{(T)}.$$

Отметим, что смещение верхней губы $\xi_{\text{вг}}$ измеряется в неподвижной системе координат XOY. Горизонтальное смещение точки вращения *нижней челюсти* есть

$$\Delta\xi_{xJ} = x_J^{(0)} - x_J^{(T)}.$$

Формально положим, что изменение коэффициентов k_n при собственных функциях языка также встречает упругое сопротивление, хотя это и не имеет прямого физического смысла, как для всех предыдущих смещений:

$$\Delta\xi_n = k_n^{(0)} - k_n^{(T)}, \quad n = 1, \ldots, 5.$$

В результате имеем 11 смещений при 13 регулируемых координатах точек артикуляторного аппарата.

Расположение контролируемых сечений, по-видимому, должно задаваться не в абсолютных величинах, а в относительных, поскольку размеры речевых трактов у разных людей различны. Но для каждого конкретного тракта эти величины могут быть вычислены как абсолютные, и каждое контрольное сечение задается набором координат (x_j, y_j) точек на верхней губе или неподвижной поверхности речевого тракта. Затем нужно определить направление, в котором измеряется заданное

расстояние R_j в j-й точке. Эта задача решается аналогично тому, как это было сделано в предыдущем разделе. Например, для точки 1 на губах направление кратчайшего расстояния — вертикаль, и

$$R_1 = y_{\text{вг}} - y_{\text{нг}},$$

где $y_{\text{вг}}$, $y_{\text{нг}}$ — координаты ближайших точек поверхностей верхней и нижней губ, измеренные в системе XOY. Из описания кинематики нижней челюсти имеем

$$y_{\text{нг}} = (y_{1\,\text{нг}} + y_{01} - y_J + \Delta y_{\text{нг}})\cos\alpha - (x_{1\,\text{нг}} + x_{01})\sin\alpha + y_J,$$

где $x_{1\,\text{нг}}$, $y_{1\,\text{нг}}$ — координаты прикрепления нижней губы к челюсти в системе $X_1O_1Y_1$, x_{01}, y_{01} — координаты начала подвижной системы, связанной с нижней челюстью, x_J, y_J — координаты точки вращения нижней челюсти, α — угол поворота нижней челюсти, $\Delta y_{\text{нг}}$ — независимое смещение нижней губы относительно нижней челюсти. Поскольку угол α даже для наибольшего опускания челюсти мал, то можно осуществить замену $\sin\alpha \approx \alpha$, $\cos\alpha \approx 1$. Отсюда получаем

$$y_{\text{нг}} = A_1 - B_1\alpha + \Delta y_{\text{нг}},$$

где $A_1 = y_{1\,\text{нг}} + y_{01}$, $B_1 = x_{1\,\text{нг}} + x_{01}$ — постоянные для данного речевого тракта величины. Тогда условие на губах, или ограничение в первом контрольном сечении есть

$$r_1 = y_{\text{вг}}^{(T)} - \Delta y_{\text{нг}}^{(T)} + B_1\alpha - A_1. \qquad (9.9)$$

Ограничения в других точках записываются как

$$r_j^2 = [x_j - x(\varphi)]^2 + [y_j - y(\varphi)]^2, \qquad (9.10)$$

где $x(\varphi)$, $y(\varphi)$ — координаты поверхности языка, а угол φ определяется направлением из точки (x_j, y_j) в точку (x_p, y_p), т. е. точку центра полярной системы языка. Из рис. 9.3 видно, что

$$\gamma = \operatorname{arctg}\frac{y_j - y_p}{x_j - x_p},$$

$$\theta = \operatorname{arctg}\frac{y_{1\,\text{кон}} - y_{1\,\text{кор}}}{x_{1\,\text{кон}} - x_{1\,\text{кор}}}, \qquad (9.11)$$

где $x_{1\,\text{кон}}$, $y_{1\,\text{кон}}$ — координаты кончика языка в системе $X_1O_1Y_1$, $x_{1\,\text{кор}}$, $y_{1\,\text{кор}}$ — координаты корня языка в этой же системе, $\varphi = \gamma + \theta$.

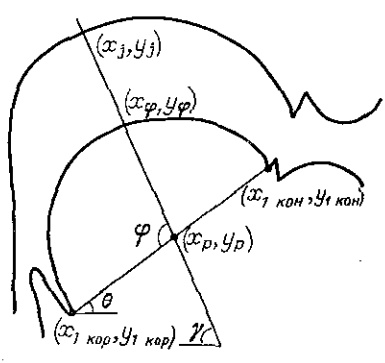


Рис. 9.3. Управляемые координаты языка

Ограничение типа равенства (9.10) предъявляется не ко всем точкам j, а лишь к некоторым, а именно к тем, которые определяют место артикуляции данного звука, т. е. приоритетным сечениям. Например, для согласных требуется лишь

два — три ограничения типа равенства — в точке смычки или щели, в точке 5, определяющей твердость или мягкость и, может быть, в точке 7, значение расстояния в которой больше для звонких взрывных, чем для глухих. В остальных точках должны применяться ограничения типа неравенства

$$r_j^2 \geqslant [x_j - x(\varphi)]^2 + [y_j - y(\varphi)]^2. \qquad (9.12)$$

Это означает, что в этих точках расстояния не строго определены, а относительно произвольны. Условие состоит лишь в том, чтобы никакое расстояние не было меньше наименьшего приоритетного расстояния. В противном случае место артикуляции сместится в эту точку и фонетическое качество звука изменится. Если одно из приоритетных расстояний равно нулю (как для взрывных согласных), то никакое другое расстояние r_j не должно быть меньше того, при котором возникают турбулентные шумы, которые могли бы замаскировать фонетическое качество звука в момент взрыва смычки.

Координаты поверхности языка $x(\varphi)$, $y(\varphi)$, с учетом того, что угол γ отсчитывается по направлению часовой стрелки, есть

$$x(\varphi) = x_p - s(\varphi)\cos\gamma,$$
$$y(\varphi) = y_p + s(\varphi)\sin\gamma,$$

где

$$s(\varphi) = \sum_{n=1}^{5} k_n \psi_n(\varphi),$$

ψ_n — собственные функции языка. Из тригонометрических соотношений имеем

$$\cos\gamma = \cos\left(\operatorname{arctg}\frac{y_j - y_p}{x_j - x_p}\right) = \frac{x_j - x_p}{r_{jp}},$$
$$\sin\gamma = \sin\left(\operatorname{arctg}\frac{y_j - y_p}{x_j - x_p}\right) = \frac{y_j - y_p}{r_{jp}},$$

где r_{jp} — расстояние между точкой (x_j, y_j) и центром полярной системы координат языка (x_p, y_p):

$$r_{jp} = \sqrt{(x_j - x_p)^2 + (y_j - y_p)^2}. \qquad (9.13)$$

Отсюда имеем

$$x(\varphi) = x_p + \frac{x_j - x_p}{r_{jp}} s(\varphi),$$
$$y(\varphi) = y_p + \frac{y_j - y_p}{r_{jp}} s(\varphi). \qquad (9.14)$$

Координаты x_p, y_p в неподвижной системе XOY записываются как

$$x_p = x_{1p}\cos\alpha + (y_{1p} - y_J)\sin\alpha + x_J,$$
$$y_p = (y_{1p} - y_J)\cos\alpha + x_{1p}\sin\alpha + y_J,$$

или с учетом малости угла α

$$x_p = x_{1p} + x_J + (y_{1p} - y_J)\alpha,$$
$$y_p = y_{1p} + (x_{1p} - x_J)\alpha,$$

где x_{1p}, y_{1p} — координаты центра полярной системы языка в системе $X_1 O_1 Y_1$. В предыдущей главе мы приняли, что

$$x_{1p} = 0{,}5(x_{1\,\text{нз}} + x_{1\,\text{кор}}),$$
$$y_{1p} = 0{,}5(y_{1\,\text{нз}} + y_{1\,\text{кор}}),$$

где $x_{1\,\text{нз}}$, $y_{1\,\text{нз}}$ — координаты точки на нижнем зубе, с которой совмещается кончик языка в нейтральном состоянии. Эта координата постоянна для выбранного речевого тракта. Поэтому окончательное выражение для x_p, y_p выглядит как

$$\begin{aligned} x_p &= 0{,}5(x_{1\,\text{нз}} + x_{1\,\text{кор}}) + (D + 0{,}5y_{1\,\text{кор}})\alpha + x_J, \\ y_p &= 0{,}5(y_{1\,\text{нз}} + y_{1\,\text{кор}}) + (C + 0{,}5x_{1\,\text{кор}})\alpha, \end{aligned} \tag{9.15}$$

где $C = 0{,}5x_{1\,\text{нз}}$, $D = 0{,}5y_{1\,\text{нз}} - y_J$. Заметим, что x_J и y_J входят в (9.15) несимметрично потому, что $y_J = \text{const}$, а $x_J \geqslant 0$, и при x_J, отличном от нуля, вся подвижная система координат $X_1 O_1 Y_1$ смещается вперед. Таким образом, координаты поверхности языка $x(\varphi)$, $y(\varphi)$, которые входят в ограничения (9.10) и (9.12) с помощью (9.13), (9.14) и (9.15) выражаются через константы y_J, $x_{1\,\text{нз}}$, $y_{1\,\text{нз}}$ и регулируемые координаты корня языка $x_{1\,\text{кор}}$, $y_{1\,\text{кор}}$ кончика языка $x_{1\,\text{кон}}$, $y_{1\,\text{кон}}$, и угла поворота нижней челюсти α.

Итак, мы получили задачу минимизации функции (9.8) при ограничениях (9.10) и (9.12). Как сама функция, так и ограничения являются нелинейными формами от артикуляторных параметров. Следовательно, задача минимизации должна решаться средствами нелинейного программирования. Один из таких способов основан на модификации метода множителей Лагранжа. Как известно, метод множителей Лагранжа позволяет найти экстремум целевой функции $U(x_1, x_2, \ldots, x_n)$ при ограничениях $V_j(x_1, x_2, \ldots, x_m) = 0$, $j = 1, \ldots, m < n$, путем формирования новой функции

$$\Phi(x_1, x_2, \ldots, x_n, \lambda_1, \lambda_2, \ldots, \lambda_m) = U + \sum_{j=1}^{m} \lambda_j V_j$$

и последующего решения системы уравнений

$$\partial W / \partial x_i = 0, \quad i = 1, \ldots, n,$$
$$\partial W / \partial \lambda_j = 0, \quad j = 1, \ldots, m.$$

В классическом методе множителей Лагранжа на переменные x_i не накладывается никаких ограничений. Это и дает возможность найти глобальный максимум целевой функции путем анализа частных производных по x_i и λ_j. В нашей задаче все артикуляторные параметры заключены в границы возможного существования типа $x_{\min} \leqslant x \leqslant x_{\max}$. В этом случае,

как мы видели на примерах с движениями губ и нижней челюсти, минимум работы может достигаться не только в середине области допустимых значений параметров, но и на их границах. Это обстоятельство существенно усложняет решение задачи, однако и на этот случай имеется обобщение метода множителей Лагранжа.

Произведем замену переменных на $\bar{x}$ таким образом, чтобы $\bar{x} \geqslant 0$, т. е. $\bar{x} = x - x_{\min}$. Это дает дополнительные ограничения

$$g_m(\bar{x}) = x_{\max} - x_{\min} - \bar{x} \geqslant 0. \tag{9.16}$$

Обозначив

$$g_j = [x_j - x(\varphi)]^2 + [y_j - y(\varphi)]^2,$$

получим функцию Лагранжа

$$\Phi = W + \lambda_1 (r_1 - g_1) + \sum_{j=2}^{J} \lambda_j (r_j^2 - g_j) + \sum_{m=J+1}^{J+M} \lambda_m g_m, \tag{9.17}$$

где J — равно числу контролируемых сечений (мы приняли $J = 7$), а M — число управляемых артикуляторных параметров ($M = 13$), g_m определяется по (9.16).

Формально, для сохранения единообразия, произведем замену переменных для угла поворота нижней челюсти α и горизонтального смещения точки ее вращения x_y, хотя $\alpha_{\min} = 0$ и $x_{J\min} = 0$. Для поиска экстремума целевой функции Φ в нелинейном программировании пользуются теоремой Куна — Таккера, утверждающей, что этот экстремум достигается в точке (X^*, λ^*), которая является глобальной седловой точкой функции $\Phi(X, \lambda)$, т. е.

$$\Phi(X, \lambda^*) \leqslant \Phi(X^*, \lambda^*) \leqslant \Phi(X^*, \lambda).$$

При этом должны удовлетворяться следующие условия:

$$x_i^* \left[\frac{\partial \Phi(X, \lambda)}{\partial x_i} \right] \Bigg|_{x_i = x_i^*} = 0, \tag{9.18}$$

$$\frac{\partial \Phi(X, \lambda)}{\partial x_i} \Bigg|_{x_i = x_i^*} \leqslant 0, \quad \text{при} \quad x \geqslant 0, \tag{9.19}$$

$$\lambda_i^* \left[\frac{\partial \Phi(X, \lambda)}{\partial \lambda_i} \right] \Bigg|_{\lambda_i = \lambda_i^*} = 0, \tag{9.20}$$

для ограничений типа (9.10)

$$\frac{\partial \Phi(X, \lambda)}{\partial \lambda_i} \Bigg|_{\lambda_i = \lambda_i^*} = 0, \tag{9.21}$$

для ограничений типа (9.12)

$$\frac{\partial \Phi(X, \lambda)}{\partial \lambda_i} \Bigg|_{\lambda_i = \lambda_i^*} \leqslant 0, \quad \text{при} \quad \lambda_i^* \leqslant 0. \tag{9.22}$$

Выпишем теперь частные производные по всем артикуляторным параметрам и множителям Лагранжа:

$$\frac{\partial \Phi}{\partial \bar{x}_{1\,\text{кор}}} = 2c_{\text{кор}}\Delta\bar{x}_{1\,\text{кор}} - \sum_{j=2}^{J}\lambda_j\frac{\partial g_j}{\partial \bar{x}_{1\,\text{кор}}} - \lambda_{J+1},$$

$$\frac{\partial \Phi}{\partial \bar{y}_{1\,\text{кор}}} = 2c_{\text{кор}}\Delta\bar{y}_{1\,\text{кор}} - \sum_{j=2}^{J}\lambda_j\frac{\partial g_j}{\partial \bar{y}_{1\,\text{кор}}} - \lambda_{J+2},$$

$$\frac{\partial \Phi}{\partial \bar{x}_{1\,\text{кон}}} = 2c_{\text{кон}}\Delta\bar{x}_{1\,\text{кон}} - \sum_{j=2}^{J}\lambda_j\frac{\partial g_j}{\partial \bar{x}_{1\,\text{кон}}} - \lambda_{J+3},$$

$$\frac{\partial \Phi}{\partial \bar{y}_{1\,\text{кон}}} = 2c_{\text{кон}}\Delta\bar{y}_{1\,\text{кон}} - \sum_{j=2}^{J}\lambda_j\frac{\partial g_j}{\partial \bar{y}_{1\,\text{кон}}} - \lambda_{J+4},$$

$$\frac{\partial \Phi}{\partial \bar{y}_{\text{вг}}} = 2c_{\text{вг}}\Delta\bar{y}_{\text{вг}} - \lambda_1 - \lambda_{J+5},$$

$$\frac{\partial \Phi}{\partial \bar{y}_{\text{нг}}} = 2c_{\text{нг}}\Delta\bar{y}_{\text{нг}} + \lambda_1 - \lambda_{J+6},$$

$$\frac{\partial \Phi}{\partial \bar{\alpha}} = 2c_{\alpha}\Delta\bar{\alpha} + \lambda_1\bar{x}_J - \sum_{j=2}^{J}\lambda_j\frac{\partial g_j}{\partial \bar{\alpha}} - \lambda_{J+7},$$

$$\frac{\partial \Phi}{\partial \bar{x}_J} = 2c_J\Delta\bar{x}_J + \lambda_1\bar{\alpha} - \sum_{j=2}^{J}\lambda_j\frac{\partial g_j}{\partial \bar{x}_J} - \lambda_{J+8},$$

$$\frac{\partial \Phi}{\partial \bar{k}_n} = 2c_n\Delta\bar{k}_n - \sum_{j=2}^{J}\frac{\partial g_j}{\partial \bar{k}_n} - \lambda_{J+8+n}, \quad n=1,\ \ldots,\ 5,$$

$$\frac{\partial \Phi}{\partial \lambda_j} = r_j^2 - g_j, \quad j=1,\ \ldots,\ J,$$

$$\frac{\partial \Phi}{\partial \lambda_{J+1}} = x_{1\,\text{кор max}} - x_{1\,\text{кор min}} - \bar{x}_{1\,\text{кор}}, \tag{9.23}$$

$$\frac{\partial \Phi}{\partial \lambda_{J+2}} = y_{1\,\text{кор max}} - y_{1\,\text{кор min}} - \bar{y}_{1\,\text{кор}},$$

$$\frac{\partial \Phi}{\partial \lambda_{J+3}} = x_{1\,\text{кон max}} - x_{1\,\text{кон min}} - \bar{x}_{1\,\text{кон}},$$

$$\frac{\partial \Phi}{\partial \lambda_{J+4}} = y_{1\,\text{кон max}} - y_{1\,\text{кон min}} - \bar{y}_{1\,\text{кон}},$$

$$\frac{\partial \Phi}{\partial \lambda_{J+5}} = y_{1\,\text{вг max}} - y_{1\,\text{вг min}} - \bar{y}_{1\,\text{вг}},$$

$$\frac{\partial \Phi}{\partial \lambda_{J+6}} = y_{1\,\text{нг max}} - y_{1\,\text{нг min}} - \bar{y}_{1\,\text{нг}},$$

$$\frac{\partial \Phi}{\partial \lambda_{J+n}} = k_{n\,\max} - k_{n\,\min} - \bar{k}_n, \quad n=1,\ \ldots,\ 5.$$

Здесь Δ означает разность между целевым и начальным значениями координаты, например, $\Delta \bar{x}_{1\,\text{кор}} = x^{(T)}_{1\,\text{кор}} - x^{(0)}_{1\,\text{кор}}$. Знак плюс перед первым членом производных по артикуляторным параметрам означает, что осуществляется поиск минимума.

Система (9.23) содержит $2M+J=33$ неизвестных, из которых 13 являются искомыми артикуляторными параметрами. Заметим, однако, что нам никогда не потребуется решать систему уравнений со всеми 33 неизвестными. Если начать поиск решения с предположения, что минимум находится внутри всех областей определения артикуляторных параметров, то множители Лагранжа с индексами, бо́льшими J, должны обращаться в нуль, поскольку из условий (9.20) имеем, что при $\partial\Phi/\partial\lambda_{J+m}=0$, соответствующий параметр находится на границе минимального значения. Например, из $\partial\Phi/\partial\lambda_{J+1}=0$ следует $x_{1\,\text{кор max}} - x_{1\,\text{кор min}} - \bar{x}_{1\,\text{кор}} = 0$, или, возвращаясь к старой переменной $x_{1\,\text{кор}} = x_{1\,\text{кор max}}$. С другой стороны, условие $\partial\Phi/\partial\lambda_{J+M}=0$ несовместимо с условием $\bar{x}_i^*=0$, так как при этом $x_i = x_{i\,\min}$, а точка оптимума не может одновременно принадлежать и максимальному, и минимальному значению одной и той же переменной. Поэтому при поиске оптимума внутри области определения переменной мы должны решать задачу для $J+M$ неизвестных:

$$\begin{aligned} \partial\Phi/\partial\bar{x}_i &= 0, \quad i=1, \ldots, M, \\ \lambda_j^*\left[\partial\Phi/\partial\lambda_j\right] &= 0, \quad j=1, \ldots, J, \\ \lambda_{J+m} &= 0, \quad m=1, \ldots, M. \end{aligned}$$

Если оптимум внутри области определения переменных не найден, то нужно осуществить перебор по ограничениям, учитывая тот факт, что если $\bar{x}_i^*=0$, то и $\lambda_{J+i}=0$, а если $\partial\Phi/\partial\lambda_{J+M}=0$, то и $\partial\Phi/\partial x_m=0$. В тех случаях, когда

$$\partial\Phi/\partial\bar{y}_{\text{вг}}=0 \quad \text{и} \quad \partial\Phi/\partial\bar{y}_{\text{нг}}=0,$$

можно уменьшить число искомых переменных, выразив λ_1 и $\Delta\bar{y}_{\text{вг}}$ через $\Delta\bar{y}_{\text{нг}}$:

$$\lambda_1 = -2c_{\text{вг}}\Delta\bar{y}_{\text{вг}} - \lambda_{J+5},$$

$$\lambda_1 = 2c_{\text{нг}}\Delta\bar{y}_{\text{нг}} + \lambda_{J+6},$$

откуда

$$\Delta\bar{y}_{\text{вг}} = -\frac{c_{\text{н}}}{c_{\text{вг}}}\Delta\bar{y}_{\text{нг}} - \frac{\lambda_{J+5}+\lambda_{J+6}}{2c_{\text{вг}}}.$$

Частные производные от функции g по артикуляторным параметрам необходимо вычислять не только в методе множителей Лагранжа, но и в других методах поиска экстремума. К счастью, в нашем случае эти производные можно найти аналитически, что снимает проблему определения величины приращения по переменным, которая возникает при численных

способах. Выпишем эти производные:

$$\frac{\partial g_j}{\partial \bar{x}_{1\,\text{кор}}} = -2\left\{\left[x_j - x(\varphi)\right]\frac{\partial x(\varphi)}{\partial \bar{x}_{1\,\text{кор}}} + \left[y_j - y(\varphi)\right]\frac{\partial y(\varphi)}{\partial \bar{x}_{1\,\text{кор}}}\right\},$$

где

$$\frac{\partial x(\varphi)}{\partial \bar{x}_{1\,\text{кор}}} = 0{,}5 + \frac{1}{r_{jp}^2}\left\{r_{jp}\left[\left(x_j - x_p\right)\frac{\partial s}{\partial \bar{x}_{1\,\text{кор}}} - 0{,}5s\right] - \frac{\partial r_{jp}}{\partial \bar{x}_{1\,\text{кор}}}\left(x_j - x_p\right)s\right\},$$

$$\frac{\partial y(\varphi)}{\partial \bar{x}_{1\,\text{кор}}} = 0{,}5\alpha + \frac{1}{r_{jp}^2}\left\{r_{jp}\left[\left(y_j - y_p\right)\frac{\partial s}{\partial \bar{x}_{1\,\text{кор}}} - 0{,}5\alpha s\right] - \frac{\partial r_{jp}}{\partial \bar{x}_{1\,\text{кор}}}\left(y_j - y_p\right)s\right\},$$

$$\frac{\partial r_{jp}}{\partial \bar{x}_{1\,\text{кор}}} = -\frac{x_j - x_p + (y_j - y_p)\alpha}{2r_{jp}},$$

$$\frac{\partial s}{\partial \bar{x}_{1\,\text{кор}}} = -s'\left[0{,}5\,\frac{(x_j - x_p)\alpha - (y_j - y_p)}{r_{jp}^2} + \frac{y_{1\,\text{кон}} + y_{1\,\text{кон min}} - (y_{1\,\text{кор}} + y_{1\,\text{кор min}})}{r_{TR}^2}\right].$$

Здесь s' — производная от суммы собственных функций языка $\psi(\varphi)$ по аргументу φ,

$$r_{\text{кк}}^2 = \left(x_{1\,\text{кон}} + x_{1\,\text{кон min}} - x_{1\,\text{кор}} - x_{1\,\text{кор min}}\right)^2 + \\ + \left(y_{1\,\text{кон}} + y_{1\,\text{кон min}} - y_{1\,\text{кон}} - y_{1\,\text{кон min}}\right)^2,$$

а r_{jp} — определяется по (9.13).

Аналогично находим производные по $\bar{y}_{1\,\text{кор}}$:

$$\frac{\partial g_j}{\partial \bar{y}_{1\,\text{кор}}} = -2\left\{\left[x_j - x(\varphi)\right]\frac{\partial x(\varphi)}{\partial \bar{y}_{1\,\text{кор}}} + \left[y_j - y(\varphi)\right]\frac{\partial y(\varphi)}{\partial \bar{y}_{1\,\text{кор}}}\right\},$$

где

$$\frac{\partial x(\varphi)}{\partial \bar{y}_{1\,\text{кор}}} = 0{,}5\alpha + \frac{1}{r_{jp}^2}\left\{r_{jp}\left[\left(x_j - x_p\right)\frac{\partial s}{\partial \bar{y}_{1\,\text{кор}}} - 0{,}5\alpha s\right] - \frac{\partial r_{jp}}{\partial \bar{y}_{1\,\text{кор}}}\left(x_j - x_p\right)s\right\},$$

$$\frac{\partial y(\varphi)}{\partial \bar{y}_{1\,\text{кор}}} = 0{,}5 + \frac{1}{r_{jp}^2}\left\{r_{jp}\left[\left(y_j - y_p\right)\frac{\partial s}{\partial \bar{y}_{1\,\text{кор}}} - 0{,}5s\right] - \frac{\partial r_{jp}}{\partial \bar{y}_{1\,\text{кор}}}\left(y_j - y_p\right)s\right\},$$

$$\frac{\partial r_{jp}}{\partial \bar{y}_{1\,\text{кор}}} = \frac{(x_j - x_p)\alpha + y_j - y_p}{2r_{jp}},$$

$$\frac{\partial s}{\partial \bar{y}_{1\,\text{кор}}} = -s'\left[0{,}5\,\frac{x_j - x_p - (y_j - y_p)\alpha}{r_{jp}^2} + \frac{x_{1\,\text{кон}} + x_{1\,\text{кон min}} - (x_{1\,\text{кор}} + x_{1\,\text{кор min}})}{r_{\text{кк}}^2}\right].$$

Производные от g по $\bar{x}_{1\,\text{кон}}$ и $\bar{y}_{1\,\text{кон}}$ выглядят проще, поскольку ряд членов обращается в нуль, будучи независимыми от этих переменных:

$$\frac{\partial g_j}{\partial \bar{x}_{1\,\text{кон}}} = -2\left\{\left[x_j - x(\varphi)\right]\frac{\partial x(\varphi)}{\partial \bar{x}_{1\,\text{кон}}} + \left[y_j - y(\varphi)\right]\frac{\partial y(\varphi)}{\partial \bar{x}_{1\,\text{кон}}}\right\},$$

$$\frac{\partial g_j}{\partial \bar{y}_{1\,\text{кон}}} = 2\left\{\left[x_j - x(\varphi)\right]\frac{\partial x(\varphi)}{\partial \bar{y}_{1\,\text{кон}}} + \left[y_j - y(\varphi)\right]\frac{\partial y(\varphi)}{\partial \bar{y}_{1\,\text{кон}}}\right\},$$

где

$$\frac{\partial x(\varphi)}{\partial \bar{x}_{1\,\text{кон}}}=\frac{x_j-x_p}{r_{jp}}\frac{\partial s}{\partial \bar{x}_{1\,\text{кон}}},\quad \frac{\partial x(\varphi)}{\partial \bar{y}_{1\,\text{кон}}}=\frac{x_j-x_p}{r_{jp}}\frac{\partial s}{\partial \bar{y}_{1\,\text{кон}}},$$

$$\frac{\partial y(\varphi)}{\partial \bar{x}_{1\,\text{кон}}}=\frac{y_j-y_p}{r_{jp}}\frac{\partial s}{\partial \bar{x}_{1\,\text{кон}}},\quad \frac{\partial y(\varphi)}{\partial \bar{y}_{1\,\text{кон}}}=\frac{y_j-y_p}{r_{jp}}\frac{\partial s}{\partial \bar{y}_{1\,\text{кон}}},$$

и

$$\frac{\partial s}{\partial \bar{x}_{1\,\text{кон}}}=-s'\frac{y_{1\,\text{кон}}+y_{1\,\text{кон min}}-(y_{1\,\text{кор}}+y_{1\,\text{кор min}})}{r_{\text{кк}}^2},$$

$$\frac{\partial s}{\partial \bar{y}_{1\,\text{кон}}}=s'\frac{x_{1\,\text{кон}}+x_{1\,\text{кон min}}-(x_{1\,\text{кор}}+x_{1\,\text{кор min}})}{r_{\text{кк}}^2}.$$

Производная от g по углу поворота нижней челюсти $\bar{\alpha}$ есть

$$\frac{\partial g_j}{\partial \bar{\alpha}}=-2\left\{[x_j-x(\varphi)]\frac{\partial x(\varphi)}{\partial \bar{\alpha}}+[y_j-y(\varphi)]\frac{\partial y(\varphi)}{\partial \bar{\alpha}}\right\},$$

где

$$\frac{\partial x(\varphi)}{\partial \bar{\alpha}}=\frac{\partial x_p}{\partial \bar{\alpha}}+\frac{1}{r_{jp}^2}\left\{r_{jp}\left[(x_j-x_p)\frac{\partial s}{\partial \bar{\alpha}}-\frac{\partial x_p}{\partial \bar{\alpha}}s\right]-\frac{\partial r_{jp}}{\partial \bar{\alpha}}(x_j-x_p)s\right\},$$

$$\frac{\partial y(\varphi)}{\partial \bar{\alpha}}=\frac{\partial y_p}{\partial \bar{\alpha}}+\frac{1}{r_{jp}^2}\left\{r_{jp}\left[(y_j-y_p)\frac{\partial s}{\partial \bar{\alpha}}-\frac{\partial x_p}{\partial \bar{\alpha}}s\right]-\frac{\partial r_{jp}}{\partial \bar{\alpha}}(y_j-y_p)s\right\},$$

и

$$\frac{\partial s}{\partial \bar{\alpha}}=-\frac{s'}{r_{jp}^2}\left[(x_j-x_p)\frac{\partial y_p}{\partial \bar{\alpha}}-(y_j-y_p)\frac{\partial x_p}{\partial \bar{\alpha}}\right],$$

$$\frac{\partial r_{jp}}{\partial \bar{\alpha}}=-\frac{1}{r_{jp}}\left[(x_j-x_p)\frac{\partial x_p}{\partial \bar{\alpha}}+(y_j-y_p)\frac{\partial y_p}{\partial \bar{\alpha}}\right],$$

причем

$$\frac{\partial x_p}{\partial \bar{\alpha}}=D+0{,}5(y_{1\,\text{кор}}+y_{1\,\text{кор min}}),$$

$$\frac{\partial y_p}{\partial \bar{\alpha}}=C+0{,}5(x_{1\,\text{кор}}+x_{1\,\text{кор min}}),$$

а C и D были определены в (9.15).

Производная от g по горизонтальному смещению точки вращения нижней челюсти x_J есть

$$\frac{\partial g_j}{\partial x_J}=-2\left\{[x_j-x(\varphi)]\frac{\partial x(\varphi)}{\partial \bar{x}_J}+[y_j-y(\varphi)]\frac{\partial y(\varphi)}{\partial x_J}\right\},$$

где

$$\frac{\partial x(\varphi)}{\partial \bar{x}_j}=\frac{1}{r_{jp}^2}\left\{r_{jp}\left[(x_j-x_p)\frac{\partial s}{\partial \bar{x}_j}-s\right]+\frac{(x_j-x_p)^2}{r_{jp}}s\right\}+1,$$

$$\frac{\partial y(\varphi)}{\partial \bar{x}_j}=\frac{1}{r_{jp}^2}\left[r_{jp}(y_j-y_p)\frac{\partial s}{\partial \bar{x}_j}+\frac{(x_j-x_p)(y_j-y_p)}{r_{jp}}s\right],$$

$$\frac{\partial s}{\partial \bar{x}_y}=-s'\frac{y_j-y_p}{r_{jp}^2}.$$

И, наконец, получим производные от g по коэффициентам k_n при собственных функциях языка. В силу независимости многих переменных от k_n эти производные выглядят весьма просто:

$$\frac{\partial g_i}{\partial k_n}=-2\frac{\psi_n}{r_{jp}^2}\left\{(x_j-x_p)[x_j-x(\varphi)]+(y_j-y_p)[y_j-y(\varphi)]\right\}.$$

Таким образом, мы определили все компоненты, необходимые для решения систем (9.18)—(9.22) при поиске минимума функции (9.17). В общем виде условия $\partial\Phi/\partial x_i=0$, $\partial\Phi/\partial\lambda_j=0$ могут быть представлены как нелинейная система уравнений, например, с $2m$ неизвестными:

$$\begin{aligned}&f_1(x_1,\ \ldots,\ x_m,\ \lambda_1,\ \ldots,\ \lambda_m)=0,\\&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\&f_{2M}(x_1,\ \ldots,\ x_m,\ \lambda_1,\ \ldots,\ \lambda_m)=0.\end{aligned}\qquad(9.24)$$

Решение такой системы возможно различными способами, в том числе и методом Ньютона, методом итераций, методом скорейшего спуска и т. д. Например, в методе Ньютона, обозначив, как и ранее, через $X=\{x_1,\ \ldots,\ x_m,\ \lambda_1,\ \ldots,\ \lambda_m\}$ вектор неизвестных величин, через $F=\{f_1,\ \ldots,\ f_{2m}\}$ — вектор функций, формируют матрицу Якоби системы (9.24), т. е.

$$G(x)=\begin{vmatrix}\frac{\partial f_1}{\partial x_1}\cdots\frac{\partial f_1}{\partial x_m} & \frac{\partial f_1}{\partial\lambda_1}\cdots\frac{\partial f_1}{\partial\lambda_m}\\ \frac{\partial f_2}{\partial x_1}\cdots\frac{\partial f_2}{\partial x_m} & \frac{\partial f_2}{\partial\lambda_1}\cdots\frac{\partial f_2}{\partial\lambda_m}\\ \ldots\ldots\ldots & \ldots\ldots\ldots\\ \frac{\partial f_{2m}}{\partial x_1}\cdots\frac{\partial f_{2m}}{\partial x_m} & \frac{\partial f_{2m}}{\partial\lambda_1}\cdots\frac{\partial f_{2m}}{\partial\lambda_m}\end{vmatrix},$$

а затем находят обратную ей матрицу $G^{-1}(X)$. Значения

неизвестных на $(p+1)$-й итерации вычисляются как

$$X^{(p+1)}=X^{(p)}-G^{-1}\left[X^{(p)}\right]f\left[X^{(p)}\right].$$

Обратная матрица $G^{-1}(X)$ получается из матрицы $G(X)$ путем замены каждого ее элемента на его алгебраическое дополнение (минор со знаком), деленное на определитель матрицы $G(X)$. В модифицированном методе Ньютона обратная матрица $G^{-1}(X)$ вычисляется лишь один раз. В матрицу $G(X)$ войдут вторые производные от g_m по неизвестным параметрам и множителям Лагранжа. Эти производные также получаются аналитически, но ввиду громоздкости их выражений, мы их опустим.

Решение системы уравнений такой высокой размерности требует большого числа вычислений. Количество вычислений можно уменьшить, если разбить процесс решения на несколько этапов, как это делается, например, в методе динамического программирования. В нашей задаче имеется естественный параметр, в функции которого изменяются все артикуляторные параметры. Этот параметр — время. Учитывая задержку сигналов обратной связи, целесообразно разделить переходный процесс из одного артикуляторного состояния в другое на интервалы длительностью $\Delta T=20$—40 мс. Тогда, в соответствии со схемой поиска оптимального пути в методе динамического программирования, через каждые ΔT будем выполнять следующие операции. На каждом интервале времени сначала оцениваем положение каждого артикуляторного органа $\xi_j^{(0)}$ относительно границ диапазона его величин и находим наибольшее положительное и отрицательное приращения:

$$\Delta\xi_j^{+}=\xi_{j\max}-\xi_j^{(0)},$$

$$\Delta\xi_j^{-}=\xi_j^{(0)}-\xi_{j\min}.$$

Величины $\Delta\xi_j^{+}$ и $\Delta\xi_j^{-}$ характеризуют наибольшие положительные и отрицательные усилия для изменения $\xi_j^{(0)}$. Очевидно, реально необходимое усилие находится внутри этого диапазона:

$$c_j\Delta\xi_j^{+}\leqslant F_j\leqslant c_j\Delta\xi_j^{-},$$

при условии статической задачи, т. е. неподвижного положения соответствующего параметра $\xi_j^{(0)}$ в начальный момент времени. Проквантуем диапазон возможных усилий $F_j^{+}=c_j\Delta\xi_j^{+}$ и $F_j^{-}=c_j\xi_j^{-}$ на N уровней и вычислим возможное смещение каждого артикуляторного параметра ξ_j для каждого из квантованных усилий F_j. Заметим, что поскольку движение каждого артикуляторного параметра подчиняется решению обыкновенного дифференциального уравнения второго порядка с разными коэффициентами, то и изменение величин ξ_j через интервал времени ΔT будет различным. Для того чтобы обеспечить движение артикуляторных параметров к требуемым положениям, воспользуемся ограничениями (9.10) и (9.12) так, чтобы,

используя часть артикуляторных параметров в качестве свободных переменных, остальные параметры вычислять в соответствии с ограничениями. Для избежания трудностей, связанных с ограничениями типа неравенства (9.12), будем считать их на первом шаге строгими равенствами. Тогда имеем J нелинейных уравнений (в нашем случае $J=7$) и, задав J артикуляторных параметров, остальные $M-J$ параметров найдем путем решения этой системы уравнений. Таким образом, вместо однократного решения системы размерности $2M$ и перебора по граничным условиям в методе Лагранжа решаем задачу почти втрое меньшей размерности. Задавая на каждом интервале времени множество усилий $\{F_j\}$, вычисляем значения J артикуляторных параметров как результат решения дифференциальных уравнений второго порядка, а оставшиеся $M-J$ параметров находим через решение системы нелинейных уравнений, составленных из ограничений типа (9.10). На каждом шаге ΔT для всех ξ_j и всех квантованных значений F_j вычисляется работа $\Delta\xi_j F_j$, равная сумме работ по всей последовательности предыдущих состояний, приведших к данному состоянию. Среди множества всех вычисленных значений определяется минимальная величина и соответствующие ей значения усилий F_j и артикуляторных параметров. Процесс заканчивается либо тогда, когда исчерпывается отведенное на него время, либо когда разница между достигнутыми расстояниями $r_j^{(T)}$ и требуемыми расстояниями в ограничениях (9.10) станет меньше некоторой заранее заданной величины допустимой погрешности ε, а условия (9.12) будут выполнены. После этого осуществляется выбор $F_j(t)$ как дискретных отсчетов через интервалы ΔT, таких, что для выбранных $F_j(t)$ в каждый момент времени (и, следовательно, на всем переходном процессе) выполняемая работа минимальна. В результате мы получаем не только требуемые значения артикуляторных параметров, но и определяем усилия $F_j(t)$, необходимые для достижения заданных целей в пространстве расстояний в речевом тракте.

§ 9.3. Обратная задача для формы речевого тракта

Рассмотренной выше зависимостью между артикуляторными параметрами и расстояниями в среднесагиттальной плоскости речевого тракта следует пользоваться при оптимизации параметров синтезатора, добиваясь наилучшего звучания путем подбора расстояний в критических сечениях. Существует и другой источник информации о звуках речи — амплитудно-частотные характеристики речевых сигналов. Располагая данными о формантных частотах и спектрах шумных звуков, можно попытаться решить обратную задачу, и найти требуемые площади (а с ними и расстояния) в критических сечениях речевого тракта. В общем случае это очень трудная проблема,

и лишь для некоторых классов уравнений существуют методы решения обратной задачи [53, 62]. Однако в нашем случае имеется дополнительная информация в виде ограничений на артикуляторные параметры и расстояния в речевом тракте, которые позволяют надеяться на получение устойчивого решения.

Можно попытаться найти простые аналитические отношения между формой речевого тракта и формантными частотами. Например, в [140] методами факторного анализа были получены формулы для расчета амплитуд двух главных компонент w_1 и w_2, описывающих форму языка как функции от частот первых трех формант F_1, F_2 и F_3:

$$w_1 = c_1 \frac{F_2}{F_3} + c_2 \frac{F_1}{F_3} + c_3 \frac{F_3}{F_1} + c_4,$$

$$w_2 = c_5 \frac{F_1}{F_2} + c_6 \frac{F_2}{F_1} + c_7 \frac{F_3}{F_1} + c_8,$$

где

$$c_1 = 2{,}309, \quad c_2 = 2{,}105, \quad c_3 = 0{,}117, \quad c_4 = -2{,}446,$$
$$c_5 = 1{,}913, \quad c_6 = -0{,}245, \quad c_7 = 0{,}188, \quad c_8 = 0{,}584.$$

Компоненты w_1 и w_2 описывают степень подъема языка в передней или задней части. Кроме того, расстояние между губами вычисляется как

$$x_{\text{г}} = c_9 F_2 + c_{10} F_2 F_3 + c_{11} \frac{F_1}{F_2} + c_{12},$$

где

$$c_9 = 3 \cdot 10^{-4}, \quad c_{10} = -3{,}43 \cdot 10^{-7}, \quad c_{11} = 4{,}143, \quad c_{12} = -2{,}865.$$

В [190] также использовались две собственные функции языка, но решение получалось путем минимизации функции

$$J(X_k) = \| Y_k - h(X_k) \|^2 + \| X_k \|^2 + \| X_k - X_{k-1} \|^2, \qquad (9.25)$$

где X_k — вектор артикуляторных параметров в k-й момент времени, $h(X_k)$ — расчетные частотные характеристики речевого сигнала, Y_k — измеренная частотная характеристика. Вектор X_k включает в себя две компоненты для языка, ширину гортани, площадь прохода в носовую полость, смещение нижней челюсти, расстояние между губами и их выпячивание — всего шесть параметров. Здесь $\| X_k \|^2$ обозначает квадратическую форму от артикуляторных параметров с некоторыми диагональными взвешивающими матрицами. Первый член в (9.25) минимизирует ошибку в вычисленных и измеренных частотных характеристиках речевого сигнала, в качестве которых используются кепстральные коэффициенты, второй член необходим для исключения необычных артикуляций, а третий сохраняет непрерывность артикуляторных параметров во времени. Этот подход дает возможность непрерывной оценки артикуляторных

параметров, тогда как предыдущий метод применим только к статическим состояниям, точнее, к отдельно произнесенным гласным. И в том, и в другом методе число артикуляторных параметров слишком мало для достижения требуемой точности в управлении формой речевого тракта.

Допустим, что мы имеем некоторое множество функций площади поперечного сечения речевого тракта $\{S_0(x)\}$ и соответствующее им множество собственных функций $\{\psi^{(0)}(x)\}$ и собственных чисел $\{\lambda^{(0)}\}$. Тогда для небольшого изменения некоторой функции $S_0(x)$ методом малых возмущений можно найти новые собственные функции $\psi_i(x)$ и числа λ_i. Пусть возмущение площади сечения тракта есть $S_1(x)$ такое, что $S_1(x) \ll S_0(x)$ при любом x, и новая функция площади $S(x)$ есть $S(x)=S_0(x)+S_1(x)$. Возмущенные значения собственных чисел λ_i представляются как

$$\lambda_i^2=\lambda_i^{(0)2}(1+\eta_i),$$

где

$$\eta_i=\frac{a_i}{\lambda_i^{(0)2}}-b_i-c_i,$$

$$a_i=\int_0^l S_1(x)\left[\psi_i^{(0)\prime}(x)\right]^2 dx,$$

$$b_i=\int_0^l S_1(x)\left[\psi_i^{(0)}(x)\right]^2 dx,$$

$$c_i=S_1(l)\psi_i^{(0)}(l)\psi_i^{(0)\prime}(l)-S_1(0)\psi_i^{(0)}(0)\psi_i^{(0)\prime}(0),$$

l — длина речевого тракта от голосовой щели до губ; штрих означает производную по x. При закрытой голосовой щели можно принять граничные условия как на жесткой стенке и второй член в коэффициенте c_i исчезает.

Ранее мы условились контролировать форму речевого тракта в отдельных сечениях, причем число этих сечений сравнительно невелико — около 7. Поэтому опишем возмущающие функции $S_1(x)$ в виде интерполяционного полинома с узлами в n контролируемых сечениях:

$$S_1(x)=\sum_{j=1}^{n} L_j(x)S_1(x_j),$$

где x_j — координаты контролируемого сечения, а

$$L_j(x)=\frac{(x-x_1)\ldots(x-x_{j-1})(x-x_{j+1})\ldots(x-x_{n+1})}{(x_j-x_1)\ldots(x_j-x_{j-1})(x_j-x_{j+1})\ldots(x_j-x_{n+1})}.$$

Как видно, знаменатель $L_j(x)$ есть число постоянное, поскольку координаты x_j контролируемых сечений фиксированы для речевого тракта заданной длины l. Числитель $L_j(x)$ есть

полином n-й степени, т. е.

$$L_j(x)=\frac{1}{d_j}(e_{1j}+e_{2j}x+\ldots+x^n),$$

где

$$d_j=(x_j-x_1)\ldots(x_j-x_{j-1})(x_j-x_{j+1})\ldots(x_j-x_n),$$
$$e_{1j}=x_1x_2\ldots x_n,$$

а остальные коэффициенты e_{ij} получаются как сумма произведений различного порядка координат x_1, x_2, ..., x_n. Таким образом, коэффициенты a_i могут быть записаны как

$$a_i=S_1(x_1)\int_0^l(e_{11}+e_{21}x+\ldots+x^n)\left[\psi_i^{(0)'}(x)\right]^2dx+\ldots$$
$$\ldots+S_1(x_n)\int_0^l(e_{1n}+e_{2n}x+\ldots+x^n)\left[\psi_i^{(0)'}(x)\right]^2dx.$$

Подынтегральные функции сводятся к произведениям вида $x^j\left[\psi_i^{(0)'}(x)\right]^2$, поэтому их интегрирование может быть выполнено раз и навсегда. При этом удобно перейти к безразмерной координате $\bar{x}=x/l$, так что интегрирование осуществляется на интервале $[0,1]$, и результат не зависит от длины речевого тракта. Аналогично поступаем и с подынтегральными функциями для коэффициентов b_i. В результате для каждого i можно записать линейную форму

$$\eta_i=\frac{1}{\lambda_i^{(0)^2}}\left[g_{a_{i1}}S_1(x_1)+\ldots+g_{a_{i2}}S_1(x_2)+\ldots+g_{a_{in}}S_1(x_n)\right]-$$
$$-\left[g_{b_{i1}}S_1(x_1)+\ldots+g_{b_{in}}S_1(x_n)\right]-c_i,$$

где $g_{a_{ij}}$ и $g_{b_{ij}}$ — определенные интегралы для a_i и b_i, умноженные на коэффициенты e. Группируя члены с одними и теми же отсчетами $S_1(x)$, получаем систему

$$\eta_1=\left(\frac{g_{a_{11}}}{\lambda_1^{(0)^2}}-g_{b_{11}}\right)S_1(x_1)+\left(\frac{g_{a_{12}}}{\lambda_1^{(0)^2}}-g_{b_{12}}\right)S_1(x_2)+\ldots$$
$$\ldots+\left(\frac{g_{a_{1n}}}{\lambda_1^{(0)^2}}-g_{b_{1n}}\right)S_1(x_n)-c_1, \quad (9.26)$$

$$\eta_2=$$
$$=\left(\frac{g_{a_{21}}}{\lambda_2^{(0)^2}}-g_{b_{21}}\right)S_1(x_1)+\left(\frac{g_{a_{22}}}{\lambda_2^{(0)^2}}-g_{b_{22}}\right)S_1(x_2)+\ldots+\left(\frac{g_{a_{2n}}}{\lambda_2^{(0)^2}}-g_{b_{2n}}\right)S_1(x_n)-c_2,$$

.....

$$\eta_m=$$
$$=\left(\frac{g_{a_{m1}}}{\lambda_m^{(0)^2}}-g_{b_{m1}}\right)S_1(x_1)+\left(\frac{g_{a_{m2}}}{\lambda_m^{(0)^2}}-g_{b_{m2}}\right)S_1(x_2)+\ldots+\left(\frac{g_{a_{mn}}}{\lambda_m^{(0)^2}}-g_{b_{mn}}\right)S_1(x_n)-c_m.$$

Эту систему можно использовать различными способами. Если число измеренных собственных функций (формантных частот) m равно числу контрольных сечений n, то возмущения $S_1(x_j)$ находятся непосредственным решением системы (9.26). Точность решения зависит от числа заранее запомненных базовых функций площади $S_0(x)$ — чем их больше, тем меньше оказывается требуемое возмущение, и тем лучше выполняются условия, при которых получена система (9.26). Повышение точности вычислений при этом оплачивается увеличением занимаемого объема памяти. Оценим этот объем при $m=3$, т. е. если число измеренных формантных частот равно 3. Тогда при непосредственном решении системы (9.26) можно найти возмущения площади сечения $S_1(x)$ лишь в трех точках. Допустим, что мы неравномерно (например, по степенному закону) квантуем значения площади в каждом сечении на 10 уровней. Тогда число возможных вариантов конфигураций речевого тракта составляет 10^3. При отсчетах функции площади через каждые 0,5 см для ее записи требуется около 40 чисел. Собственные функции запоминать не надо, а в памяти храним лишь результаты интегрирования — в данном случае по 9 коэффициентов для a и b. В итоге получаем 6×10^4 чисел, т. е. около 120 Кбайт, если на каждое число отводится по 2 байта. Это довольно большой объем памяти, хотя он находится вполне в пределах возможностей существующей и, тем более, будущей электронной технологии.

Другой способ использования системы (9.26) вытекает из необходимости определить требуемые возмущения во всех контролируемых сечениях. Их, как мы условились, около 7, тогда как число надежно измеряемых формантных частот редко превышает 3. В этом случае система (9.26) дает дополнительные ограничения для расстояний r_j в схеме вариационной задачи управления артикуляцией, рассмотренной выше. При этом переход от $S_1(x_j)$ к r_j осуществляется с помощью алгоритма, связывающего расстояния и площади поперечного сечения речевого тракта, описанного в § 9.2.

Если в процессе решения вариационной задачи окажется, что нет необходимости изменять расстояния во всех n сечениях, а можно ограничиться изменением в двух — трех точках, и эти точки будут определены, то дальнейшие вычисления выполняются путем непосредственного решения системы (9.26), как это было описано выше. Иными словами, проблема сводится к определению таких m сечений, возмущения в которых давали бы наилучшее приближение вычисленных формантных частот к измеренным. Можно приближенно определить эти сечения, если воспользоваться свойствами δ-функции, предположив, что возмущение функции площади есть

$$S_1(x_j)=\varepsilon_j\delta(x-\ldots_j).$$

Тогда поправки на собственные числа определяются как

$$\eta_i = \frac{1}{\lambda_i^{(0)2}} \varepsilon_j \left\{ \left[\psi_i^{(0)'}(x_j) \right]^2 - \left[\lambda_i^{(0)} \psi_i^{(0)}(x_j) \right]^2 \right\} - c_i.$$

Отсюда можно найти такое сечение, изменение площади которого в наибольшей степени приблизит искомое собственное число λ_i к наблюдаемому. При этом воспользуемся тем свойством, что сужение речевого тракта вблизи узла i-й собственной функции повышает ее собственное число, а при сужении в области ее пучности собственное число уменьшается. Определив знак искомого приращения η_i из

$$\eta_i = \frac{\lambda_i^2}{\lambda_i^{(0)2}} - 1,$$

найдем и сечения, в которых необходимо произвести возмущение.

Можно определить сечения с наибольшим влиянием на собственные числа и более точно, положив

$$S_1(x) = \sum_{j=1}^{m} \varepsilon_j \delta(x - x_j).$$

В этом случае задача сводится к решению системы линейных уравнений, но требуется перебор вариантов. Для определения наилучшего набора сечений в памяти необходимо хранить, помимо эталонных площадей и значений определенных интегралов, еще и координаты узлов и пучностей собственных функций, т. е. таких точек, где либо $\psi_i = 0$, либо $\psi_i' = 0$. Выбор площади нулевого приближения $S_0(x)$ осуществляется путем поиска таких $\lambda_i^{(0)}$, которые оказались бы ближайшими в некотором пространстве к измеренным собственным числам λ_i. Вопрос о мере сходства сигналов с несколькими периодическими компонентами относится к области слухового восприятия. Учитывая нелинейную деформацию оси частот в слуховой зоне восприятия, степень близости измеренной формантной частоты F_i и вычисленной или хранящейся в памяти синтезатора $F_i^{(0)}$, может быть определена как

$$\rho(F_i, F_i^{(0)}) = \lg F_i - \lg F_i^{(0)},$$

а для m формант, как

$$\rho = \sum_{i=1}^{m} \lg \frac{F_i}{F_i^{(0)}}.$$

В методе возмущений сравнительная простота вычисления сопровождается необходимостью хранить в памяти большое число количества информации. Этого можно избежать, несколько увеличив объем вычислений. Такую возможность предоставляет метод Галеркина [59]. Напомним, что в методе Галеркина собственные функции $\psi_k(x)$ для речевого тракта

с переменной площадью сечения раскладываются в ряд по собственным функциям $\psi_i^{(0)}(x)$ однородной акустической системы (с постоянной площадью сечения) с теми же граничными условиями:

$$\psi_k(x)=\sum_{i=1}^{I} a_i\psi_i^{(0)}(x).$$

Неизвестные коэффициенты a_i и собственные числа λ_i определяются путем решения линейной системы

$$\begin{aligned}
&a_1(A_{11}+\lambda^2 B_{11})+a_2(A_{12}+\lambda^2 B_{12})+\ldots+a_I(A_{1I}+\lambda^2 B_{1I})=0,\\
&a_1(A_{21}+\lambda^2 B_{21})+a_2(A_{22}+\lambda^2 B_{22})+\ldots+a_I(A_{2I}+\lambda^2 B_{2I})=0,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&a_1(A_{m1}+\lambda^2 B_{m1})+a_2(A_{m2}+\lambda^2 B_{m2})+\ldots+a_I(A_{mI}+\lambda^2 B_{mI})=0.
\end{aligned}$$

Возможности этого подхода, в общем, такие же, как и метода малых возмущений.

Для синтезаторов решение обратной задачи связано с необходимостью определения артикуляторных параметров, обеспечивающих заданные акустические характеристики звуков речи. Если удастся найти алгоритм устойчивого решения обратной задачи, то проблема формирования управляющих команд для артикуляторного синтезатора без использования рентгенографических измерений будет в значительной степени решена. Рассмотрим результаты одного исследования, посвященного восстановлению артикуляторных параметров для гласных звуков русского языка. В этом исследовании речевой сигнал записывался одновременно с измерениями координат 12 точек на поверхности языка посредством микролучевого рентгеноскопа лаборатории управления движением университета штата Висконсин, США. Все гласные звуки помещались в трехсложные звукосочетания с одним и тем же согласным /Б/, например, БАБАБА, БОБОБО, БУБУБУ и т. д., причем последний гласный тянулся, и в это время осуществлялась запись координат поверхности языка. Формантные частоты гласных определялись с помощью алгоритма линейного предсказания на 14 коэффициентов в Кайзеровом окне длительностью 10 мс.

В алгоритме решения обратной задачи использовалась кинематическая модель речевого тракта, описанная в гл. 8. Затем вычислялась площадь поперечного сечения речевого тракта, и с помощью алгоритма пристрелки (см. § 7.4) вычислялись резонансные частоты. В процессе решения артикуляторные параметры варьировались таким образом, чтобы достичь минимума критерия оптимальности Ф:

$$\Phi=\sum_{i=1}^{14} b_i\xi_i^2+\alpha_j c_k\rho,$$

где ξ_i — изменение того или иного артикуляторного параметра, b_i — упругое сопротивление этому изменению, ρ — расстояние

между вычисленными резонансными частотами F_k и измеренными частотами формант F_k^* в равномерной метрике:

$$\rho = \max_k |F_k - F_k^*|, \quad k = 1, \ 2, \ 3, \ 4.$$

Первый член критерия оптимальности Ф есть, как видно, работа, совершаемая при изменении артикуляторных параметров. Коэффициент $c_k = 1$, если $F_k^* < 1000$ Гц, и $c_k = 0{,}5$ при $F_k^* \geqslant 1000$ Гц. Таким способом учитывается перцептивно более важная роль нижних формант. Член с ρ на самом деле является ограничением (требуется $\rho = 0$), но, согласно методу штрафных функций, он вводится в критерий оптимальности, причем коэффициент α_j меняется в процессе оптимизации от малых до больших величин (j — номер итерации). Алгоритм оптимизации использует покоординатный спуск, а поиск

Таблица 9.1. Вычисленные и измеренные формантные частоты

Гласный	А					О				
Частота форманг	Измерено	Вычислено				Измерено	Вычислено			
Число форманг		1	2	3	4		1	2	3	4
F_1	686	686	686	668	668	518	319	352	352	317
F_2	1286	1310	1286	1310	1300	871	1230	1040	1040	1167
F_3	2200	2110	2110	2170	2170	1960	2010	1810	1810	1890
F_4	2670	2600	2600	2700	2700	2850	2510	2310	2310	2430
Гласный	У					И				
Частота форманг	Измерено	Вычислено				Измерено	Вычислено			
Число форманг		1	2	3	4		1	2	3	4
F_1	360	360	420	420	360	257	257	325	369	242
F_2	794	1150	1056	1056	864	1960	1160	2087	1980	1710
F_3	1810	2010	2010	1826	1830	2670	1840	—	2380	2600
F_4	2900	2540	2520	2310	2920	3360	2430	—	2700	3150
Гласный	Ы					Э				
Частота форманг	Измерено	Вычислено				Измерено	Вычислено			
Число форманг		1	2	3	4		1	2	3	4
F_1	273	273	268	238	369	530	530	546	530	573
F_2	1770	1320	1610	1610	1605	1590	1320	1560	1590	1700
F_3	2140	2000	2000	2180	2330	2320	2000	2000	2310	2250
F_4	3280	2550	2520	2540	3076	3380	2550	2520	2530	3280

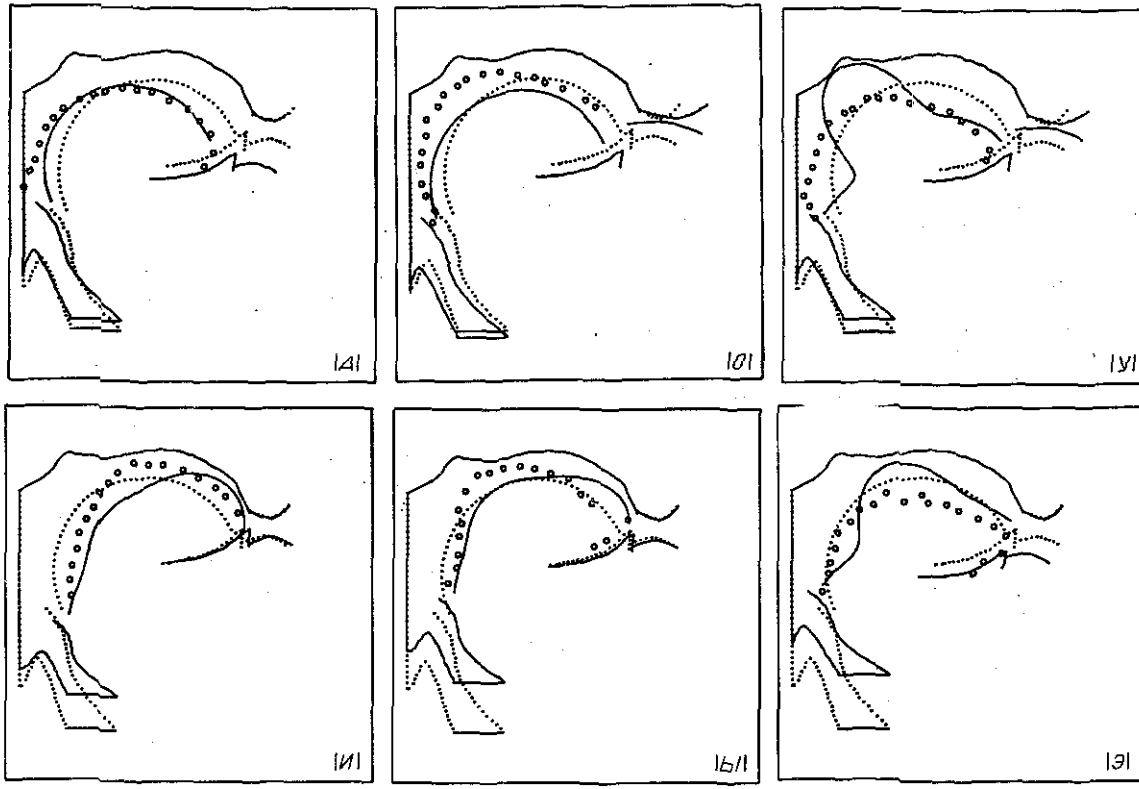


Рис. 9.4. Измеренная (ооо) и вычисленная (——) формы речевого тракта для 4-х опорных формант, (···)— *нейтральная позиция*

минимума по каждому параметру осуществляется методом золотого сечения (см. § 11.2). Недостатки этих методов известны, но эти алгоритмы были использованы из-за их вычислительной простоты.

В экспериментах начальной точкой служила нейтральная позиция речевого тракта и использовалось от одной до четырех формант для того, чтобы определить влияние их числа на точность восстановления формы речевого тракта и на возможность достижения требуемых акустических характеристик звуков речи. Результаты вычислений формантных частот для мужского голоса показаны в табл. 9.1. Отметим, что иногда даже по одной — первой форманте можно предсказать частоты остальных формант с удовлетворительной точностью, как, например, для /А/ или /Э/. При этом вычисленная и измеренная формы тракта также оказываются довольно близкими. Все же для большей точности требуется 3—4 форманты (см. рис. 9.4), при этом и речевые сигналы, синтезированные с помощью формантного синтезатора для измеренных и вычисленных формантных частот, оказываются похожими на слух. Хуже всего аппроксимируются форма тракта и формантные частоты для гласного /О/. В этом случае сказались недостатки выбранного алгоритма покоординатного спуска.

Анализируя результаты этого эксперимента, следует обратить внимание на то, что были получены правильные артикуляторные признаки такие, как подъем и опускание гортани для /И/ и /У/, а также огубление для /У/. Место артикуляции, т. е. область наибольшего сужения, также обычно находится вблизи от реально наблюдаемого. В целом эти результаты можно оценить как обнадеживающие, и в дальнейшем рассчитывать на метод оптимизации в решении обратной задачи по крайней мере для предварительной оценки артикуляторных параметров.

§ 9.4. Регулирование артикуляторных движений

После того как с помощью внутренней модели определены цели в пространстве артикуляторных параметров, нужно позаботиться о том, чтобы реальные движения артикуляторов к заданным целям происходили в точном соответствии с желаемым фонетическим составом речевого сообщения и без генерации лишних звуков. Эта задача решается средствами теории автоматического регулирования.

Поскольку все объекты регулирования описываются дифференциальными уравнениями второго порядка, рассмотрим некоторые свойства этих уравнений. Пусть уравнение записано в виде

$$mx'' + rx' + cx = F,$$

где x — регулируемая координата, F — управляющее воздейст-

вие, m, r и c — механические параметры артикулятора. Это уравнение может быть представлено также и в других формах:

$$T^2x'' + 2qx' + x = kF,$$
$$x'' + 2q\omega_0 x' + \omega_0^2 x = k\omega_0^2 F,$$

где $2q = r/c$, $\omega_0^2 = 1/T^2 = c/m$, $k = 1/c$. Форма переходного процесса из одного положения артикулятора в другое зависит от параметров T и q. Если внешнее воздействие F имеет вид ступеньки, то переходный процесс характеризуется величиной перерегулирования σ и временем переходного процесса $T_{\text{пп}}$. Величина перерегулирования определяется как

$$\sigma = \frac{x_{\max} - x_\infty}{x_\infty} 100\%,$$

где $x_{\max}$ — максимальное значение координаты x во время переходного процесса, x_∞ — установившееся значение после затухания переходного процесса. Время переходного процесса $T_{пп}$ определяется как момент времени, после которого относительная разность значений регулируемой координаты x и установившегося значения x_∞ по модулю не превышает некоторой величины, 5%, т. е.

$$|x(t) - x_\infty| / x_\infty \leqslant 5\%.$$

В зависимости от параметра T и q характеристики переходного процесса σ и $T_{\text{пп}}$ различны, как это видно из рис. 9.5, где по оси абсцисс отложено нормированное время $\tau = t/T$, а по оси ординат — нормированный отклик x/kF. Чем меньше коэффициент затухания q, тем круче нарастает переходный процесс, перерегулирование становится больше, и процесс может приобрести колебательный характер. При больших потерях переходный процесс становится апериодическим, т. е. протекает без перерегулирования. Оптимальным значением является $q = 0{,}707 = \sqrt{2}/2$, при котором перерегулирование не превышает $0{,}05k$, а время переходного процесса — наименьшее, $T_{\text{пп}} \approx 3T$.

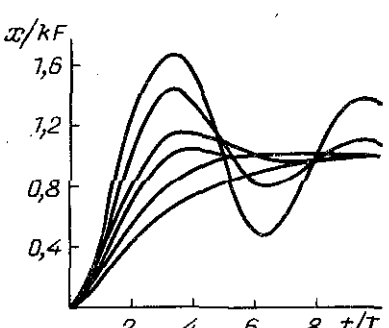


Рис. 9.5. Нормированные переходные процессы в системе второго порядка

В установившемся состоянии, после затухания переходного процесса $x' = 0$ и $x'' = 0$. Отсюда найдем величину установившегося состояния, как $x_\infty = kF$. Значит, для того чтобы регулируемая координата приняла целевое значение x^*, нужно входное воздействие F задать равным x^*/k или cx^*.

Механические параметры артикуляторов не обязательно гарантируют оптимальный переходный процесс, но можно изменить характеристики переходного процесса путем введения обратной связи по координате x и ее скорости x'. Для отрицательной обратной связи имеем

$$mx''+rx'+cx=F-c_{\text{ос}}x-r_{\text{ос}}x'.$$

Группируя члены, содержащие x и x', найдем, что параметры k, T и q стали другими:

$$k=\frac{1}{c+c_{\text{ос}}},\quad T=\sqrt{\frac{m}{c+c_{\text{ос}}}},\quad 2q=\frac{r+r_{\text{ос}}}{c+c_{\text{ос}}}.$$

При $c_{\text{ос}}\neq 0$ постоянная времени T, а следовательно, и время переходного процесса, становится меньше, но и коэффициент затухания q тоже становится меньше, что может привести к увеличению перерегулирования. Этот эффект может быть скомпенсирован соответствующим увеличением коэффициента $r_{\text{ос}}$ в цепи обратной связи по скорости. Таким образом, подбирая коэффициенты в цепях обратной связи, можно всегда добиться оптимального переходного процесса для каждого артикулятора.

Фактически обратную связь даже не нужно моделировать — просто параметр q задается равным примерно 0,7, а длительность переходного процесса определяется величиной T или ω_0, исходя из наблюдений за движениями артикуляторных органов. Тогда в канонической рекурсивной схеме для дифференциального уравнения второго порядка, которую мы анализировали в гл. 2, т. е.

$$x''+2gx'+\omega_0^2x=f,$$

коэффициент g равен $0{,}7\omega_0$, а управляющее воздействие f находится как x^*/ω_0^2. Зная, что переходный процесс заканчивается через $T_{\text{пп}}\approx 3/\omega_0$, можно так сдвинуть во времени моменты скачкообразного изменения управляющих воздействий F_i, чтобы скоординировать движения артикуляторов и достичь требуемого значения площади поперечного сечения речевого тракта в заданной области в нужный момент времени.

Требование апериодичности переходного процесса должно быть обязательно выполнено при артикуляции фрикативных звуков, где даже небольшое перерегулирование может привести к полной смычке вместо щели. В то же время для смычки согласных величина перерегулирования не играет роли, поскольку дальнейшее движение артикуляторных органов после смычки все равно невозможно. При этом в момент соприкосновения меняются и механические параметры артикуляторов. Чтобы не усложнять чрезмерно процесс регулирования переменными параметрами дифференциальных уравнений, можно просто задавать несколько иные (большие или меньшие) значения

управляющих воздействий. При этом переходные процессы могут ускориться в соответствии с возрастанием эффективной команды.

В процессе речеобразования целевые команды F_i сменяют друг друга во времени. Рекурсивная схема решения дифференциального уравнения такова, что в ней автоматически учитываются начальные условия и, если предыдущая команда была F_{i1}, а последующая — F_{i2}, то переходный процесс будет соответствовать разности этих команд: $\Delta F_i = F_{i2} - F_{i1}$. Таким образом, отпадает необходимость в специальном механизме отслеживания текущего положения координаты и ее сравнения с вновь поступившей управляющей командой. Если новая команда F_{i2} появится раньше, чем через время $T_{\text{пп}i}$ после старой команды F_{i1}, то переходный процесс не успеет завершиться, установившееся значение x_∞ не будет достигнуто, и произойдет так называемое «недорегулирование» (при $F_{i2} \leqslant F_{i1}$). Если же $F_{i2} > F_i$, то время пребывания координаты x в области x_∞ может сильно сократиться и стационарное положение вообще не реализуется. В результате недорегулирования, свойственного беглой неформальной речи, вместо смычки образуется щель, величина щели для фрикативных увеличивается, небная занавеска поднимается не полностью. Если последующая команда F_{i2} не очень отличается от предыдущей F_{i1}, то переходные процессы слабы, регулируемая координата практически меняется мало, и величина перерегулирования или недорегулирования не имеет большого значения.

В действительности, схема автоматического регулирования артикуляторных координат гораздо сложнее — она включает задержки в цепях прямого управления и обратной связи, нелинейность механорецепторов, нелинейность самого объекта управления. Поведение такой системы регулирования описывается в [59]. Если в задачу синтеза входит моделирование тонких дикторских различий, эмоциональных состояний и патологии речеобразования, то все вышеперечисленные факторы нужно обязательно учитывать. В простейшем случае, однако, можно ограничиться описанной выше схемой управления.

Рассмотрим теперь задачу координации движений артикуляторов в слитном потоке речи, т. е. задачу построения партитуры управляющих команд. В слитной речи каждый звук имеет три сегмента, последовательно расположенных во времени: переход от предыдущего звука, стационарное состояние и переход к последующему звуку. Такое разбиение на три сегмента весьма условно и часто не имеет прямого физического отображения в движениях артикуляторов, имеющих гораздо более сложную структуру. Однако на качественном уровне удобно оперировать обобщенными понятиями переходов и стационарного состояния. Стационарное состояние — это наиболее характерное для данного звука положение

артикуляторных органов, обеспечивающее его фонетическое отличие от других звуков. Для гласных и фрикативных согласных стационарное состояние полностью описывает их фонетическое качество, тогда как для взрывных согласных и аффрикат /Ц, Ч/ требуется еще и определение способа формирования переходов. В дополнение к известным звукам речи следует ввести нейтральное состояние речевого тракта, причем различаются состояния с закрытым и открытым ртом. Нейтральным состоянием начинается и заканчивается любое высказывание, а стационарное состояние безударных и редуцированных гласных в меньшей степени отличается от нейтрального состояния, чем у гласных, находящихся под ударением.

Как было показано в гл. 8, имеется 16 геометрических координат и параметров, определяющих положение и форму артикуляторных органов, а также четыре параметра, определяющих частоту основного тона, форму и амплитуду импульсов голосового источника, т. е. всего 20 параметров. Вновь перечислим эти параметры: высота голосовой щели, координаты $(x_{1\,кор},\ y_{1\,кор})$ корня языка и координаты $(x_{1\,кон},\ y_{1\,кон})$ кончика языка в подвижной системе координат $X_1O_1Y_1$, угол поворота нижней челюсти α и горизонтальное смещение точки ее вращения x_J, угол поворота небной занавески α_N, длина l_L и изменение длины губ Δl_L, а также прогиб нижней губы Δy_L, пять коэффициентов при собственных функциях языка, расстояние между задними концами голосовых складок, длина и натяжение голосовых складок и подсвязочное давление. Кроме того, для каждого параметра должны быть заданы длительность команды и ее сдвиг относительно некоторой точки отсчета, например, команды на движение основного артикулятора. Предполагается, что все команды носят ступенчатый характер, и это не является чрезмерным упрощением, поскольку изменение управляющей команды вследствие действия сигналов обратной связи уже учтено в специально подобранных параметрах дифференциальных уравнений артикуляторов.

Согласные звуки имеют в речевом тракте четко выраженное сужение — место артикуляции, и если это место приходится на губы, то максимум 5 геометрических параметров из 16 требуют спецификации для образования согласного: расстояние между голосовыми складками (определяющее звонкий или глухой тип согласного), углы поворота небной занавески и нижней челюсти, прогиб нижней губы. Остальные параметры произвольны и могут принимать значения, соответствующие соседним гласным. В то же время согласные звуки имеют более сложную временную структуру, связанную с необходимостью указания сдвига фаз в движениях небной занавески и голосовых складок относительно движения основного артикулятора. Именно поэтому оказывается условным деление

на переходные и стационарные участки. На артикуляторном уровне не существует четких границ между сегментами, принадлежащими разным звукам,— одни параметры могут принадлежать текущему звуку, а другие уже принимают значение, характерное для последующего звука.

Принято считать, что при артикуляции гласных звуков не существует явно выраженного сужения в речевом тракте, подобно тому, как это наблюдается для согласных звуков, но в действительности оказывается, что невозможно обеспечить требуемые частотные характеристики (распределение формантных частот), если минимальная площадь сужения превышает 0,4—0,6 $см^2$ (кроме /Ы/ и /Э/, см. Приложение). Вследствие этого между гласными и согласными (кроме губных) возникает конкуренция за каналы управления, особенно в том случае, когда согласный окружен гласными разного фонетического качества. При этом может оказаться, что переходный процесс по какому-либо параметру от одного состояния к другому может привести к созданию ложного места артикуляции. Если для расчета управляющих команд не используется процедура глобальной оптимизации, описанная в предыдущем разделе, то для каждого аллофона необходимо выполнять проверку на появление ложных звуков и в случае необходимости корректировать команды управления.

Каждое речевое сообщение, отделенное паузами от других сообщений, начинается и заканчивается нейтральной позицией, для которой характерна опущенная небная занавеска, опущенная гортань, разведенные голосовые складки, нулевые значения коэффициентов при собственных функциях языка. На рис. 9.6 показаны движения некоторых параметров для изолированного произнесения гласного /А/. Видно, что колебания голосового источника начинаются еще до того, как наиболее медленные артикуляторные органы займут положение, характерное для этого гласного. На конечном участке, наоборот, движения в направлении нейтрального состояния начинаются еще до прекращения фонации. Нейтрализация формантных частот хорошо видна на рис. 6.7, где частота первой форманты понижается, а второй — повышается, так что их отношение стремится к гармоническому, что свойственно однородной акустической трубе, соответствующей нейтральному артикуляторному состоянию.

Конечный сегмент высказывания перед паузой характеризуется еще одним эффектом, проявляющимся в акустической области, о чем мы говорили в гл. 5. Разведение голосовых складок в направлении нейтрального состояния приводит к тому, что импульсы голосового источника приобретают все более синусоидальный характер. Спектр сигнала возбуждения сужается, вследствие чего прекращается сначала возбуждение высших, а затем и первой форманты, и в течение нескольких десятков миллисекунд звучит лишь постепенно понижающийся

тональный сигнал на частоте основного тона. Этот сигнал служит маркером конца высказывания перед паузой наряду с другими признаками конца сообщения — удлинения последнего сегмента, оглушения звонких согласных, падения частоты основного тона.

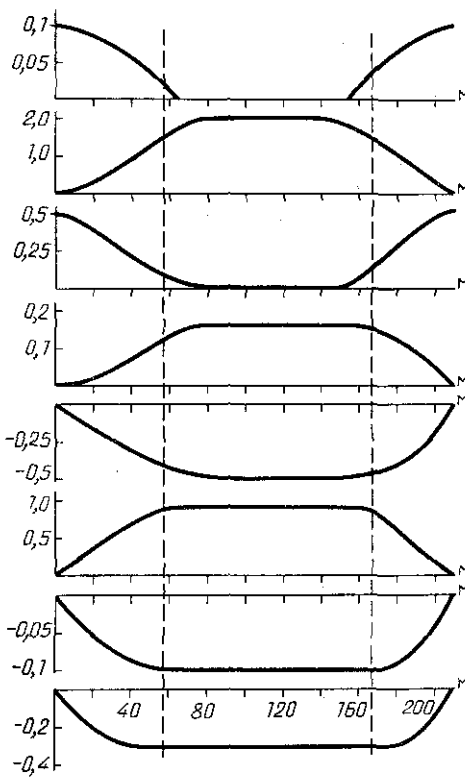


Рис. 9.6. Артикуляторные параметры в изолированном произнесении гласного /А/. *а* — угол поворота небной занавески, *б* — высота гортани, *в* — расстояние между задними концами голосовых складок, *г* — угол поворота нижней челюсти, *д* — коэффициент при первой собственной функции языка, *е* — коэффициент при второй собственной функции языка, *ж* — коэффициент при четвертой собственной функции языка, *з* — коэффициент при пятой собственной функции языка

Сдвиг по фазе между движениями артикуляторных органов иллюстрируется рис. 9.7, на котором показана временная диаграмма для слога УПИ. Разница во времени между началом движения основных для /П/ артикуляторов — нижней губы и нижней челюсти, и разведением голосовых складок нужна для того, чтобы на сегменте перехода к смычке не возникал турбулентный источник возбуждения, который может замаскировать формантные переходы. Такая тактика, однако, свойственна не всем дикторам — часто наблюдаются аспиративные участки с турбулентными шумами на интервале перехода от гласного к согласному, особенно для глухих согласных. Непостоянство характеристик возникает благодаря тому, что в русском языке неаспиративные и аспиративные согласные фонетически не противопоставляются и могут появляться в речи в произвольных соотношениях. В противоположность этому частотные характеристики взрыва с турбулентным шумом в момент раскрытия смычки служат признаками места артикуляции глухого взрывного. Поэтому при раскрытии губ голосовая щель остается открытой, и лишь затем ее площадь уменьшается до величины, при которой начинается фонация.

Поскольку изменения формы и положения языка в данном случае не препятствуют организации губной смычки, то переходные процессы по коэффициентам при собственных

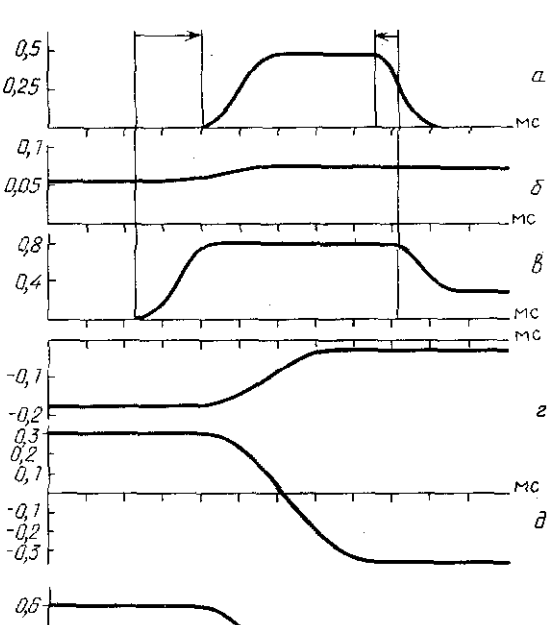


Рис. 9.7. Артикуляторные параметры для слога УПИ. а — расстояние между задними концами голосовых складок, *б* — угол поворота нижней челюсти, *в* — прогиб нижней губы, *г* — коэффициент при первой собственной функции языка, *д* — коэффициент при второй собственной функции языка, *е* — коэффициент при третьей собственной функции языка, *ж* — коэффициент при пятой собственной функции языка

функциях языка начинаются одновременно с движением нижней губы, так что коартикуляционные эффекты проявляются уже на подходе к смычке. Нижняя челюсть входит в число основных артикуляторов для согласного /П/, поэтому она управляется только командой от согласного, и ее движения не участвуют в коартикуляции гласных /У/ и /И/. Параметры подъема нижней губы и поворота нижней челюсти для согласных /П, Б, М, В, Ф/ являются приоритетными, они не могут быть замещены другими в процессах коартикуляции и должны быть особо отмечены в списке параметров этих звуков.

Организация процесса формирования управляющих команд может быть различной. Прежде всего, для каждого звука необходимо составить списки эталонных параметров с указанием длительности стационарного участка и сдвига команд по отдельным параметрам относительно команды на основной (приоритетный) параметр. Простейший способ формирования партитуры команд состоит в представлении каждого сегмента, на котором значение любого параметра сохраняется неизменным в виде столбца из 20 геометрических параметров и длительности этого сегмента. Изменение любого параметра порождает новый столбец и т. д., так что последовательность столбцов обеспечивает развертку команд во времени. Этот способ нагляден и позволяет визуально контролировать состав и длительность команд для заданного речевого сообщения. Он легко реализуется в программном обеспечении синтезатора. Однако при этом требуется много лишней работы по заполнению тех позиций столбцов, в которых параметры от сегмента не меняются, а введение новых сегментов также весьма трудоемко.

Другой способ организации данных для управления синтезатором, предложенный в [188], показан в табл. 9.2. Здесь

Таблица 9.2 Организация данных для управления синтезаторов

Номер сегмента	Значение параметра	Длительность команды	Сдвиг по началу	Относительно какого параметра

сдвиг команды относительно главного артикулятора задается в явном виде, в результате чего изменение временных характеристик по главному артикулятору автоматически приведет к пересчету моментов подачи команд на зависимые артикуляторы. Это избавляет от трудоемкой работы по согласованию движений артикуляторов и уменьшает вероятность ошибки. Становится легче и изменение временных характеристик в соответствии с просодическими законами, и при смене стиля произношения. Для того чтобы объединить преимущества обоих способов организации данных, нетрудно от второго способа перейти к первому при помощи несложных программных средств.

ГЛАВА 10

БЕГУЩИЕ ВОЛНЫ

§ 10.1. Схема Келли — Локбаума

Для решения уравнения в частных производных, которым описываются акустические процессы в речевом тракте, обычно пользуются универсальными методами решений, некоторые из которых рассматривались в гл. 7. Вместе с тем, известно, что для каждого конкретного случая можно найти специальный, более эффективный способ решения. Поскольку в акустике речеобразования доминируют нестационарные и параметрические процессы, характеризующиеся большой скоростью, то наиболее приемлемо непосредственное решение уравнения речевого тракта в пространственно-временной области, без промежуточного вычисления собственных чисел, подразумевающего стационарность системы. Одним из таких способов является схема Келли — Локбаума [127], эффективно описывающая распространение бегущих волн в речевом тракте (см. также [35], поскольку публикация [127] мало доступна). В своем первоначальном виде эта схема привлекает исключительной простотой вследствие однородности вычислительных процессов, и хотя необходимость описания реальных свойств речеобразования существенно ее усложняет, она не теряет своих основных достоинств. Идея схемы Келли — Локбаума состоит в представлении речевого тракта в виде последовательности цилиндрических секций одинаковой длины и расчете трансформаций бегущих волн при пересечении границ этих секций.

В речевом тракте имеются участки с различными свойствами площади поперечного сечения. Наиболее простым случаем является трахея, где площадь поперечного сечения постоянна по всей длине и не изменяется во времени. В носовой полости площадь поперечного сечения также не меняется во времени (за исключением участка небной занавески), но площадь соседних участков различна. В ротовой полости площадь сечения меняется и в пространстве, и во времени, но если движения артикуляторных органов сравнительно медленные, то площадь голосовой щели изменяется довольно быстро. Эти свойства необходимо учитывать для достижения

наибольшей вычислительной эффективности при расчете акустических процессов. Мы начнем описание метода бегущих волн с самого простого случая — участка речевого тракта с неизменной во времени площадью поперечного сечения $S=S(x)$, причем для ясности изложения пока не будем принимать во внимание ряд существенных явлений, таких как потери, колебания стенок и изменение длины речевого тракта.

Итак, пусть площадь $S(x)$ представлена в виде последовательности цилиндрических секций $S_i=S(i\Delta x)$, где Δx — длина, одинаковая для всех секций, $i=0,1, \ldots, N$, а отсчет секций ведется от легких к губам. Такая акустическая система описывается уравнениями сохранения количества движения и неразрывности потока:

$$S\frac{\partial P}{\partial x}=-\rho_0\frac{\partial U}{\partial t}, \tag{10.1}$$

$$S\frac{\partial P}{\partial t}=-\rho_0 c_0^2\frac{\partial U}{\partial x}, \tag{10.2}$$

где P — давление, U — объемная скорость воздуха, ρ_0 — плотность воздуха, c_0 — скорость звука в воздухе. Как известно, решение этой системы может быть представлено в виде бегущих волн, распространяющихся в противоположных направлениях, что проверяется непосредственной подстановкой соотношений

$$P(x, t)=p^+(t-x/c_0)+p^-(t+x/c_0), \tag{10.3}$$

$$U(x, t)=u^+(t-x/c_0)-u^-(t+x/c_0) \tag{10.4}$$

в систему (10.1), (10.2). Физический смысл (10.4) — это поток жидкости или газа, движущегося в одну сторону. Даже при отсутствии постоянного потока имеется смещение, обусловленное акустическими колебаниями — так называемый «акустический ветер». Примем, что волны p^+ распространяются от легких к губам, а волны p^- — от губ к легким. Для любой дифференцируемой функции f выполняется следующее соотношение:

$$\frac{\partial f(t\pm x/c_0)}{\partial x}=\pm\frac{1}{c_0}\frac{\partial f(t\pm x/c_0)}{\partial t}.$$

Воспользовавшись этим соотношением, из (10.1), (10.2) получим

$$\frac{\partial p^+(t-x/c_0)}{\partial x}+\frac{\partial p^-(t+x/c_0)}{\partial x}=\frac{\rho_0 c_0}{S}\left[\frac{\partial u^+(t-x/c_0)}{\partial x}+\frac{\partial u^-(t+x/c_0)}{\partial x}\right],$$

$$\frac{\partial p^+(t-x/c_0)}{\partial x}-\frac{\partial p^-(t+x/c_0)}{\partial x}=\frac{\rho_0 c_0}{S}\left[\frac{\partial u^+(t-x/c_0)}{\partial x}-\frac{\partial u^-(t+x/c_0)}{\partial x}\right].$$

Складывая и вычитая почленно эти уравнения, имеем

$$\frac{\partial p^+(t-x/c_0)}{\partial x}=\frac{\rho_0 c_0}{S}\frac{\partial u^+(t-x/c_0)}{\partial x},$$

$$\frac{\partial p^-(t+x/c_0)}{\partial x}=\frac{\rho_0 c_0}{S}\frac{\partial u^-(t+x/c_0)}{\partial x}.$$

В силу того что в каждой цилиндрической секции $S=\text{const}$, эту систему уравнения можно проинтегрировать по x:

$$p^+(t-x/c_0)=\frac{\rho_0 c_0}{S}u^+(t-x/c_0)+C_1, \tag{10.5}$$

$$p^-(t+x/c_0)=\frac{\rho_0 c_0}{S}u^-(t+x/c_0)+C_2, \tag{10.6}$$

где C_1 и C_2 — некоторые постоянные. Поскольку мы рассматриваем только акустические колебания, то постоянные C_1 и C_2 можно положить равными нулю. Это означает, что в данной схеме невозможно получить значения давления и потока, имеющиеся в системе в отсутствии акустических волн. Это важный момент и его необходимо иметь в виду в дальнейшем. Учитывая (10.3), получим

$$P(x, t)=\frac{\rho_0 c_0}{S}\left[u^+(t-x/c_0)+u^-(t+x/c_0)\right], \tag{10.7}$$

откуда видно, что давление P и объемная скорость U в нашей системе описываются одними и теми же компонентами бегущих волн u^+ и u^-.

Поместим центр пространственных координат посередине каждой секции. Тогда координата правого конца есть $\Delta x/2$, а левого — $\Delta x/2$, и волна, движущаяся со скоростью звука c_0, достигает их через время $\tau=\Delta x/2c_0$. На границе между двумя секциями должны выполняться условия неразрывности давления и объемной скорости:

$$U_i(t, \Delta x/2)=U_{i+1}(t, -\Delta x/2), \tag{10.8}$$

$$P_i(t, \Delta x/2)=P_{i+1}(t, -\Delta x/2), \tag{10.9}$$

где i — номер секции. Подставляя (10.4) в (10.8) и (10.5), (10.6) в (10.9), получим

$$u_i^+(t-\tau)-u_i^-(t+\tau)=u_{i+1}^+(t+\tau)-u_{i+1}^-(t-\tau), \tag{10.10}$$

$$\frac{u_i^+(t-\tau)+u_i^-(t+\tau)}{S_i}=\frac{u_{i+1}^+(t+\tau)+u_{i+1}^-(t-\tau)}{S_{i+1}}. \tag{10.11}$$

Умножим (10.11) на S_i и сложим с (10.10), а затем умножим (10.11) на S_{i+1} и вычтем из него (10.10). Это приводит к рекурсивной схеме трансформации бегущих волн, предложенной Келли и Локбаумом. Мы запишем эту схему в несколько

иной форме. Для объемной скорости имеем

$$u_{i+1}^{+}(t+\tau)=u_i^{+}(t-\tau)+\mu_i\left[u_i^{+}(t-\tau)+u_{i+1}^{-}(t-\tau)\right], \quad (10.12)$$

$$u_i^{-}(t+\tau)=u_{i+1}^{-}(t-\tau)-\mu_i\left[u_i^{+}(t-\tau)+u_{i+1}^{-}(t-\tau)\right]. \quad (10.13)$$

Выражая объемную скорость через волны давления и действуя аналогично, для давления получаем

$$p_{i+1}^{+}(t+\tau)=p_i^{+}(t-\tau)+\mu_i\left[p_{i+1}^{-}(t-\tau)-p_i^{+}(t-\tau)\right], \quad (10.14)$$

$$p_i^{-}(t+\tau)=p_{i+1}^{-}(t-\tau)+\mu_i\left[p_{i+1}^{-}(t-\tau)-p_i^{+}(t-\tau)\right]. \quad (10.15)$$

Величина μ_i называется коэффициентом отражения:

$$\mu_i=\frac{S_{i+1}-S_i}{S_{i+1}+S_i}. \quad (10.16)$$

Если соотношение между давлением и объемной скоростью выразить через импеданс секции Z_i (где $P_i=U_iZ_i$), то коэффициент отражения записывается как

$$\mu_i=\frac{Z_i-Z_{i+1}}{Z_i+Z_{i+1}}, \quad (10.17)$$

где $Z_i=\rho_0 c_0/S_i$. Представление коэффициента отражения через импеданс удобно при анализе потерь и податливости стенок.

Вследствие равенства длин секций распространение бегущих волн по всем секциям происходит синхронно. Каждый такт состоит в сдвиге волн от одного конца секции к другому (от левого к правому и от правого к левому). В элементарный такт включается также и пересечение границ секций с вычислением новых значений на внутренних границах каждой секции по (10.12), (10.13) или (10.14), (10.15). Для формирования установившегося состояния давления или объемной скорости в каждой секции требуется два таких такта, поэтому отсчеты речевой волны на выходе речевого тракта берутся с частотой, *вдвое меньшей, чем частота сдвига волн.*

Как видно, для расчета волн давления или объемной скорости в этой схеме требуется одно и то же число операций — 1 умножение и 3 сложения.

Допустим, что площадь одной из секций стала равной нулю, $S_i=0$, и рассмотрим трансформацию волн справа и слева от этой секции, а также изменение амплитуд волн в i-й секции в процессе приближения ее площади к нулю. Коэффициент отражения справа от i-й секции $\mu_{i+1}=+1$, и из (10.12) для объемной скорости имеем

$$u_{i+1}^{+}(t+\tau)=2u_i^{+}(t-\tau)+u_{i+1}^{-}(t-\tau).$$

Поскольку внутри закрытой секции волна $u_i^{+}(t-\tau)$ должна равняться нулю, то $u_{i+1}^{+}(t+\tau)=u_{i+1}^{-}(t-\tau)$, т. е. при отражении от закрытой секции волна объемной скорости сохраняет свой знак. Это согласуется с требованием, чтобы на жесткой стенке

объемная скорость равнялась нулю, так что из

$$U_{i+1}(t)=u_{i+1}^{+}(t-\tau)-u_{i+1}^{-}(t+\tau)=0,$$

снова получаем $u_{i+1}^{+}(t-\tau)=u_{i+1}^{-}(t+\tau)$. Слева от закрытой секции коэффициент отражения $\mu_{i-1}=-1$ и $u_{i-1}^{-}(t+\tau)=2u_i^{+}(t-\tau)+u_{i-1}^{-}(t-\tau)$, т. е. $u_i^{-}(t+\tau)=u_{i-1}^{+}(t-\tau)$.

Посмотрим, что происходит в i-й секции при $S_i\to 0$. Коэффициент отражения на правой границе $\mu_i\to 1$, а на левой границе $\mu_{i-1}\to -1$. При этом для правой границы имеем

$$u_{i+1}^{+}(t+\tau)\to u_{i+1}^{-}(t-\tau)+2u_{i+1}^{-}(t-\tau), \tag{10.18}$$

$$u_i^{-}(t+\tau)\to -u_i^{+}(t-\tau), \tag{10.19}$$

а для левой границы

$$u_i^{+}(t+\tau)\to -u_i^{-}(t-\tau), \tag{10.20}$$

$$u_{i-1}^{-}(t+\tau)\to u_{i-1}^{+}(t-\tau)+2u_i^{-}(t-\tau). \tag{10.21}$$

Выражения (10.19) и (10.20) характеризуют отражение волны объемной скорости от конца, выходящего в свободное пространство, так как отношение площади S_i при $S_i\to 0$ к площади секции с ненулевыми S_{i-1} и S_{i+1} очень мало. При этом волна объемной скорости меняет свой знак, поскольку известно, что давление на конце трубы, выходящей в свободное пространство, равно нулю, и из

$$\frac{S_i}{\rho_0 c_0}P_i(t)=u_i^{+}(t-\tau)+u_i^{-}(t+\tau)=0,$$

получаем $u_i^{+}(t-\tau)=-u_i^{-}(t+\tau)$. Обратим, однако, внимание на то, что внутри i-й секции волны не убывают, а продолжают циркулировать даже и после достижения площадью S_i нулевого значения. При этом и в выражениях (10.18), (10.21) волны от i-й секции не обращаются в нуль, и схема бегущих волн теряет устойчивость — происходит неограниченное возрастание амплитуды бегущих волн.

Если в секции с уменьшающейся до нуля площадью поперечного сечения попытаться определить акустическое давление по (10.7), то оно будет неограниченно возрастать по мере уменьшения S_i. Очевидно, что такое физически неправильное поведение является следствием отсутствия потерь на вязкое трение, которые неограниченно растут при стремлении площади сечения секции к нулю, и, таким образом, сводят до нуля амплитуды волн, циркулирующие в смыкающейся секции. Следовательно, потери на трение принципиально необходимо учитывать в каждой секции, площадь которой может стать равной нулю.

Анализируя поведение волн давления, найдем, что при отражении от закрытой секции бегущая волна меняет свой знак на обратный, а при отражении от конца, выходящего

в свободное пространство, знак отраженной волны не меняется. В системе без потерь для давления волны справа и слева от секции с площадью, стремящейся к нулю, описываются как

$$p_{i+1}^{+}(t+\tau)\to p_{i+1}^{-}(t-\tau), \tag{10.22}$$

$$p_{i}^{-}(t+\tau)\to 2p_{i+1}^{-}(t-\tau)-p_{i}^{+}(t-\tau), \tag{10.23}$$

$$p_{i}^{+}(t+\tau)\to 2p_{i}^{+}(t-\tau)-p_{i+1}^{-}(t-\tau), \tag{10.24}$$

$$p_{i-1}^{-}(t+\tau)\to p_{i}^{+}(t-\tau). \tag{10.25}$$

Из (10.22) и (10.25) мы видим, что справа и слева от секции с $S_i\to 0$ волны давления ведут себя правильно, тогда как в самой i-й секции продолжается циркуляция волн с увеличением амплитуды и последующей потерей устойчивости решения. Из сравнения поведения волн объемной скорости и давления в системе без потерь можно предположить, что схема для давления предъявляет несколько менее сильные требования к точности описания потерь, чем схема для объемной скорости, но и в том, и в другом случае рано или поздно система без потерь теряет устойчивость.

Дальнейшее описание распространения бегущих волн в речевом тракте требует анализа граничных условий со стороны легких и губ.

§ 10.2. Граничные условия со стороны легких

Простейшая модель легких — это емкость с некоторым сопротивлением (см. рис. 10.1). Эта модель получается из схемы резонатора Гельмгольца с длиной горла, равной нулю. Для того чтобы яснее представить процессы, происходящие при отражении волн от легких, начнем с этой, на самом деле чрезмерно упрощенной, модели. Импеданс такой схемы есть

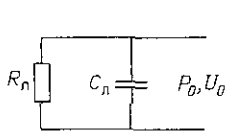


Рис. 10.1. Простейшая электрическая схема легких

$$Z_л=\frac{1+j\omega R_л C_л}{j\omega C_л}, \tag{10.26}$$

где $C_л=V_л/\rho_0 c_0^2$, $V_л$ — объем легких (равный примерно 3000—6000 см^3), а сопротивление $R_л$, вообще говоря, зависит от величины объемной скорости потока, вытекающего из легких. Это сопротивление находится в диапазоне 3—15 г/(см$^4\cdot$с).

Из граничных условий $P_0=-Z_л U_0$ (знак минус взят с учетом принятого положительного направления распространения волн) получаем

$$Z_0(u_0^{+}+u_0^{-})=-Z_л(u_0^{+}-u_0^{-}),$$

где $Z_0=\rho_0 c_0/S_0$, S_0 — площадь трахеи. Учитывая, что умножение на $j\omega$ соответствует однократному дифференцирова-

нию по времени, имеем

$$T_{л2}u_0^{+\prime}+u_0^{+}=T_{л1}u_0^{-\prime}+u_0^{-}, \qquad (10.27)$$

где $T_{л1}=C_л(R_л-Z_0)$, $T_{л2}=C_л(R_л+Z_0)$. Взяв преобразование Лапласа от (10.27), получим передаточную функцию легких для отраженной волны:

$$W_0(s)=-\frac{u^+(s)}{u^-(s)}=\frac{1+T_{л1}s}{1+T_{л2}s},$$

где s — комплексная частота. Такая передаточная функция соответствует параллельному соединению апериодического звена $1/(1+T_{л1}s)$ и звена с реальным дифференцированием $T_{л2}s/(1+T_{л1}s)$ и легко реализуется во временной области.

Из (10.27) видно, что в установившемся режиме волна отражается от легких с сохранением своего знака, т. е. как от свободного пространства. Это объясняется следствием выбора модели в виде полости большого объема и соответствующего ей импеданса (10.26). Постоянная времени $T_{л2}\approx 0{,}05-0{,}09$ с, что соответствует характеристической частоте 3—6 Гц, близкой к экспериментально измеренной резонансной частоте легких.

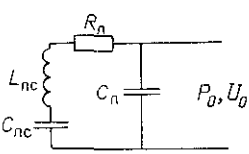


Рис. 10.2. Эквивалентная электрическая схема легких

Одним из важных свойств легких является податливость их стенок. Включая податливость в модель, получим электрический аналог, показанный на рис. 10.2, где в $R_л$ включены также и потери на колебание стенок. Учитывая, что емкость $C_л\approx 0{,}003$ см$^4\cdot$с^2/г для объема легких $V_л=4500$ см^3 значительно меньше емкости $C_{лс}=0{,}1$ см$^4\cdot$с^2/г, получим импеданс такой системы

$$Z_л=\frac{1+j\omega C_{лс}R_л+j\omega C_{лс}j\omega L_л}{j\omega C_{лс}(1+j\omega C_л R_л+j\omega C_л j\omega L_л)}.$$

Во временной области это соответствует уравнению

$$u_0^{+\prime\prime}+2g_л u_0^{+\prime}+\omega_л^2 u_0^{+}+\frac{C_л}{Z_0C_{лс}L_л}\int u_0^{+}dt=F_0, \qquad (10.28)$$

где

$$F_0=\frac{C_л}{Z_0C_{лс}L_л}\int u_0^{-}dt+\frac{(R_л-Z_0)}{Z_0L_л}u_0^{-}+\frac{(L_л-Z_0R_лC_л)}{Z_0L_л}u_0^{-\prime}-u_0^{-\prime\prime},$$

$$2g_л=\frac{(L_л+Z_0R_лC_л)C_л}{Z_0L_л},\quad \omega_л^2=\frac{(R_л+Z_0)C_л}{Z_0L_л}.$$

Перенесем член с интегралом в левой части (10.28) направо и запишем конечно-разностную схему для этого уравнения:

$$u_{0i}^{+}=a_0\overline{F_0}+a_1u_{0i-1}^{+}+a_2u_{0i-2}^{+},$$

где

$$\overline{F_0} = F_0 - \frac{C_л}{Z_0 C_{лс} L_л} \int u_0^+ dt,$$

а u_{0i-1}^+ и u_{0i-2}^+ — значения волны u_0^+ в момент времени $t - \Delta t_s$ и $t - 2\Delta t_s$. Представим интеграл от объемной скорости как

$$\int u_0^+ dt = 0{,}5\, u_{0i}^+ + I_{i-1},$$

где $I_{i-1} = 0{,}5\, u_{i-1}^+ + I_{i-2}$. Перенося обратно в левую часть член с u_{0i}^+, получим

$$u_{0i}^+ \left(1 + \frac{0{,}5 a_0 C_л}{Z_0 C_{лс} L_л}\right) = a_0 (F_0 + I_{i-1}) + a_1 u_{0i-1}^+ + a_2 u_{0i-2}^+.$$

Индуктивность податливых легких, по оценке [159], $L_л \approx 0{,}036$ г/см^2, поэтому $\omega_л \approx 0{,}3 - 0{,}5$ с$^{-1}$, но это не есть резонансная частота легких.

В действительности легкие не являются некой полостью объемом $V_л$, а состоят из множества дихотомически ветвящихся бронхиол, причем на каждом ветвлении происходит отражение акустических волн. Однако полная акустическая модель легких весьма сложна и в артикуляторных синтезаторах пока не используется.

§ 10.3. Граничные условия со стороны губ

В основе всех способов оценки граничных условий со стороны губ лежит анализ излучения поршня радиусом r_N в сфере гораздо большего радиуса, выполненный в [37]. Импеданс такого излучателя описывается через Бесселевы функции, однако с точностью до 10% его можно аппроксимировать простой формулой

$$Z_l = \frac{(\omega r_N)^2}{2c_0^2} + j\frac{8\omega r_N}{3\pi c_0}. \qquad (10.30)$$

Отметим, что Z_l — безразмерный импеданс. Как видно, потери на излучение пропорциональны квадрату радиуса поршня r_N, т. е. его площади, и квадрату частоты ω. Зависимость потерь на излучение от площади губного отверстия имеет ясный физический смысл — чем больше это отверстие, тем больше звуковой энергии уходит из речевого тракта. Зависимость сопротивления излучения от квадрата частоты не представляет затруднений в методах синтеза речи, использующих собственные частоты, в частности, в формантных синтезаторах. Метод бегущих волн оперирует лишь пространственно-временными соотношениями, и поэтому зависимость сопротивления от частоты вносит некоторые трудности. Для того чтобы избавиться от такой зависимости, в [65] было предложено перейти от последовательного соединения сопротивления и индуктив-

ности, которое служит электрическим аналогом (10.30), к параллельному, при сохранении тех же электрических свойств. Запишем импеданс губного отверстия как $Z_{l1}=R_{l1}+j\omega L_{l1}$, а импеданс схемы с параллельным соединением — как

$$Z_{l2}=\frac{j\omega R_{l2}L_{l2}}{R_{l2}+j\omega L_{l2}},$$

и приравняем мнимые и действительные части этих импедансов, получив при этом

$$R_{l1}=\frac{R_{l2}}{1+\left(\frac{R_{l2}}{\omega L_{l2}}\right)^2},\quad L_{l1}=\frac{L_{l2}}{1+\left(\frac{\omega L_{l2}}{R_{l2}}\right)^2}.$$

Взяв отношение этих выражений, имеем

$$\frac{R_{l1}}{L_{l1}}=\frac{\omega L_{l2}}{R_{l2}}.$$

Если $R_{l1}<<L_{l1}$, что выполняется для частот ниже 1 кГц, то и $\omega L_{l2}<<R_{l2}$, откуда получаем $L_{l2}\approx L_{l1}$, $R_{l2}=\omega^2 L_{l2}^2/R_{l1}=128/9\pi^2$. Положим, что импеданс излучения $Z_l=Z_{l2}$, и найдем выражение для трансформации бегущих волн, отражающихся от губного отверстия. Из равенства $P_l=U_l Z_l$ для волн объемной скорости имеем

$$\frac{\rho_0 c_0}{S_N}\left(u_N^+ + u_N^-\right)=\frac{\rho_0 c_0}{S_N}Z_l\left(u_N^+ - u_N^-\right).$$

Множитель $\rho_0 c_0/S_N$ при Z_l появился потому, что Z_l — удельный безразмерный импеданс. Отсюда получаем

$$T_{l1}u_N^{-'}+u_N^-=T_{l2}u_N^{+'}-u_N^+, \qquad (10.31)$$

где

$$T_{l2}=L_{l2}\frac{R_{L2}+1}{R_{l2}}\approx 0{,}5\cdot 10^{-5}\sqrt{S_N},$$

$$T_{l1}=L_{l2}\frac{R_{L2}-1}{R_{l2}}\approx 2{,}3\cdot 10^{-5}\sqrt{S_N}.$$

В диапазоне параметров $S_N\leqslant 10$ см^2 и $f\leqslant 9$ кГц волна u_N^-, отраженная от губного отверстия, изменяет свой знак, что похоже на отражение от жесткой стенки. Передаточная функция излучающего отверстия

$$W_N(s)=\frac{u_N^-}{u_N^+}=-\frac{1-T_{l1}s}{1+T_{l2}s},$$

где s — комплексная частота. Как видно, $W_N(s)$ представляет собой параллельное соединение реального дифференцирующего звена и апериодического звена, в результате чего отраженную волну u_N^- можно найти без дифференцирования прямой

волны u_N^+:

$$u_N^-(t)=\left[3\bar{u}(t)-4\bar{u}(t-\Delta t_s)+\bar{u}(t-2\Delta t_s)\right]\frac{T_{t1}}{2\Delta t_s}-\bar{u}(t),$$

где

$$\bar{u}(t)=\bar{u}(t-\Delta t_s)+\left[u^+(t)-\bar{u}(t-\Delta t_s)\right]\left(1-e^{-\Delta t_s/T_{t2}}\right).$$

Напомним, что $\Delta t_s=2\Delta t$, так как частота отсчетов сигнала вдвое меньше частоты сдвига волн.

Аппроксимируя импеданс излучения Z_l с помощью гармонических функций, можно посредством минимизации среднеквадратической ошибки получить другую форму [141]:

$$Z_l\approx a\frac{\rho_0 c_0}{S_N}\frac{1-z^{-1}}{1-bz^{-1}},$$

где z^{-1} — оператор задержки сигнала на один шаг по времени, а коэффициенты a и b представляются как

$$a=0{,}0779+0{,}2373\sqrt{S_N},$$
$$b=-0{,}843+0{,}3062\sqrt{S_N}.$$

Эта форма справедлива при $0{,}5\leqslant S_N\leqslant 6$ см^2. Пользуясь этим импедансом, получаем соотношения для прямой и отраженной волн в последней секции речевого тракта:

$$(1-a)u_N^--(a+b)z^{-1}u_N^-=(a-1)u_N^++(b-a)z^{-1}u_N^+,$$

или

$$(1-a)u_N(t)-(a+b)u_N^-(t-\Delta t_s)=(a-1)u_N^+(t)+(b-a)u_N^+(t-\Delta t_s),$$

что с точностью до коэффициентов при переменных u_N^+ и u_N^- совпадает с (10.31), представленной в конечно-разностной форме, если взять левую производную по времени, т. е. $U'=\left[U(t)-U(t-\Delta t_s)\right]/\Delta t_s$.

В свободное пространство излучается волна с линейной колебательной скоростью V_s:

$$V_s=(u_N^+-u_N^-)/S_N.$$

Как было показано в [65], экранирующее влияние головы приводит к следующей зависимости звукового давления P_s в свободном пространстве от колебательной скорости V_s на выходе речевого тракта:

$$P_s=V_s\frac{\rho_0 r_N c_0}{r_a}\frac{j\omega\frac{r_N}{c_0}}{1+j\omega\frac{r_N}{c_0}}\exp\left\{-j\omega\frac{r_a-r_N}{c_0}\right\}, \qquad (10.32)$$

где r_a — расстояние от губ до приемника звукового давления. Передаточная функция для P_s представляет собой звено

с реальным дифференцированием, или последовательное соединение идеального дифференцирующего звена с апериодическим звеном. Следовательно, во временной области можно записать

$$P_s(t)=a\left\{\bar{P}(t-\Delta t_s)+\left[V_s(t)-\bar{P}_s(t-\Delta t_s)\right]\left(1-\exp\left\{-\frac{\Delta t_s c_0}{r_N}\right\}\right)\right\}, \tag{10.33}$$

где

$$\bar{P}(t)=3V_s(t)-4V_s(t-\Delta t_s)+V_s(t-2\Delta t_s),$$

$$a=\frac{\rho_0 S_N}{\pi r_a}.$$

Особый случай представляет излучение из носовой полости через стенки ноздрей, а также через стенки речевого тракта. Мы рассмотрим это излучение в разделе, посвященном податливости стенок.

§ 10.4. Разветвление речевого тракта

Акустическая система речевого тракта разветвляется по крайней мере в двух случаях: при артикуляции носовых звуков и звука /Л/. В первом случае небная занавеска опускается, открывая доступ в носовую полость, причем и сама носовая полость состоит из двух полостей, разделенных перегородкой. Обычно разветвлением внутри носовой полости пренебрегают, но это правильно лишь в случае абсолютной идентичности носовых полостей. Если эквивалентная длина полостей или их форма различны, то в месте разветвления возникает отражение волн и характеристики звука изменяются. Во втором случае разветвление тракта происходит в результате касания кончиком языка твердого неба с опусканием боковых участков языка. Речевой тракт разветвляется на небольшом протяжении и снова соединяется в одну полость. Характеристики звуков /М, Н, Л/ оказываются до некоторой степени похожими, что подтверждается ошибками восприятия речи в условиях помех [59]. При этом группа назальных звуков /М, Н/ не смешивается ни с какими звуками, кроме /Л/. Это сходство, несомненно, объясняется тем, что и в том, и в другом случае речевой тракт разветвляется, хотя геометрические характеристики ветвей весьма различны.

Рассмотрим прохождение и отражение бегущих волн в сечении, где речевой тракт разветвляется. Пусть номер секции, предшествующей разветвлению, есть $j-1$ и площадь этой секции S_{j-1}, а ветви имеют площади S_{1j} и S_{2j}. Тогда на границе $(j-1)$-й и j-й секции должны соблюдаться условия

равенства давления

$$P_{j-1}(t,\ \Delta x/2)=P_{1j}(t,\ -\Delta x/2)=P_{2j}(t,\ -\Delta x/2), \quad (10.34)$$

и неразрывности объемной скорости

$$U_{j-1}(t,\ \Delta x/2)=U_{1j}(t,\ -\Delta x/2)+U_{2j}(t,\ -\Delta x/2). \quad (10.35)$$

Пользуясь соотношениями (10.4), (10.5) и (10.6), из (10.34) и (10.35) получим систему из трех уравнений для волн объемной скорости

$$\begin{aligned} u^+_{1j}+u^+_{2j}+u^-_{j-1} &= u^+_{j-1}+u^-_{1j}+u^-_{2j},\\ S_{j-1}u^+_{1j}-S_{1j}u^-_{j-1} &= S_{1j}u^+_{j-1}-S_{j-1}u^-_{1j},\\ S_{j-1}u^+_{2j}-S_{2j}u^-_{j-1} &= S_{2j}u^+_{j-1}-S_{j-1}u^-_{2j}. \end{aligned} \quad (10.36)$$

Решая эту систему, найдем трансформацию бегущих волн, проходящих в разветвление и обратно:

$$\begin{aligned} u^+_{1j} &= u^+_{j-1}+u^+_{2j}+\mu_{1j}(u^+_{j-1}+u^-_{1j}+u^-_{2j}),\\ u^+_{2j} &= u^+_{j-1}+u^-_{1j}+\mu_{2j}(u^+_{j-1}+u^-_{1j}+u^-_{2j}),\\ u^-_{j-1} &= u^-_{1j}+u^-_{2j}-\mu_j(u^+_{j-1}+u^-_{1j}+u^-_{2j}), \end{aligned} \quad (10.37)$$

где

$$\mu_{1j}=\frac{S_{1j}-(S_{j-1}+S_{2j})}{S_{1j}+S_{j-1}+S_{2j}}, \quad (10.38)$$

$$\mu_{2j}=\frac{S_{2j}-(S_{j-1}+S_{1j})}{S_{1j}+S_{j-1}+S_{2j}}, \quad (10.39)$$

$$\mu_j=\frac{(S_{1j}+S_{2j})-S_{j-1}}{S_{1j}+S_{j-1}+S_{2j}}. \quad (10.40)$$

Если площадь одной из ветвей равна нулю ($S_{1j}=0$ или $S_{2j}=0$), то все коэффициенты отражения переходят в (10.16), что соответствует неразветвленному тракту.

Действуя аналогичным образом, получим рекурсивные *формулы для трансформации бегущих волн давления*:

$$\begin{aligned} p^+_{1j} &= p^+_{j-1}+p^-_{2j}+p_{j,j-1},\\ p^+_{2j} &= p^+_{j-1}+p^-_{1j}+p_{j,j-1},\\ p^-_{j-1} &= p^-_{1j}+p^-_{2j}+p_{j,j-1}, \end{aligned} \quad (10.41)$$

где

$$p_{j,j-1}=\mu_{1j}p^-_{1j}+\mu_{2j}p^-_{2j}-\mu_j p^+_{j-1},$$

а коэффициенты отражения μ_{1j}, μ_{2j} и μ_j — те же, что и для волн объемной скорости. Как видно, в схеме для давления (10.37) требуется на два умножения больше, чем в схеме для объемной скорости. По сравнению с неразветвленным трактом при расчете волн давления требуется на 2 умножения и 6 сложений больше, а для волн объемной скорости — только на 6 сложений больше. Увеличение количества вычислительных операций при разветвлении акустической системы, однако,

оказывает небольшое влияние на сложность алгоритма расчета бегущих волн, так как имеется всего лишь несколько участков с разветвлением, тогда как количество секций измеряется несколькими десятками.

§ 10.5. Потери в речевом тракте

Потери в речевом тракте приводят к затуханию колебаний и определяют ширину формант. Как мы видели ранее, пренебрежение потерями в секции, площадь которой стремится к нулю, приводит к неограниченному возрастанию колебаний. *Рассмотрим различные способы введения потерь в рекурсивную* схему для бегущих волн. Сначала воспользуемся определением коэффициента потерь через импеданс цилиндрической секции (10.17):

$$\mu_i = \frac{Z_i - Z_{i+1}}{Z_i + Z_{i-1}}.$$

Определим импеданс секции с потерями как

$$Z_i = Z_i^{(0)} + R_i, \quad Z_{i+1} = Z_{i+1}^{(0)} + R_{i+1},$$

где $Z_i^{(0)} = \rho_0 c_0 / S_i$, $Z_{i+1} = \rho_0 c_0 / S_{i+1}$, R_i и R_{i+1} — активное сопротивление в каждой секции (полное, а не на единицу длины). Тогда коэффициент отражения есть

$$\mu_i = \frac{Z_i^{(0)} - Z_{i+1}^{(0)} + R_i - R_{i+1}}{Z_i^{(0)} + Z_{i+1}^{(0)} + R_i + R_{i+1}}. \tag{10.43}$$

Поскольку $(R_i - R_{i+1}) < (R_i + R_{i+1})$, то коэффициент отражения для системы с потерями всегда меньше коэффициента отражения для системы без потерь и, таким образом, амплитуда бегущих волн уменьшается. При стремлении площади секции к нулю ее сопротивление стремится к бесконечности, и волны, подходящие справа и слева от секции с нулевой площадью, отражаются с изменением знака для объемной скорости и с сохранением знака для давления. Однако внутри секции с $S \to 0$ волны по-прежнему не затухают. Это естественно, так как и в этом случае мы считаем, что все изменения волн происходят на границе секций, а внутри каждой секции волны распространяются, сохраняя неизменной свою форму и амплитуду. Следовательно, выражение (10.43) для коэффициента отражения в системе с потерями пригодно лишь для малых потерь.

Малые потери на вязкое трение возникают при достаточно большой площади сечения, причем сопротивлением пропорционально корню квадратному от частоты:

$$R = \frac{\mathscr{L} \Delta x}{S} \sqrt{\frac{\rho_0 \mu}{2}}\, \omega,$$

где $\mathscr{L}$ — периметр сечения, Δx — длина участка, μ — коэффициент

вязкого трения. В [146] зависимость квадратного корня от частоты аппроксимируется как

$$R = R_0 \frac{z-b}{z-a}, \tag{10.44}$$

где a и b подбираются для наилучшей аппроксимации, а z^{-1} по-прежнему соответствует задержке сигнала на Δt_s. При малых площадях сечения сопротивление не зависит от частоты, а обратно пропорционально кубу площади сечения S^3, так что при стремлении S к нулю R стремится к бесконечности. При переходе от одной формы зависимости сопротивления от частоты и площади сечения к другой необходимо позаботиться об отсутствии разрывов и достаточной гладкости, поскольку несогласование приводит к скачкам в коэффициенте отражения, и, следовательно, к резкому изменению амплитуды бегущих волн, что ухудшает качество синтетической речи.

Рассмотрим теперь другой способ использования потерь, исходя из основной системы уравнений

$$-S\frac{\partial P}{\partial x} = \rho_0 \frac{\partial U}{\partial t} + r_1 U, \tag{10.45}$$

$$-\rho_0 c_0^2 \frac{\partial U}{\partial x} = S\frac{\partial P}{\partial t} + r_2 P. \tag{10.46}$$

Возьмем производную по x от (10.45) и производную по t от (10.46), приравняем одинаковые члены и, вновь используя (10.46) для получения однородности относительно давления, имеем

$$\frac{\partial^2 P}{\partial x^2} = \frac{1}{c_0^2}\frac{\partial^2 P}{\partial t^2} + \frac{1}{c_0^2}\left(\frac{r_1}{\rho_0} + \frac{r_2}{S}\right)\frac{\partial P}{\partial t} + \frac{r_1 r_2}{S\rho_0 c_0^2} P. \tag{10.47}$$

Возьмем производную по t от (10.45) и производную по x от (10.46), приравняем одинаковые члены и, используя (10.45) для получения однородности относительно объемной скорости, получим

$$\frac{\partial^2 U}{\partial x^2} = \frac{1}{c_0^2}\frac{\partial^2 U}{\partial t^2} + \frac{1}{c_0^2}\left(\frac{r_1}{\rho_0} + \frac{r_2}{S}\right)\frac{\partial U}{\partial t} + \frac{r_1 r_2}{S\rho_0 c_0^2} U. \tag{10.48}$$

В проведенных выше преобразованиях считалось, что площадь поперечного сечения не меняется ни как функция времени t, ни как функция x, поскольку имелась в виду цилиндрическая секция. Сравнивая (10.47) и (10.48), видим, что они имеют один и тот же коэффициент потерь

$$g = \frac{1}{2c_0^2}\left(\frac{r_1}{\rho_0} + \frac{r_2}{S}\right).$$

При малых потерях, когда коэффициент g значительно меньше частоты первого резонанса акустической системы, форма

бегущих волн не изменяется, а лишь уменьшается их амплитуда. Поэтому можно записать

$$P(x, t)=e^{-g(t+x/c_0)}p^{+}(t-x/c_0)+e^{-g(t-x/c_0)}p(t+x/c_0), \quad (10.49)$$

$$U(x, t)=e^{-g(t+x/c_0)}u^{+}(t-x/c_0)-e^{-g(t-x/c_0)}u^{-}(t+x/c_0). \quad (10.50)$$

и в дальнейшем исключить члены с потерями из (10.45) и (10.46). Действуя так же, как было описано в разделе 10.1, получим

$$P(x, t)=\frac{\rho_0 c_0}{S}\left[e^{-g(t+x/c_0)}u^{+}(t-x/c_0)+\right.$$
$$\left.+e^{-g(t-x/c_0)}u^{-}(t+x/c_0)\right]. \quad (10.51)$$

Граничные условия для секций с потерями теперь выглядят как

$$u_i^{+}(t-\tau)e^{-g_i\tau}-u_i^{-}(t+\tau)e^{g_i\tau}=u_{i+1}^{+}(t+\tau)e^{g_{i+1}\tau}-u_{i+1}^{-}(t-\tau)e^{g_{i+1}\tau},$$

$$\frac{u_i^{+}(t-\tau)e^{-g_i\tau}}{S_i}+\frac{u_i^{-}(t+\tau)e^{g_i\tau}}{S_i}=\frac{u_{i+1}^{+}(t+\tau)e^{g_{i+1}\tau}}{S_{i+1}}+\frac{u_{i+1}^{-}(t-\tau)e^{g_{i+1}\tau}}{S_{i+1}}.$$

Преобразуя эту систему к рекурсивной форме, получим

$$u_{i-1}^{+}(t+\tau)=\left\{u_i^{+}(t-\tau)e^{-g_i\tau}+\mu_i\left[u_i^{+}(t-\tau)e^{-g_i\tau}+\right.\right.$$
$$\left.\left.+u_{i+1}^{-}(t-\tau)e^{-g_{i+1}\tau}\right]\right\}e^{-g_{i+1}\tau},$$
$$u_i^{-}(t+\tau)=\left\{u_{i+1}^{-}(t-\tau)e^{-g_{i+1}\tau}-\mu_i\left[u_i^{+}(t-\tau)e^{-g_i\tau}+\right.\right. \quad (10.52)$$
$$\left.\left.+u_{i+1}^{-}(t-\tau)e^{-g_{i+1}\tau}\right]\right\}e^{-g_i\tau}.$$

Если потери в соседних секциях одинаковы, т. е. $g_i=g_{i+1}=g$, то система (10.52) упрощается:

$$u_{i+1}^{+}(t+\tau)=\left\{u_i^{+}(t-\tau)+\mu_i\left[u_i^{+}(t-\tau)+u_{i+1}^{-}(t-\tau)\right]\right\}e^{-2g\tau},$$
$$u_i^{-}(t+\tau)=\left\{u_{i+1}^{-}(t-\tau)-\mu_i\left[u_i^{+}(t-\tau)+u_{i+1}^{-}(t-\tau)\right]\right\}e^{-2g\tau}.$$

Здесь коэффициент отражения имеет тот же смысл, что и для системы без потерь, т. е.

$$\mu_i=(S_{i+1}-S_i)/(S_{i+1}-S_i).$$

Анализируя систему (10.52), видим, что она не только учитывает потери путем уменьшения амплитуды волн, пересекающих границу секций, но также обеспечивает сведение до нуля амплитуды волн, циркулирующих внутри секции, площадь которой стремится к нулю. И хотя при этом нарушается условие малости потерь, тем не менее рекурсивная схема не теряет устойчивости при возникновении смычки в речевом тракте.

Для волн давления в системе с потерями получаем аналогичную рекурсивную схему:

$$p_{i+1}^{+}(t+\tau)=\{p_i^{+}(t-\tau)e^{-g_i\tau}+\mu_i\left[p_{i+1}^{-}(t-\tau)e^{-g_{i+1}\tau}-\right.$$
$$\left.-p_i^{+}(t-\tau)e^{-g_i\tau}\right]\}e^{-g_{i+1}\tau},$$

$$p_i^{-}(t-\tau)=\{p_{i+1}^{-}(t-\tau)e^{-g_{i+1}\tau}+\mu_i\left[p_{i+1}^{-}(t-\tau)e^{-g_{i+1}\tau}-\right.$$
$$\left.-p_i^{+}(t-\tau)e^{-g_i\tau}\right]\}e^{-g_i\tau}.$$

Эта схема обладает теми же свойствами, что и (10.52) относительно потерь в области, где площадь поперечного сечения стремится к нулю.

При малых потерях можно воспользоваться разложением экспоненты в ряд Тейлора

$$e^{-x}=1-x+\frac{x^2}{2!}-\frac{x^3}{3!}+\ldots,$$

и ограничиться, например, первыми двумя членами:

$$e^{-g\tau}\approx 1-g\tau.$$

И в этом случае нужно позаботиться о том, чтобы переход от приближенного представления экспоненты к точному не сопровождался скачками в амплитуде бегущих волн.

При малых потерях (10.52) можно представить в виде двух компонент, одна из которых соответствует распространению волн в системе без потерь, а другая — поправкам на потери:

$$u_{i+1}^{+}(t+\tau)=u_i^{+}(t-\tau)+\mu_i\left[u_i^{+}(t-\tau)+u_{i+1}^{-}(t-\tau)\right]-$$
$$-\{u_i^{+}(t-\tau)g_i+\mu_i\left[u_i^{+}(t-\tau)g_i+u_{i+1}^{-}(t-\tau)g_{i+1}\right]\}(1-g_{i+1}\tau)\tau,$$
$$u_i^{-}(t+\tau)=u_{i+1}^{-}(t-\tau)-\mu_i\left[u_i^{+}(t-\tau)+u_{i+1}^{-}(t-\tau)\right]- \qquad (10.53)$$
$$-\{u_{i+1}^{-}(t-\tau)g_{i+1}-\mu_i\left[u_i^{+}(t-\tau)g_i+u_{i+1}^{-}(t-\tau)g_{i+1}\right]\}(1-g_i\tau)\tau.$$

Аппроксимируя зависимость коэффициента затухания от частоты с помощью (10.44) из (10.53) получим конечно-разностное уравнение, которое мы не будем приводить ввиду его громоздкости. В этом уравнении используются не только значения бегущих волн в момент $t-\tau$ и $t+\tau$, но также и их значения, задержанные на $\Delta t=2\tau$ и $2\Delta t$. Усложнение вычислительной схемы оправдывается необходимостью использования частотно-зависимых потерь с целью подавления паразитных сигналов в рекурсивной схеме.

Физически каждая секция речевого тракта представляет собой фильтр низких частот, тогда как в рекурсивной схеме с частотно-независимыми потерями шумы вычислений не фильтруются, а могут накапливаться, ухудшая качество син-

тезированного сигнала. Можно попытаться сократить количество вычислений, разместив фильтры в отдельных участках речевого тракта, но при этом необходимо учитывать связь между потерями и пучностями собственных функций мод акустических колебаний. Размещение фильтра с потерями в узле какой-либо моды не повлияет на ширину соответствующей форманты.

Мы рассмотрели два вида формул для потерь на вязкое трение вблизи стенок речевого тракта — зависимой от корня квадратного от частоты для больших площадей поперечного сечения и зависимой лишь от площади поперечного сечения для узких щелей:

$$R = 12\mu d^2 \Delta x / S^3,$$

где d — высота щели. В речевом тракте, однако, имеются и другие потери. Потери на теплопроводность, представленные членом с коэффициентом r_2 в (10.46), имеют такую же зависимость от частоты, что и потери на вязкое трение при больших площадях сечения. При малых площадях сечения, еще до того, как начинает доминировать капиллярное вязкое трение, возникает динамическое сопротивление, которое в несколько раз больше сопротивления вязкого трения:

$$R_д = c_R \rho_0 v_0 / (2S),$$

где c_R — коэффициент, зависящий от формы сужения вдоль пространственной координаты x, v_0 — линейная скорость постоянного воздушного потока, протекающего через сужение. При больших площадях сечения линейная скорость v_0 мала и динамическое сопротивление мало. Но при площади сужения, близкой к 0,2 см^2, динамическое сопротивление на порядок больше сопротивления вязкого трения.

§ 10.6. Колебания стенок

Податливость стенок играет важную роль в процессах речеобразования — частота первого резонанса во время смычки определяется исключительно податливостью стенок; увеличение объема речевого тракта при возрастании внутриротового давления способствует продлению колебаний голосовых складок; площадь узких щелей увеличивается; возникает эффект запирания для низких частот [59]. В диапазоне звуковых частот импеданс мягких стенок речевого тракта носит инерционный характер, и резонансная частота колебаний стенок с учетом упругости воздуха находится в пределах 150—350 Гц, со средним значением около 200 Гц. Эта частота достаточно низка для того, чтобы считать стенки сосредоточенной системой, т. е. пренебречь различиями фаз в колебаниях стенок вдоль оси речевого тракта.

Рассмотрим систему уравнений для речевого тракта с учетом податливости стенок:

$$
\begin{aligned}
-S\frac{\partial P}{\partial x} &= \rho_0 \frac{\partial U}{\partial t}, \\
-\rho_0 c_0^2 \frac{\partial U}{\partial x} &= S\frac{\partial P}{\partial t} + \rho_0 c_0^2 \frac{\partial S}{\partial t}, \\
S &= y\mathscr{L} + S_0, \\
m\frac{\partial^2 y}{\partial t^2} &+ b\frac{\partial y}{\partial t} + ky = P,
\end{aligned}
\qquad (10.54)
$$

где S_0 — площадь сечения тракта, создаваемая положением артикуляторных органов, $\Delta S = y\mathscr{L}$ — изменение площади поперечного сечения из-за податливости стенок, $\mathscr{L}$ — периметр сечения, y — смещение стенок в сечении, m, b, k — механические параметры стенок в расчете на единицу параметра. Пренебрежем пока скоростью изменения площади сечения $\partial S_0/\partial t$ вследствие движений артикуляторных органов.

Поскольку мы считаем стенки колеблющимися как единое целое, то исключим последнее уравнение из системы (10.54), так как смещение y можно вычислить один раз, взяв, например, давление в фарингиальной области, где все моды акустических колебаний имеют пучность. Если форма поперечного сечения близка к кругу с радиусом r_s, то периметр $\mathscr{L} = 2\pi(r_s + y)$ и $\partial \Delta S/\partial t = 2\pi\partial\left[y(r_s + y)\right]/\partial t$, и при большой площади сечения $r_s \gg y$, так что

$$
\partial \Delta S/\partial t = 2\pi y_t'(r_s + 2y) \approx 2\pi r_s y_t'.
$$

Если форма сечения близка к прямоугольной со сторонами d и h, то $\mathscr{L} = 2(d + h + y)$ и

$$
\partial \Delta S/\partial t = 2y_t'(d + h).
$$

Учитывая, что твердое нёбо и задняя стенка гортани имеют импеданс другого типа, чем мягкие ткани, необходимо уменьшить дополнительную объемную скорость, создаваемую колебаниями стенок, примерно вдвое, так что $\partial \Delta S/\partial t \approx y_t'$. Оценка объемной скорости для кругового сечения представляется завышенной, и мы будем пользоваться в дальнейшем оценкой для прямоугольного сечения. Поскольку ΔS зависит от акустического давления, будем решать систему (10.54) итеративно, полагая на первом шаге давление P равным его значению для предыдущего отсчета времени.

Представляя, как и ранее, $U(x, t) = u^+(t - x/c_0) - u^-(t + x/c_0)$, $P(x, t) = p^+(t - x/c_0) + p^-(t + x/c_0)$, подставим их в первые два уравнения системы (10.54), почленно сложим и вычтем их,

получив

$$\frac{\partial p^+(t-x/c_0)}{\partial x}=\frac{\rho_0 c_0}{S}\left[\frac{\partial u^+(t-x/c_0)}{\partial x}+\frac{1}{2}f(x,\ t)\right],$$

$$\frac{\partial p^-(t+x/c_0)}{\partial x}=\frac{\rho_0 c_0}{S}\left[\frac{\partial u^-(t+x/c_0)}{\partial x}-\frac{1}{2}f(x,\ t)\right],$$

где $f(x,\ t)=\rho_0 c_0^2\partial\Delta S/\partial t$. Для каждой цилиндрической секции, где $S=\text{const}$, эту систему можно проинтегрировать по x:

$$p_i^+(t-\tau)=\frac{\rho_0 c_0}{S_i}\left[u_i^+(t-\tau)+\frac{\Delta x}{2}f_i(t)\right],$$

$$p_i^-(t+\tau)=\frac{\rho_0 c_0}{S_i}\left[u_i^-(t+\tau)-\frac{\Delta x}{2}f_i(t)\right].$$

Разрешая эту систему относительно $u_i^{\pm}$, получим

$$u_i^+(t-\tau)=\frac{S_i}{\rho_0 c_0}p_i^+(t-\tau)-\frac{\Delta x}{2}f_i(t),$$

$$u_i^-(t+\tau)=\frac{S_i}{\rho_0 c_0}p_i^-(t+\tau)+\frac{\Delta x}{2}f_i(t).$$

Акустический поток равен разности полного потока и потока, создаваемого колебаниями стенок

$$U(x,\ t)=\frac{S_i}{\rho_0 c_0}\cdot\left[p_i^+(t-\tau)-p_i^-(t+\tau)-\Delta x f_i(t)\right].$$

Граничные условия между секциями должны учитывать полный поток U, т. е. сумму объемной скорости акустических колебаний и объемной скорости от колебаний стенок:

$$U_i(t,\ \Delta x/2)=U_{i+1}(t,\ -\Delta x/2),$$

т. е.

$$u_i^+(t-\tau)-u_i^-(t+\tau)+\Delta x f_i(t-\tau)=$$
$$=u_{i+1}^+(t+\tau)-u_{i+1}^-(t-\tau)+\Delta x f_{i+1}(t+\tau).$$

Граничные условия по давлению не зависят от потока $\Delta x f$:

$$\frac{p_i^+(t-\tau)+p_i^-(t+\tau)}{S_i}=\frac{p_{i+1}^+(t+\tau)+p_{i+1}^-(t-\tau)}{S_{i+1}}.$$

Рекурсивная схема для объемной скорости есть

$$u_{i+1}^+(t+\tau)=u_i^+(t-\tau)+\mu_i\left[u_i^+(t-\tau)+u_{i+1}^-(t-\tau)\right]-\frac{\Delta x}{2}(1+\mu_i)\Delta f_{i,\,i+1},$$

$$u_i^-(t+\tau)=u_{i+1}^-(t-\tau)-\mu_i\left[u_i^+(t-\tau)+u_{i+1}^-(t-\tau)\right]-\frac{\Delta x}{2}(1-\mu_i)\Delta f_{i,\,i+1},$$

где $\Delta f_{i,\,i+1}=f_{i+1}(t+\tau)-f_i(t-\tau)$. Как уже обсуждалось, из-за

низкой частоты во всех сечениях речевого тракта стенки колеблются синфазно, но даже в случае равенства амплитуд смещений и, соответственно, объемных скоростей f_i и f_{i+1} в соседних секциях, дополнительное возбуждение создается вследствие разницы отсчетов f_i и f_{i+1} во времени, отстоящих друг от друга на 2τ. Решая относительно волн давления и учитывая, что вследствие синфазности колебания стенок $f_i(t)=f_{i+1}(t)$, получим

$$p_{i+1}^{+}(t+\tau)=p_i^{+}(t-\tau)+\mu_i\left[p_{i+1}^{-}(t-\tau)-p_i^{+}(t-\tau)\right]-\frac{\rho_0 c_0 \Delta x}{S_i+S_{i+1}}\Delta f_{i,i+1},$$

$$p_i^{-}(t+\tau)=p_{i+1}^{-}(t-\tau)+\mu_i\left[p_{i+1}^{-}(t-\tau)-p_i^{+}(t-\tau)\right]-\frac{\rho_0 c_0 \Delta x}{S_i+S_{i+1}}\Delta f_{i,i+1}.$$

Завершив первую итерацию расчета бегущих волн, вычислим акустическое давление, где-нибудь в области гортани

$$P(x,\ t)=\frac{\rho_0 c_0}{S(x)}\left[\bar{u}^{+}(t-\tau)+\bar{u}(t+\tau)\right],$$

и воспользуемся этой величиной совместно с давлением, создаваемым постоянным воздушным потоком для коррекции смещения стенок по последнему уравнению системы (10.54). Для получения сходящихся значений f, как правило, требуется всего лишь несколько итераций.

Во время звонкой смычки излучение акустического сигнала в пространство происходит через стенки речевого тракта. Поскольку частота резонанса механических колебаний весьма низкая, то и излучение происходит на низких частотах. Это же относится и к излучению сигнала стенками ноздрей. Поскольку толщина этих стенок в несколько раз меньше, например, толщины щек, то и диапазон излучаемых частот шире, что вносит существенный вклад в излучение из носовой полости. В [153] отмечалось, что если учитывать только излучение из ноздрей, то степень назализации звуков получается недостаточной. Для повышения эффекта назализации в этой работе предлагается учитывать гайморовы полости. В действительности же, канал, соединяющий гайморовы полости с носовыми полостями, обычно заполнен слизью, так что акустическая связь этих полостей затруднена. Если же учитывать излучение стенок ноздрей, то эффект назализации значительно повышается.

Введенная в рекурсивную схему податливость стенок обеспечивает воспроизведение эффекта запирания на низких частотах при малых площадях сечения, если учитывать упругость воздуха в этой области. Для этого, однако, нужно перейти от описания колебаний стенок как системы с сосредоточенными параметрами к описанию распределенных механических колебаний.

§ 10.7. Динамическая схема

Рекурсивная схема Келли—Локбаума была создана в то время, когда считалось, что скорость изменения формы речевого тракта достаточно мала для того, чтобы использовать квазистационарные решения, используя метод «замороженных коэффициентов». Однако схема Келли—Локбаума дает неверное решение даже в случае очень медленного изменения поперечного сечения трубы — в зависимости от того, какие процессы взяты за основу — объемная скорость или давление, получается, что либо объемная скорость постоянна, а давление изменяется, либо давление постоянно, объемная скорость изменяется. На самом же деле должны изменяться и объемная скорость, и давление.

Как показано в ряде работ, скорость изменения формы речевого тракта действительно имеет влияние лишь второго порядка малости на изменения резонансных частот, однако ширина полосы резонансов зависит от этой скорости, и при некоторых условиях может стать отрицательной. Иными словами, вместо затухания резонансных колебаний может произойти их нарастание [126]. Физический смысл этого явления состоит в том, что при отрицательной скорости изменения площади поперечного сечения происходит сужение, сопровождающееся сжатием воздуха и передачей ему дополнительной энергии. Участки с нарастающими колебаниями на интервале закрытой голосовой щели в самом деле изредка наблюдаются в речевом сигнале. Эти явления имеют не только теоретический интерес, поскольку стационарная схема трансформации бегущих волн, не учитывающая изменения площади сечения секций за время пробега ее волной, синтезирует речевой сигнал с шумами, ухудшающими его качество. Причиной этих шумов является несогласование между граничными условиями соседних секций при изменении площади их сечения. Даже в простейшем случае акустической системы, состоящей всего из двух труб с площадями сечения, изменяющимися в противоположных направлениях, возникают искажения сигнала. В экспериментах, выполненных в [146], скорость этих изменений была очень мала, а акустические колебания с помощью специального генератора возбуждались только на частоте первого резонанса акустической системы. Через некоторое время оказалось, что возникли колебания и на частоте следующего резонанса, что свидетельствует об искажении формы акустических колебаний в стационарной рекурсивной схеме.

С целью устранения этих искажений в [60, 146, 151, 194] были предложены различные схемы согласования граничных условий между секциями при изменении площади их сечения. Анализируя способы динамического согласования, оценим сначала относительное изменение площади поперечного сечения тракта за один такт рекурсивной схемы Δt.

Максимальная суммарная скорость кончика языка и нижней челюсти около 20 см/с, так что за интервал времени Δt произойдет смещение его примерно на 3×10^{-4} см. Для фрикативных звуков минимальное расстояние между кончиком языка и нёбом составляет около 0,1 см. Следовательно, максимальное относительное изменение площади поперечного сечения составляет 0,3%. Скорость колебаний голосовых складок находится в диапазоне 50—300 см/с для разных частот основного тона при амплитуде колебаний 0,1—0,2 см. При этом относительное изменение площади голосовой щели за интервал Δt составляет несколько процентов. Скорость колебаний податливых стенок речевого тракта сравнима со скоростью колебаний голосовых складок, но относительное изменение площади поперечного сечения меньше 1% для большинства звуков. Эти, казалось бы, совершенно незначительные изменения площади в соседних секциях приводят к весьма заметным искажениям речевого сигнала, и не только потому, что ошибки накапливаются из-за большого числа секций. Стоит пренебречь динамикой поперечного сечения в районе голосовой щели, и качество речевого сигнала заметно ухудшается. Это связано с чувствительностью конечно-разностных схем к ошибкам в граничных условиях.

При согласовании граничных условий между секциями возможны различные способы дискретизации площади поперечного сечения во времени. Один из них состоит в сохранении площади постоянной в течение всего времени пробега волны, т. е. $S_i = S_i(t-\tau)$, и скачкообразном ее изменении в момент пересечения волной границы между секциями, т. е. $S_i = S_i(t+\tau)$. Это относится и к дополнительной объемной скорости, создаваемой колебаниями стенок. Обозначив

$$\bar{u}^{+}(t+\tau) = u^{+}(t+\tau)e^{-g\tau},$$
$$\bar{u}^{+}(t-\tau) = u^{+}(t-\tau)e^{g\tau},$$

обратимся к граничным условиям (10.8) и (10.9), откуда получим

$$\bar{u}_i^{+}(t-\tau) + \bar{u}_i^{-}(t+\tau) = \bar{u}_{i+1}^{-}(t+\tau) - \bar{u}_{i+1}^{-}(t-\tau), \qquad (10.56)$$

$$\frac{\bar{u}_i^{+}(t-\tau)}{S_i(t-\tau)} + \frac{\bar{u}_i(t+\tau)}{S_i(t+\tau)} = \frac{\bar{u}_{i+1}^{+}(t+\tau)}{S_{i+1}(t+\tau)} + \frac{\bar{u}_{i+1}^{-}(t-\tau)}{S_{i+1}(t-\tau)}. \qquad (10.57)$$

В величинах $f^{\pm}$ теперь будем учитывать и дополнительную объемную скорость, создаваемую вследствие перетекания воздуха при движении артикуляторных органов. Умножив (10.57) на $S_i(t+\tau)$ и сложим его с (10.56), а затем умножим (10.57) на $S_{i+1}(t+\tau)$ и вычтем из него (10.56). Теперь получаем рекурсивную схему для трансформации бегущих волн на границе между i-й и $(i+1)$-й секциями для произвольно меняющейся во времени площади поперечного сечения речевого

тракта:

$$u_{i+1}^{+}(t+\tau)=\frac{S_{i+1}(t+\tau)}{S_{i+1}(t-\tau)}\left[\frac{S_i(t+\tau)}{S_i(t-\tau)}\nu_i\bar{u}_i^{+}(t-\tau)+\mu_{1i}A_i\right]e^{-g_{i+1}\tau}, \quad (10.58)$$

$$u_i^{-}(t+\tau)=\frac{S_i(t+\tau)}{S_i(t-\tau)}\left[\frac{S_{i+1}(t+\tau)}{S_{i+1}(t-\tau)}\nu_i\bar{u}_{i+1}^{-}(t-\tau)-\mu_{2i}A_i\right]e^{-g_i\tau}, \quad (10.59)$$

где

$$A_i=\bar{u}_i^{+}(t-\tau)+\bar{u}_{i+1}^{-}(t-\tau),$$

$$\mu_{1i}=\frac{S_{i+1}(t-\tau)-S_i(t+\tau)}{B_i}, \quad \mu_{2i}=\frac{S_{i+1}(t+\tau)-S_i(t-\tau)}{B_i},$$

$$\nu_i=\frac{S_{i+1}(t-\tau)+S_i(t-\tau)}{B_i}, \quad B_i=S_i(t+\tau)+S_i(t+\tau).$$

Мы видим, что по сравнению со статической схемой (10.12), (10.13) в динамической схеме (10.58), (10.59) появились отношения значений *площади одной и той же секции* в разные моменты времени и, кроме того, в числителях выражений *для коэффициентов отражения* μ_{1i} *и* μ_{2i} *разность площадей* секций также берется для разных моментов времени. Поэтому μ_{1i} отличается от статического коэффициента отражения на $\Delta S_{i+1}/B_i$, а μ_{2i} – на $-\Delta S_i/B_i$, где $\Delta S_i=S_i(t+\tau)-S_i(t-\tau)$ – изменение площади поперечного сечения секции за время $\Delta t=2\tau$. Если же площади секций не меняются во времени, то $\nu_i=1$, $S(t+\tau)/S(t-\tau)=1$ и $\mu_{1i}=\mu_{2i}=\mu_i$.

В речевом тракте встречаются участки, где относительное изменение площади сечения для каждой пары секций одинаково:

$$q_i=\frac{S_i(t+\tau)}{S_i(t-\tau)}=\frac{S_{i+1}(t+\tau)}{S_{i+1}(t-\tau)}=q_{i+1},$$

тогда

$$q_i\nu_i=q_i\frac{S_{i+1}(t-\tau)+S_i(t-\tau)}{q_iS_{i+1}(t-\tau)+q_iS_i(t-\tau)}=1,$$

$$q_{i+1}\nu_i=q_{i+1}\frac{S_{i+1}(t-\tau)+S_i(t-\tau)}{q_{i+1}S_{i+1}(t-\tau)+q_{i+1}S_i(t-\tau)}=1.$$

Это дает возможность упрощения рекурсивной схемы:

$$u_{i+1}^{+}(t+\tau)=\frac{S_{i+1}(t+\tau)}{S_{i+1}(t-\tau)}\left[\bar{u}_i^{+}(t-\tau)+\mu_{1i}A_i\right]e^{-g_{i+1}\tau},$$

$$u_i^{-}(t+\tau)=\frac{S_i(t+\tau)}{S_i(t-\tau)}\left[\bar{u}_{i+1}^{-}(t-\tau)-\mu_{2i}A_i\right]e^{-g_i\tau}.$$

В области голосовой щели также возможно упрощение, исходя из того, что площадь голосовой щели значительно меньше площадей выше- и нижележащих секций. Поэтому $(S_n+S_{n+1})/S_{n+1}\approx 1$ и $(S_n+S_{n-1})/S_{n-1}\approx 1$, где n – номер секции,

приходящейся на голосовую щель. Отсюда получаем для вышележащей секции

$$u_{n+1}^{+}(t+\tau)=\frac{S_n(t+\tau)}{S_n(t-\tau)}\left[\bar{u}_n^{+}(t-\tau)+\mu_{1n}A_n\right]e^{-g_{n+1}\tau},$$

$$u_n^{-}(t+\tau)=\frac{S_n(t+\tau)}{S_n(t-\tau)}\left[\bar{u}_{n+1}^{-}(t-\tau)-\mu_{2n}A_n\right]e^{-g_n\tau},$$

а для нижележащей секции

$$u_n^{+}(t+\tau)=\frac{S_n(t+\tau)}{S_n(t-\tau)}\left[\bar{u}_n^{+}(t-\tau)+\mu_{1n-1}A_{n-1}\right]e^{-g_n\tau},$$

$$u_{n-1}^{-}(t+\tau)=\frac{S_n(t+\tau)}{S_n(t-\tau)}\left[\bar{u}_n^{-}(t-\tau)-\mu_{2n-1}A_{n-1}\right]e^{-g_{n-1}\tau}.$$

Следует, однако, отметить, что такое упрощение все же приводит к некоторому ухудшению качества синтетического речевого сигнала.

Для волн давления, действуя аналогичным образом, получаем

$$p_{i+1}^{+}(t+\tau)=\left[\nu_i\bar{p}_i^{+}(t-\tau)+\mu_{1i}B_i\right]e^{-g_{i+1}\tau}, \qquad (10.60)$$

$$p_i^{-}(t+\tau)=\left[\nu_i\bar{p}_{i+1}^{-}(t-\tau)+\mu_{2i}B_i\right]e^{-g_i\tau}, \qquad (10.61)$$

где

$$\bar{p}^{\pm}(t-\tau)=\left[p^{\pm}(t-\tau)-\frac{\rho_0 c_0\Delta x}{2S(t+\tau)}f(t+\tau)\right]e^{-g\tau},$$

$$B_i=\bar{p}_{i+1}(t-\tau)-\bar{p}_i(t-\tau),$$

а коэффициенты ν_i, μ_{1i} и μ_{2i} — те же, что и для волн объемной скорости. Как видно, рекурсивная схема для трансформации

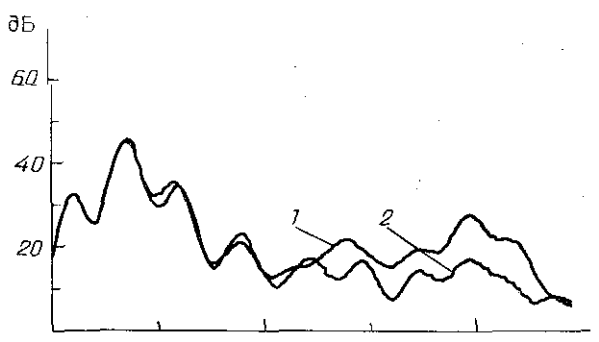


Рис. 10.3. Спектр синтезированного гласного звука /А/: *1* — статическая модель, *2* — динамическая модель

давления (10.60), (10.61) оказывается несколько проще схемы для объемной скорости (10.58), (10.59). Схемы для давления

и объемной скорости обладают разной чувствительностью к упрощению и ошибкам в описании параметров речевого тракта, поэтому для каждой конкретной схемы синтеза желательно сопоставление обеих схем с целью выбора наилучшего варианта.

Применение динамической схемы согласования граничных условий позволяет снизить уровень специфических шумов в высокочастотной области спектра более чем на 10 дб для гласных звуков (см. рис. 10.3), и уменьшает паразитные шумы непосредственно после взрыва смычки согласных звуков [60].

§ 10.8. Источники возбуждения в рекурсивной схеме

Имеющиеся в речевом тракте источники давления и объемной скорости необходимо включить в рекурсивную схему, поскольку исходная систем уравнений (10.1), (10.2) предполагает независимость источников возбуждения и акустических процессов. Источником давления могут служить либо турбулентные процессы, начинающиеся при значениях числа Рейнольдса $\mathrm{Re} > 50$, либо накопленное давление во время смычки.

Пусть источник давления $P_в$ распределен по всему речевому тракту таким образом, что в каждой i-й секции имеется давление $P_{вi}$. Включая это давление в граничные условия

$$u_i(t, \Delta x/2) = u_{i+1}(t, -\Delta x/2),$$

$$P_i(t, \Delta x/2) + P_{вi}(t, \Delta x/2) = P_{i+1}(t, -\Delta x/2) + P_{вi+1}(t, -\Delta x/2),$$

разрешая исходную систему относительно волн давления

$$S_i\left[p_i^+(t-\tau) - p_i^-(t+\tau)\right] = S_{i+1}\left[p_{i+1}^+(t+\tau) - p_{i+1}^-(t-\tau)\right],$$

$$p_i^+(t-\tau) + p_i^-(t+\tau) + P_{вi}(t-\tau) = p_{i+1}^+(t+\tau) + p_{i+1}^-(t-\tau) + P_{вi+1}(t+\tau),$$

и проводя необходимые преобразования, получим

$$\begin{aligned} p_{i+1}^+(t+\tau) &= \\ &= p_i^+(t-\tau) + \mu_i\left[p_{i+1}^-(t-\tau) - p_i^+(t-\tau)\right] - (1-\mu_i)\frac{\Delta P_{вi,\,i+1}}{2}, \\ p_i^-(t+\tau) &= \\ &= p_{i+1}^-(t-\tau) + \mu_i\left[p_{i+1}^-(t-\tau) - p_i^+(t-\tau)\right] + (1+\mu_i)\frac{\Delta P_{вi,\,i+1}}{2}, \end{aligned} \tag{10.62}$$

где $\Delta P_{вi,\,i+1} = P_{вi+1}(t+\tau) - P_{вi}(t-\tau)$.

Для волн объемной скорости имеем

$$u_i^+(t-\tau) - u_i^-(t+\tau) = u_{i+1}^+(t+\tau) - u_{i+1}^-(t-\tau),$$

$$\frac{\rho_0 c_0}{S_i}\left[u_i^+(t-\tau) + u_i^-(t+\tau)\right] + P_{вi}(t-\tau) =$$

$$= \frac{\rho_0 c_0}{S_{i+1}}\left[u_{i+1}^+(t-\tau) + u_{i+1}^-(t+\tau)\right] + P_{вi+1}(t+\tau),$$

откуда

$$u_{i+1}^{+}(t+\tau)=u_i^{+}(t-\tau)+\mu_i\left[u_i^{+}(t-\tau)+u_{i+1}^{-}(t-\tau)\right]-\\-\frac{S_i}{2\rho_0 c_0}(1+\mu_i)\,\Delta P_{\text{вi},i+1},$$

$$u_i^{-}(t+\tau)=u_{i+1}^{-}(t-\tau)-\mu_i\left[u_i^{+}(t-\tau)+u_{i+1}^{-}(t-\tau)\right]+\\+\frac{S_i}{2\rho_0 c_0}(1-\mu_i)\,\Delta P_{\text{вi},i+1}, \qquad (10.63)$$

или, используя импедансы секций $Z_i=\rho_0 c_0/S_i$, $Z_{i+1}=\rho_0 c_0/S_{i+1}$,

$$u_{i+1}^{+}(t+\tau)=u_i^{+}(t-\tau)+\mu_i\left[u_i^{+}(t-\tau)+u_{i+1}^{-}(t-\tau)\right]-\frac{\Delta P_{\text{вi},i+1}}{Z_i+Z_{i+1}},$$

$$u_i^{-}(t+\tau)=u_{i+1}^{-}(t-\tau)-\mu_i\left[u_i^{+}(t-\tau)+u_{i+1}^{-}(t-\tau)\right]+\frac{\Delta P_{\text{вi},i+1}}{Z_i+Z_{i+1}}.$$

Из (10.62) и (10.63) видно, что если источник давления распределен равномерно и давление не меняется во времени, то перепад давления между соседними секциями отсутствует, т. е. $\Delta P_{\text{вi},i+1}=0$ и не возникает никакого возбуждения ни волн давления, ни волн объемной скорости. При превышении числом Рейнольдса порога 1800 возникает турбулентный шум $P_{\text{т}}$, сосредоточенный на выходе из сужения в расширяющуюся область речевого тракта.

Поскольку формулы для давления в турбулентном потоке полуэмпирические, при синтезе речи необходимо добиваться надлежащего качества звуков путем подбора коэффициентов при $P_{\text{т}}$ и $P_{\text{в}}$.

В момент взрыва смычки на короткое время возникает перепад давления между секцией, где площадь сечения была равна нулю, и соседней секцией с ненулевой площадью. Этот перепад и служит импульсным источником возбуждения для взрывных согласных.

Для источника давления аналогичные выражения получаются и в том случае, если принять, что при возникновении турбулентных шумов давление в речевом тракте равно сумме акустического давления P и давления турбулентного источника $P_{\text{т}}$. Тогда система дифференциальных уравнений, связывающая давление и объемную скорость в речевом тракте, выглядит как

$$-S\frac{\partial(P+P_{\text{т}})}{\partial x}=\rho_0\frac{\partial U}{\partial t},\qquad -S\frac{\partial(P+P_{\text{т}})}{\partial t}=\rho_0 c_0^2\frac{\partial U}{\partial x}.$$

Действуя, как и прежде, получим

$$p_i^{+}(t-\tau)=\frac{\rho_0 c_0}{S_i}u_i^{+}(t-\tau)-\frac{P_{\text{т}}(t-\tau)}{2}-\frac{1}{2c_0}\int\frac{\partial P_{\text{т}}(t-\tau)}{\partial t}\,dx,$$

$$p_i^{-}(t+\tau)=\frac{\rho_0 c_0}{S_i}u_i^{-}(t+\tau)-\frac{P_{\text{т}}(t+\tau)}{2}+\frac{1}{2c_0}\int\frac{\partial P_{\text{т}}(t+\tau)}{\partial t}\,dx.$$

Допустим теперь, что давление от турбулентного источника распределено равномерно по i-й секции и изменяется в ней синфазно. Тогда

$$p_i^+(t-\tau)=\frac{\rho_0 c_0}{S_i}u_i^+(t-\tau)-\frac{P_т(t)}{2}-\frac{\Delta x_i}{2c_0}\frac{\partial P_т(t)}{\partial t},$$

$$p_i^-(t+\tau)=\frac{\rho_0 c_0}{S_i}u_i^-(t+\tau)-\frac{P_т(t)}{2}+\frac{\Delta x_i}{2c_0}\frac{\partial P_т(t)}{\partial t}.$$

Если длина секции Δx_i достаточно мала, то вполне можно принять, что давление шума $P_т$ в каждой паре соседних секций i и $i+1$ одинаково. Поэтому в граничных условиях

$$u_i^+(t-\tau)-u_i^-(t+\tau)=u_{i+1}^+(t+\tau)-u_{i+1}^-(t-\tau),$$

$$\frac{\rho_0 c_0}{S_i}\left[u_i^+(t-\tau)+u_i^-(t+\tau)\right]-\frac{P_т(t)}{2}-\frac{\Delta x_i}{2c_0}\frac{\partial P_т(t)}{\partial t}=$$

$$=\frac{\rho_0 c_0}{S_{i+1}}\left[u_{i+1}^+(t+\tau)+u_{i+1}^-(t-\tau)\right]-\frac{P_т(t)}{2}+\frac{\Delta x_i}{2c_0}\frac{\partial P_т(t)}{\partial t},$$

от источника шумового давления остается только производная по времени $\partial P_т/\partial t$. В результате получаем

$$u_{i+1}^+(t+\tau)=u_i^+(t-\tau)+\mu_i\left[u_i^+(t-\tau)+u_{i+1}^-(t-\tau)\right]-\frac{\Delta x_i S_i}{2\rho_0 c_0^2}(1+\mu_i)\frac{\partial P_т}{\partial t},$$

$$u_i^-(t+\tau)=u_{i+1}^-(t-\tau)-\mu_i\left[u_i^+(t-\tau)+u_{i+1}^-(t-\tau)\right]+\frac{\Delta x_i S_i}{2\rho_0 c_0^2}(1-\mu_i)\frac{\partial P_т}{\partial t}.$$

Учитывая, что в конечно-разностной форме

$$\frac{\partial P_т}{\partial t}=\frac{P_т(t+\tau)-P_т(t-\tau)}{\Delta t},$$

и $\Delta t=\Delta x/c_0$, видим, что полученные выражения совпадают с (10.63).

Если имеется распределенный источник объемной скорости $U_в$, то из граничных условий

$$U_i(t,\ \Delta x/2)+U_{вi}(t,\ \Delta x/2)=U_{i+1}(t,\ -\Delta x/2)+U_{вi}(t,\ -\Delta x/2),$$

$$P_i(t,\ \Delta x/2)=P_{i+1}(t,\ -\Delta x/2),$$

для волн объемной скорости имеем

$$u_i^+(t-\tau)-u_i^-(t+\tau)+U_{вi}(t-\tau)=$$

$$=u_{i+1}^+(t+\tau)-u_{i+1}^-(t-\tau)+U_{вi+1}(t+\tau),$$

$$\frac{u_i^+(t-\tau)+u_i^-(t+\tau)}{S_i}=\frac{u_{i+1}^+(t+\tau)+u_{i+1}^-(t-\tau)}{S_{i+1}},$$

что дает рекурсивную схему

$$u_{i+1}^{+}(t+\tau)=u_i^{+}(t-\tau)+\mu_i\left[u_i^{+}(t-\tau)+u_{i+1}^{-}(t-\tau)\right]-$$
$$-(1+\mu_i)\frac{\Delta U_{\text{в}i,\,i+1}}{2},$$

$$u_i^{-}(t+\tau)=u_{i+1}^{+}(t-\tau)-\mu_i\left[u_i^{+}(t-\tau)+u_{i+1}^{-}(t-\tau)\right]-$$
$$-(1-\mu_i)\frac{\Delta U_{\text{в}i,\,i+1}}{2}, \qquad (10.64)$$

где $\Delta U_{\text{в}i,\,i+1}=U_{\text{в}i+1}(t+\tau)-U_{\text{в}i}(t-\tau)$ — изменение производительности объемного источника при переходе из одной секции в другую. Аналогично для волн давления имеем

$$\frac{S_i}{\rho_0 c_0}\left[p_i^{+}(t-\tau)-p_i^{-}(t+\tau)\right]+U_{\text{в}i}(t-\tau)=$$
$$=\frac{S_{i+1}}{\rho_0 c_0}\left[p_{i+1}^{+}(t+\tau)-p_{i+1}^{-}(t-\tau)\right]+U_{\text{в}i+1}(t+\tau),$$

$$p_i^{+}(t-\tau)+p_i^{-}(t+\tau)=p_{i+1}^{+}(t+\tau)+p_{i+1}^{-}(t-\tau),$$

откуда

$$p_{i+1}^{+}(t+\tau)=p_i^{+}(t-\tau)+\mu_i\left[p_{i+1}^{-}(t-\tau)-p_i^{-}(t-\tau)\right]-$$
$$-\frac{\rho_0 c_0}{S_i+S_{i+1}}\Delta U_{\text{в}i,\,i+1},$$

$$p_i^{-}(t+\tau)=p_{i+1}^{-}(t-\tau)+\mu_i\left[p_{i+1}^{-}(t-\tau)-p_i^{-}(t-\tau)\right]-$$
$$-\frac{\rho_0 c_0}{S_i+S_{i+1}}\Delta U_{\text{в}i,\,i+1}, \qquad (10.65)$$

или, используя проводимости секций, $Y_i=1/Z_i$

$$p_{i+1}^{+}(t+\tau)=p_i^{+}(t-\tau)+\mu_i\left[p_{i+1}^{-}(t-\tau)-p_i^{+}(t-\tau)\right]-\frac{\Delta U_{\text{в}i,\,i+1}}{Y_i+Y_{i+1}},$$

$$p_i^{-}(t+\tau)=p_{i+1}^{-}(t-\tau)+\mu_i\left[p_{i+1}^{-}(t-\tau)-p_i^{+}(t-\tau)\right]-\frac{\Delta U_{\text{в}i,\,i+1}}{Y_i+Y_{i+1}}.$$

Схемы (10.64) и (10.65) отражают тот факт, что при установившемся течении воздушного потока, когда объемная скорость в каждом сечении речевого тракта одинакова и постоянна во времени, не возникает никакого возбуждения акустических колебаний, так как $\Delta U_{\text{в}i,\,i+1}=0$ для любой пары соседних секций.

Объемная скорость воздушного потока, протекающего по речевому тракту, может быть различной на разных участках при изменении формы речевого тракта во времени. При этом возникают неустановившиеся процессы, которые и порождают источники возбуждения с интенсивностью, зависящей от скорости изменения поперечного сечения. Наиболее интенсивный

источник объемной скорости — это, конечно, голосовая щель, площадь которой меняется очень быстро. Однако и артикуляторные движения, и колебания стенок тракта также создают источники объемной скорости, причем эти источники не сосредоточены, а распределены по всему речевому тракту. Импульсный источник возбуждения, возникающий при взрыве смычки, также может рассматриваться как источник объемной скорости, поскольку падение давления в тракте сопровождается истечением воздуха с быстро меняющейся скоростью.

Рассматривая частный случай голосового источника, можно пренебречь изменением объемной скорости потока в секциях непосредственно под голосовой щелью и над ней по сравнению с изменением объемной скорости в самой голосовой щели. Тогда для секции $n+1$ (над голосовой щелью) приращение объемной скорости есть

$$\Delta U_{вп+1} = U_{вп+1}(t+\tau) - U_{вп}(t-\tau) \approx U_{вп}(t-\tau) - U_{вп}(t+\tau).$$

Поскольку вследствие инерционности воздушного потока объемная скорость в $(n+1)$-й секции в момент $t+\tau$ остается почти такой же, как и объемная скорость в голосовой щели в момент $t-\tau$, т. е. $U_{вп+1}(t+\tau) \approx U_{вп}(t-\tau)$, то

$$\Delta U_{вп} = U_{вп}(t+\tau) - U_{вп-1}(t+\tau) \approx U_{вп}(t+\tau) - U_{вп}(t-\tau),$$

т. е. источник возбуждения равен разности объемной скорости в голосовой щели в моменты $t+\tau$ и $t-\tau$.

Поршневой источник объемной скорости, появляющийся в результате вертикальных колебаний голосовых складок, существует только в двух секциях: непосредственно перед голосовой щелью и сразу за ней. В этих секциях поршневой источник действует в противофазе, поскольку положительная объемная скорость, создаваемая в секции $n+1$, сопровождается отрицательной объемной скоростью в секции $n-1$, где n — номер секции с голосовой щелью (полагая, что голосовая щель занимает ровно одну секцию). Амплитуды объемной скорости от поршневого источника равны в обеих секциях, т. е. $U_{пп-1} = -U_{пп+1}$. Поскольку поршневые источники сосредоточены в двух секциях, разделенных между собой голосовой щелью, то для секций $n-2$, $n-1$, n, $n+1$ и $n+2$ рекурсивная схема принимает несколько иной вид:

$$u_{i+1}^{+}(t+\tau) = u_i^{+}(t-\tau) + \mu_i\left[u_i^{+}(t-\tau) + u_{i+1}^{-}(t-\tau)\right] \mp (1+\mu_i)\frac{U_п(t\pm\tau)}{2},$$

$$u_i^{+}(t+\tau) = u_{i+1}^{+}(t-\tau) - \mu_i\left[u_i^{+}(t-\tau) + u_{i+1}^{-}(t-\tau)\right] \mp (1-\mu_i)\frac{U_п(t\pm\tau)}{2},$$

где знак минус перед последним членом ставится для секций $i=n+1$ и $i=n-2$, а плюс — для секций $i=n$ и $i=n-1$. Знак минус при τ для $U_п$ ставится для секций $i=n$ и $i=n-1$, а плюс — для секций $i=n+1$ и $i=n-2$.

Подведем итоги анализа различных источников возбуждения в рекурсивных схемах для бегущих волн и запишем результирующую форму, включающую все источники, в том числе и колеблющиеся стенки речевого тракта. Для статической системы и волн объемной скорости получаем

$$u_{i+1}^{+}(t+\tau)=\left[u_i^{+}(t-\tau)e^{-g_i\tau}+\mu_i\Delta u_{i,\,i+1}-\frac{1+\mu_i}{2}F_1\right]e^{-g_{i+1}\tau},$$

$$u_i^{-}(t+\tau)=\left[u_{i+1}^{-}(t-\tau)e^{-g_{i+1}\tau}-\mu_i\Delta u_{i,\,i+1}-\frac{1-\mu_i}{2}F_2\right]e^{-g_i\tau}, \quad (10.66)$$

$$\Delta u_{i,\,i+1}=u_i^{+}(t-\tau)e^{-g_i\tau}+u_{i+1}^{-}(t-\tau)e^{-g_{i+1}\tau},$$

где

$$F_1=\Delta x\Delta f_{i,\,i+1}+\Delta U_{\text{вi},\,i+1}-\frac{S_i}{\rho_0 c_0}\Delta P_{\text{вi},\,i+1}+U_{\text{пi}}(t+\tau),$$

$$F_2=\Delta x\Delta f_{i,\,i+1}+\Delta U_{\text{вi},\,i+1}-\frac{S_{i+1}}{\rho_0 c_0}\Delta P_{\text{вi},\,i+1}+U_{\text{пi}}(t+\tau),$$

$$\Delta f_{i,\,i+1}=f_{i+1}(t+\tau)e^{-g_{i+1}\tau}-f_i(t-\tau)e^{-g_i\tau},$$

другие функции аналогичны по потерям, а условия вхождения поршневого источника обсуждались выше.

Соответственно, для волн акустического давления имеем

$$p_{i+1}^{+}(t+\tau)=\left[p_i^{+}(t-\tau)e^{-g_i\tau}+\mu_i\Delta p_{i,\,i+1}-\right.$$
$$\left.-\frac{1-\mu_i}{2}\Delta P_{\text{вi},\,i+1}-G_1\frac{\rho_0 c_0}{S_i+S_{i+1}}\right]e^{-g_{i+1}\tau},$$

$$p_i^{-}(t+\tau)=\left[p_{i+1}^{-}(t-\tau)e^{-g_{i+1}\tau}+\mu_i\Delta p_{i,\,i+1}+\right.$$
$$\left.+\frac{1+\mu_i}{2}\Delta P_{\text{вi},\,i+1}-G_1\frac{\rho_0 c_0}{S_i+S_{i+1}}\right]e^{-g_i\tau},$$

где

$$\Delta p_{i,\,i+1}=p_{i+1}^{-}(t-\tau)e^{-g_{i+1}\tau}-p_i^{-}(t-\tau)e^{-g_i\tau},$$
$$G_1=\Delta x\Delta f_{i,\,i+1}+\Delta U_{\text{вi},\,i+1}+U_{\text{пi}}(t+\tau).$$

Поскольку смещение стенок тракта происходит под воздействием как акустического давления, так и внешних источников, сначала рассчитывается давление с учетом всех факторов, а затем итеративно вводится влияние податливых стенок.

Вследствие достаточно быстрого колебания стенок и голосовых складок для внешних источников также необходимо применить динамическую схему согласования граничных условий. Поэтому в общем виде схема трансформации бегущих

волн для объемной скорости есть:

$$u_{i+1}^{+}(t+\tau)=\left\{q_{i+1}\left[q_i\nu_i u_i^{+}(t-\tau)e^{-g_i\tau}+\mu_{1i}\Delta u_{i,\,i+1}\right]-\right.$$
$$\left.-\frac{S_{i+1}(t+\tau)}{S_i(t+\tau)+S_{i+1}(t+\tau)}F_2\right\}e^{-g_{i+1}\tau},$$

$$u_i^{-}(t+\tau)=\left\{q_i\left[q_{i+1}\nu_i u_{i+1}^{-}(t-\tau)e^{-g_{i+1}\tau}-\mu_{2i}\Delta u_{i,\,i+1}\right]-\right.$$
$$\left.-\frac{S_i(t+\tau)}{S_i(t+\tau)+S_{i+1}(t+\tau)}F_2\right\}e^{-g_i\tau},$$

где

$$q_i=S_i(t+\tau)/S_i(t-\tau),$$

ν_i, μ_{1i}, μ_{2i} определены в предыдущем разделе, а $\Delta u_{i,\,i+1}$, F_1 и F_2 — те же, что и в статической схеме, причем для члена с турбулентным источником давления площади S_i и S_{i+1} берутся в момент времени $t+\tau$.

§ 10.9. Асинхронная рекурсивная схема

Длина речевого тракта непрерывно изменяется в процессе артикуляции не только из-за подъема — опускания гортани, но также и вследствие изменения формы тракта, что, как мы видели в гл. 8, приводит к изменению длины средней линии тракта. В рекурсивной схеме с фиксированной длиной секций Δx изменение длины речевого тракта остается незамеченным до тех пор, пока она не превысит Δx. При этом погрешность представления длины l речевого тракта $\varepsilon=\Delta x/l$. Следовательно, и погрешность акустических параметров в синтезируемом речевом сигнале велика. Например, при $\Delta x=1$ см и $l=17{,}5$ см, $\varepsilon\approx 5{,}5\%$ и даже при наиболее часто применяющейся величине $\Delta x=0{,}5$ см ε оказывается около 2—3%. Еще более неприятные последствия вызывают акустические возмущения, связанные с внезапным добавлением или исчезновением секции, поскольку оно соответствует скачкообразному, т. е. бесконечно быстрому изменению длины речевого тракта. В результате этого синтезированный речевой сигнал искажается, приобретая хриплый оттенок.

Один из простейших способов оперирования с переменной длиной секций использует связь между длиной секции l и частотой отсчетов речевого сигнала на выходе из тракта $f_s=c_0/2\Delta x$. При малых изменениях длины секции Δx изменение характеристик речевого сигнала можно моделировать путем изменения частоты отсчетов [206]. Если назначить некоторую среднестатическую длину секции $\Delta\bar{x}$, и вычислить по рекурсивной схеме дискретные отсчеты речевого сигнала $P(t_i)$ с интервалами времени $\Delta\bar{t}=1/f_s$, то при отклонении длины

секции на δ_x сначала путем интерполяции восстанавливают непрерывную форму речевого сигнала $P(t)$, а затем берут отсчеты с новым интервалом $\Delta t = 1/2(\Delta x + \delta_x)$ по формуле

$$P(m) = \sum_{n=N_1}^{N_2} P(n) \frac{\sin[\pi(m\Delta t - n\Delta\bar{t})]/\Delta\bar{t}}{\pi(m\Delta t - n\Delta\bar{t})/\Delta\bar{t}} W(m\Delta t - n\Delta\bar{t}),$$

где W — сглаживающее окно. Оценка искажений речевого сигнала показала возможность использования W длиной всего лишь в 5—6 отсчетов. При этом искажения, в основном, возникают на высоких частотах, а ниже 2,5 кГц искажается форма впадин между формантами в спектре речевых сигналов. Эти оценки, однако, сделаны лишь для стационарных звуков, с использованием моногармонического возбуждения, тогда как в динамике речеобразования искажения могут оказаться больше. Кроме того, требование относительной малости изменения длины каждой секции препятствует уменьшению числа секций при переходе к асинхронной рекурсивной схеме, где длины секций различны. Таким образом, и изменение длины речевого тракта, и необходимость снижения вычислительных затрат заставляют рассмотреть возможность модификации рекурсивной схемы бегущих волн с включением секций различной длины.

Пусть площадь поперечного сечения речевого тракта $S(x, t)$ аппроксимируется M цилиндрическими секциями неравной длины с координатами $j = 1, \ldots, M$. Зададим фиксированный интервал сдвига бегущих волн за один такт Δt, $\Delta x = c_0/\Delta t$. Очевидно, что в общем случае координаты границ секций x_j и координаты бегущих волн x_i не совпадают. Из рис. 10.4 видно, что не совпадают и координаты волн, бегущих в положительном и отрицательном направлениях. Более того, координаты бегущих волн смещаются от одного такта к другому. Действительно, (+)-волна, не дошедшая до конца речевого тракта на расстояние $\delta_2^+ x < \Delta x$, после отражения займет положение с координатой $l - \delta_1^- x$, где l — длина речевого тракта, $\delta_1^- x = \Delta x - \delta_2^+ x$. В свою очередь, (−)-волна, не дойдя до начала тракта на $\delta_2^+ x < \Delta x$, отразится с координатой $\delta_1^+ x = \Delta x - \delta_2^- x$. Начиная первый цикл с $\delta_1^+ x = 0$, получим $\delta_1^+ x \neq 0$, как только волна, отраженная от конца речевого тракта, вернется к его началу, так как в общем случае $\delta_2^+ x \neq \delta_1^- x$. По сравнению с синхронной схемой, где координаты бегущих волн фиксированы и совпадают с границами секций, асинхронная схема

Рис. 10.4. Сдвиг бегущих волн в асинхронной схеме

требует вычисления координат бегущих волн после каждого сдвига. Усложнения, однако, оказываются не слишком большими, поскольку эти вычисления довольно просты. Их алгоритм следующий:

$$
\begin{aligned}
&1.\ \delta_1^+ x = 0,\\
&2.\ \delta_2^+ x = \Delta x - \delta_1^+ x,\\
&3.\ \delta_1^- x = l - \left\lfloor \frac{l - \delta_1^+ x}{\Delta x} \right\rfloor - \delta_1^+ x,\\
&4.\ \delta_2^- x = l - \left\lfloor \frac{l - \delta_1^- x}{\Delta x} \right\rfloor - \delta_1^- x,\\
&5.\ \delta_1^+ x = \Delta x - \delta_2^- x,
\end{aligned}
$$

где знак $\lfloor\ \rfloor$ обозначает взятие целой части. После того, как найдены $\delta_1^+ x$, $\delta_2^+ x$, $\delta_1^- x$ и $\delta_2^- x$, правые координаты бегущих волн вычисляются, как

$$
\begin{aligned}
x_i^+ &= \delta_1^+ x + \Delta x(i-1),\\
x_i^- &= \delta_2^- x + \Delta x(i-1).
\end{aligned}
$$

По-прежнему считая, что распространение внутри цилиндрической секции сопровождается лишь задержкой и затуханием волн, трансформацию бегущих волн производим лишь на границе между j-й и $(j+1)$-й секциями при условии $x_j^+ \leqslant x_j \leqslant x_j^-$, где $i = 1, \ldots, N$; $N = \left\lfloor \frac{l - \delta_1^- (x)}{\Delta x} \right\rfloor$. Для того чтобы избежать смещения координат бегущих волн относительно начала речевого тракта, можно проквантовать его длину l, выбрав достаточно малый интервал сдвига волн Δx.

Расстояние от границы между j-й и $(j+1)$-й секциями до ближайшей координаты $(+)$-волны x_i^+ обычно не равно расстоянию до координаты соответствующей $(-)$-волны x_i^-. Это значит, что границу между секциями эти волны пересекают не одновременно. Простейшим решением этой проблемы является сдвиг границы между секциями в таком направлении, чтобы она заняла положение точно посередине между x_i^+ и x_i^-. Другое решение состоит в интерполяции бегущих волн вдоль речевого тракта и взятии таких отсчетов, которые находились бы на равном расстоянии от границы. Представляется, однако, что первый — простейший, способ обеспечивает достаточно хорошее приближение, так как погрешность для каждой границы не больше $\pm \Delta x/2$ и эта погрешность не накапливается вследствие некоррелированности разностей $x_i - x_j$ для разных секций.

Асинхронная вычислительная схема для бегущих волн дает возможность по-новому подойти к аппроксимации функции площади поперечного сечения речевого тракта $S(x, t)$ цилиндрическими секциями. При использовании равноотстоящих

отсчетов от S ошибка аппроксимации зависит от скорости изменения S вдоль пространственной координаты, т. е. от $\partial S/\partial x$, и эта ошибка никак не регулируется. Если же потребовать, чтобы коэффициент отражения между соседними секциями был постоянной величиной $\bar{\mu}$, то тогда длины секций будут различными и их число находится в обратной зависимости от $\bar{\mu}$. Ошибка аппроксимации S при этом становится зависимой от величины площади — для меньших площадей абсолютная погрешность меньше, а для бóльших — больше. Из

$$(S_{i+1}-S_i)/(S_{i+1}+S_i)=\bar{\mu}=\text{const},$$

видно, что $\Delta S=S_{i+1}-S_i$ пропорциональна $S_{i+1}+S_i$, поэтому при меньших площадях отсчеты от S будут браться чаще, чем при больших площадях. Конечно, число отсчетов зависит и от скорости изменения S вдоль пространственной координаты.

Такой способ аппроксимации площади сечения S более соответствует свойствам акустической системы, поскольку известно, что одно и то же изменение акустических характеристик, например, резонансных частот, требует меньшего изменения S при малых площадях, и большего изменения — при больших. Определив таким образом способ аппроксимации, мы достигаем примерно равного изменения характеристик речевого сигнала при переходе от одной секции к другой на протяжении всего речевого тракта. Алгоритм цилиндрической аппроксимации тогда выглядит следующим образом: $(i+1)$-й отсчет от S берется при такой координате x, где $S(x)\geqslant S_i+\bar{\mu}\times$ $\times[S(x)+S_i]$, S_i — последний отсчет. Знак $\geqslant$ вместо строгого равенства используется потому, что в речевом тракте могут встретиться участки с разрывами S, где конечная разность $S(x)-S_i$ может оказаться большей, чем $\bar{\mu}[S(x)-S_i]$. В этом месте и коэффициент отражения μ_i будет больше $\bar{\mu}$.

На неподвижных участках тракта, например, в носовой полости, постоянный коэффициент отражения $\bar{\mu}$ не только позволяет уменьшить число секций, но и делает ненужным вычисление коэффициентов отражения $\bar{\mu}_i$, экономя, таким образом, вычислительные ресурсы. Там, где площадь S меняется во времени, аппроксимация производится, исходя из условия

$$\mu_{1i}=\bar{\mu}=\frac{S_{i+1}(t+\tau)-S_i(t+\tau)}{S_{i+1}(t+\tau)+S_i(t+\tau)}=\text{const},$$

тогда как второй коэффициент может оказаться различным для разных секций, но и в этом случае сложность вычислений уменьшается.

ГЛАВА 11

ТЕСТИРОВАНИЕ СИНТЕТИЧЕСКОЙ РЕЧИ

§ 11.1. Восприятие речи человеком

Восприятие синтетической речи человеком определяется теми же законами, что и восприятие естественной речи, но искусственная природа синтетической речи вносит в процесс восприятия некоторые особенности. Готовность человека пользоваться синтезатором и даже степень доверия к сообщениям синтезатора зависят от трудности понимания синтетической речи, скорости передачи информации, натуральности и других факторов. Оценка качества синтезатора является многомерной переменной, где различные факторы нелинейно взаимодействуют, могут взаимно компенсироваться и менять степень своей важности в зависимости от условий восприятия, т. е. от конкретной задачи. Если условия задачи максимально широки, то мерой качества синтетической речи должна служить близость к характеристикам естественной речи. Для того чтобы иметь возможность оценить степень этой близости, нужно установить основные факторы, влияющие на восприятие речи, создать соответствующие методы тестирования. Известно, что дикторы радио и телевидения подвергаются испытаниям (хотя и субъективным) на качество речи. Тем более важно оценить качество того или иного синтезатора, и тем самым определить область его наиболее эффективного использования. Адекватные методы тестирования позволяют не только оценить качество уже готового синтезатора, но и достичь его оптимального значения путем управления параметрами в процессе разработки синтезатора. Рассмотрим последовательно основные факторы, влияющие на восприятие речи.

Сложность задачи. Вычислительные ресурсы мозга ограничены, и их приходится распределять в соответствии с некоторыми критериями. Распознавание речи для человека может быть лишь одной из нескольких одновременно решаемых задач, и сложность понимания речи также может быть неодинаковой. Поэтому степень распознавания или понимания речи человеком зависит от той доли ресурсов мозга, которая выделена на решение этой задачи, т. е. от сложности текущей деятельности человека.

Сложность текущей деятельности определяется множеством сообщений и реакций на них, степенью неопределенности задачи, степенью предсказуемости сообщений, наличием других источников информации (как речевых, так и неречевых), присутствием шумов, необходимостью выполнения каких-то действий и степенью сложности этих действий.

Отсюда следует, что сложность задачи восприятия речи совершенно различная для слепого человека, слушающего рассказ, и для пилота низколетящего самолета или космонавта в процессе маневра стыковки с ручным управлением. В результате отвлечения ресурсов мозга на другие задачи восприятия и мышления, на решение задачи понимания речевого сообщения выделяется меньший объем кратковременной памяти, и даже этот объем временами может сводиться к нулю. Вследствие этого речевое сообщение воспринимается кусками, при выпадении ключевых слов оно не может быть понято, и даже сам факт наличия речевого сообщения может пройти мимо сознания. В таких условиях разборчивость речи должна быть предельно высокой с тем, чтобы по воспринятым кускам сообщения можно было восстановить его смысл. Если же внимание слушателя целиком сосредоточено на восприятии речи и к тому же у него имеется сильный стимул к пониманию речевого сообщения, то допустимо некоторое снижение качества синтетической речи. Чем выше сложность задачи, тем более избыточной и предсказуемой должна быть структура речевого сообщения и выше разборчивость ключевых слов.

Социальный опыт и психолингвистический тип слушателя влияют на способность анализировать состав речевого сообщения. Люди с ограниченными жизненными интересами пользуются менее аналитическими операциями, для них минимальная единица речи — не фонема, а более крупные блоки, в том числе слоги «согласный — гласный». Это влияет на понимаемость речевого сообщения, поскольку в диалоговой речи важнее различимость слогов, а в контекстной речи (описаниях, инструкциях) важнее фонемная различимость [9]. Грамотность делает человека осведомленным о существовании наименьших смыслоразличительных единиц в речевом потоке — фонем — и, таким образом, придает способность к различению тонких признаков речи, т. е. повышает его возможности в понимании смысла речевого сообщения.

Понимаемость зависит также и от степени владения языком, в том числе от диалектных особенностей. Сообщения формантного синтезатора воспринимаются хуже людьми, для которых язык не является родным по сравнению с людьми, владеющими этим языком с детства [110]. Отсюда следует, что фактическая разборчивость будет хуже, а трудности понимания — выше, если синтезатор используется в обществе, где в большом числе присутствуют носители диалектов и других языков, поскольку все тесты на разборчивость

рассчитаны на языковую компетентность членов аудиторской бригады, т. е. на интуитивное знание акустико-лингвистических закономерностей языка.

Объем кратковременной памяти, как уже упоминалось, является одним из основных ресурсов мозга, выделяемых на решение текущих задач. Этот фактор ограничивает возможности запоминания речевого сигнала. Так, установлено, что сосредоточение внимания на речи одного диктора значительно снижает способность к обнаружению конкретных слов или фонем в речи другого диктора [41]. Опытные лекторы знают, что в речевых сообщениях с насыщенной информацией нужно делать паузы для того, чтобы слушатели успели понять сказанное. По-видимому, речевой сигнал хранится в кратковременной памяти до тех пор, пока длится его осмысление, и для нового речевого сообщения просто не остается места. В задачах на запоминание списка слов последние слова вспоминаются лучше, чем первые, причем забываемость первых слов для синтетической речи существенно больше, чем для естественной речи [41]. Степень понимания и точность транскрибирования длинных фраз в синтетической речи хуже, чем для коротких фраз. Это означает, что обработка синтетической речи в системе восприятия представляет большие трудности, требует большего объема кратковременной памяти, речевые сигналы дольше задерживаются в памяти. В ситуациях, где одновременно с восприятием синтетической речи человек получает информацию от других источников и должен выполнять какие-то действия, возникает конкуренция за активно используемый объем оперативной памяти. Если для восприятия синтетической речи требуется больший объем памяти, чем для восприятия естественной речи, то сосредоточение на речевом сообщении может ухудшить восприятие информации от других источников и снизить эффективность решения текущей задачи, например, управления самолетом.

Мотивация к восприятию речи, создаваемая внутренними стимулами, инструкциями или условиями производственной деятельности, изменяет распределение ресурсов мозга, в том числе и объем оперативной памяти. Высокий уровень мотивации, т. е. стремление понять речевое сообщение, способствует повышению надежности распознавания речи, тогда как низкая заинтересованность может привести к полному игнорированию сообщения.

Тренированность слушателя к особенностям речи сильно влияет на надежность распознавания. Родители лучше распознают речь своих детей, чем посторонние. Речь некоторых людей настолько отличается от стандартного произношения, что к ней нужно привыкать иногда в течение длительного времени. Если задача восприятия речи необычна (например, прием артикуляционных таблиц), то даже в нормальных условиях, без помех и искажений, когда диктор и аудитор

находятся рядом на расстоянии одного метра, требуется более десяти дней для достижения устойчивого высокого уровня распознавания (см. § 11.3). Способность к пониманию иностранного языка также растет в процессе тренировки.

Восприятие синтетической речи требует тренировки, тем более длительной, чем хуже качество синтезатора, причем имеются некоторые особенности, отличающиеся от процессов тренировки к восприятию естественной речи. Эти особенности будут обсуждаться в § 11.3.

Условия восприятия включают в себя громкость речи, уровень и вид помех (случайный шум, реверберация помещений, мешающие разговоры), искажения в канале связи (ограничения частотного диапазона сверху или снизу, деформация амплитудных соотношений), степень умственной нагрузки. Последний фактор мы обсуждали выше, рассматривая распределение ресурсов мозга и, в частности, объема кратковременной памяти. Очень тихая речь обладает низкой разборчивостью. Это особенно заметно при восприятии иностранной речи при слабом владении языком. Однако очень громкая речь сопровождается явлением самомаскировки, также ухудшающим разборчивость.

Повышение уровня случайного шума не только снижает разборчивость речи, но и изменяет характер ошибок — в первую очередь повреждаются фрикативные, затем взрывные согласные и только при очень высоких уровнях шумов искажаются гласные. Кроме того, в зависимости от уровня шумов меняется надежность распознавания разных частей речи, перераспределяется вклад различных лингвистических уровней в понимание речи, с ростом уровня шумов растет роль статистических характеристик речи. Вид помехи также влияет на восприятие — наиболее трудным оказывается понимание речи на фоне другого разговора.

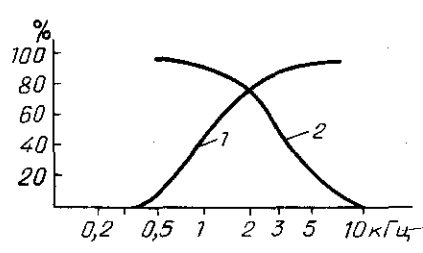


Рис. 11.1. Зависимость слоговой разборчивости от частоты ограничения спектра снизу (*1*) и сверху (*2*) (по [42])

Ограничения частотного диапазона речевого сигнала детально исследовались в связи с разработкой технических условий на характеристики телефонных каналов. Установлено, что ограничение как снизу, так и сверху по частоте монотонно ухудшает разборчивость (см. рис. 11.1).

Синтетическая речь обладает некоторым преимуществом перед естественной в отношении возможности приспособления к характеристикам канала связи. Так, для повышения разборчивости формантного синтеза при использовании телефонного канала в [137] пришлось повысить уровень фрикативных

и добавить сегмент с нейтральным гласным после раскрытия смычки конечных согласных. Такое избирательное изменение параметров трудно реализовать в естественной речи, хотя здесь имеются свои средства повышения разборчивости при разговоре по телефону.

Известно, что восприятие речи правым и левым ухом дает *неодинаковый результат. Это связывают со специализацией* полушарий коры головного мозга — левое полушарие в основном выполняет операции анализа фонетической, синтаксической и семантической информации, тогда как в правом полушарии анализируются интонационные и эмоциональные характеристики речи. Поэтому при моноуральном прослушивании речи не безразлично, на какое ухо поступает речевой сигнал. Если необходимо обеспечить наилучшее понимание сообщения, как, например, при аварийном оповещении, то правое ухо обладает преимуществом, как более тесно связанное с левым полушарием.

Структура речевого сообщения, в том числе активные или пассивные грамматические формы, положительное или отрицательное, истинное или ложное высказывание, длительность сообщения, предсказуемость и другие лингвистические факторы, вклад которых меняется в зависимости от контекста, определяют разборчивость и степень понимания речи, устойчивость к шумам и искажениям. Восприятие изолированных слов (команд), отдельных фраз или беглого разговора требует участия различных лингвистических механизмов анализа речи. При разговоре по плохому телефонному каналу, особенно с незнакомым собеседником на тему с низкой избыточностью, речь не только замедляется и становится громче, но и выбираются другие слова, упрощается структура фраз, уменьшается их длина.

Все эти факторы действительны и для синтетической речи, однако к ним добавляются комплекс просодических характеристик — *частота основного тона, длительность и громкость* сегментов. Эти характеристики наиболее информативны для определения типа фразы и акцентного слова, что решающим образом влияет на понимание смысла речевого сообщения. Понимание речи зависит от особенностей механизмов декодирования речевого сигнала, поиска информации в долговременной памяти, способов интерпретации и использования различных источников знаний, которые в известной мере связаны с лингвистическим опытом слушателя. Одному и тому же показателю разборчивости — фонемной, слоговой или словесной, может соответствовать разная степень понимания содержания сообщения. Об этом свидетельствует, в частности, так называемое автоматическое письмо, когда прослушиваемый текст записывается фонетически точно, но понимание полностью отсутствует.

Ясно, что выбор структуры речевого сообщения — лексической, синтаксической и просодической, столь же важен для

понимания синтетической речи, как и разборчивость фонетических элементов.

Даже краткий анализ факторов, влияющих на восприятие речи, который был проведен выше, показывает необходимость разработки многочисленных методов тестирования синтетической речи. Часть таких методов может быть заимствована из практики оценки качества телефонных каналов и вокодерных систем, хотя в применении к синтезу речи эти методы должны подвергнуться определенной переработке. Кроме того, нужны новые методы тестирования, специфичные именно к синтетической речи, например, оценки натуральности, задержки восприятия и т. д. Эти вопросы будут рассматриваться в последующих разделах.

В настоящее время наилучшее качество речи в классе систем синтеза по правилам обеспечивают формантные синтезаторы. Эти синтезаторы лучше других исследованы на различные показатели качества. Анализ результатов этих исследований дает важную информацию о возможностях, недостатках и направлениях дальнейших исследований в области синтеза речи. Разборчивость и натуральность формантного синтеза достигли достаточно высокого уровня, приемлемого в условиях практического применения, однако установлено, что даже для формантных синтезаторов с наиболее высокой разборчивостью (как, например, *DECTalk*) восприятие синтетической речи требует больше умственных усилий и внимания, чем естественной речи [149, 165]. По всем измеримым параметрам образцы формантного синтеза хуже естественной речи: по разборчивости, натуральности, времени задержки восприятия, устойчивости к помехам. Для иллюстрации этих свойств рассмотрим таблицу ошибок восприятия для некоторых синтезаторов в различных условиях (по [41]).

Таблица 11.1. Ошибка в процентах

Условия восприятия	Ограниченный выбор	Неограниченный выбор	Сигнал/шум 28 дБ	Бессмысленные фразы	Осмысленные фразы
Речь	0,6	2,8	3,4	0,8	2,3
DECTalk Paul	3,3	13,3	7,8	4,7	13,2
DECTalk Betty	5,6	17,5	—	9,5	24,9
Infovox	12,6	—	—	—	—
Votrax	32,8	—	63	—	—

Два первых столбца этой таблицы получены в экспериментах по восприятию синтетических слогов типа СГС с ограниченным выбором из 6 вариантов и неограниченным выбором. Третий столбец соответствует ошибкам в восприятии слогов типа «согласный — гласный» в шумах очень низкого уровня — с

отношением сигнал/шум, равным +28 дБ. В последних двух столбцах приведены ошибки в восприятии слов в осмысленных и бессмысленных фразах.

Прежде всего бросается в глаза большая разница в ошибках для разных типов синтезаторов, например, от 3,3% для *DECTalk Paul* до 32,8% для *Votrax*. Это свидетельствует о том, что принцип формантного синтеза сам по себе еще не гарантирует высокой разборчивости — для создания высококачественной системы необходимы глубокие знания свойств речеобразования и восприятия, а также то, что условно можно назвать «фонетическим слухом». Этот последний фактор играет существенную роль еще и потому, что при разработке синтезатора для какого-либо языка носителями другого языка неизбежны ошибки, даже если разработчики в совершенстве владеют неродным языком. Так, в [45] было найдено, что такие просодические характеристики, как длительность сегментов, динамика переходов от согласных к гласным, интенсивность различны в речи людей, для которых данный язык является родным или неродным. Многоязычная система *Infovox* разработана шведскими учеными Б. Гранстремом и Р. Карлсоном, безукоризненно владеющими английским языком, но, как видно из таблицы, слоговая ошибка в самых благоприятных условиях ограниченного выбора почти в 4 раза выше, чем для *DECTalk*, созданного носителем английского языка Д. Клаттом. Сам Д. Клатт указывает в [137] на трудности распространения возможностей системы *DECTalk* и *Prose* на синтез других языков. Отсюда следует, что в процессе разработки и оптимизации синтезатора для некоторого языка должны участвовать носители этого языка.

Восприятие синтетической речи даже на фоне очень слабого шума (отношение сигнал/шум + 28 дБ) значительно увеличивает ошибки, причем для системы *Votrax* ошибки становятся столь большими, что это практически приводит к тому, что в телефонии называют «срывом связи».

Наконец, сопоставление словесной разборчивости в осмысленных и бессмысленных фразах дает объективную оценку фонетического качества синтезатора — ошибки в бессмысленных фразах почти в 6 раз больше, чем для естественной речи для синтетического мужского голоса (*Paul*) и почти в 11 раз больше — для женского голоса (*Betty*). Сравнивая матрицы ошибок фонетического элемента для естественной и синтетической речи, в [41] приходят к заключению, что формантный синтез совсем не соответствует естественной речи, слегка замаскированной шумом. Отличия между естественной и синтетической речью гораздо более глубоки — синтетическая речь перцептивно беднее, в ней слабо представлены или вообще отсутствуют важные акустические признаки, многие фонетические отличия кодируются лишь одним акустическим признаком, имеются «лже-признаки», лишь поверхностно имитирующие

истинные признаки. Поэтому появляются два типа ошибок: для тех звуков, у которых различия объективно малы (например, между фрикативными /С/ и /Ш/) и для тех звуков, которые близки в пространстве признаков, создаваемых синтезатором. Низкая избыточность синтетической речи приводит к плохой помехоустойчивости, а неестественные различительные признаки затрудняют обработку сигнала в системе восприятия человека. Отсюда следует, что предстоит большая работа по совершенствованию синтезаторов речи, и наиболее перспективны артикуляторные синтезаторы, в максимальной степени моделирующие процессы речеобразования.

Приемлемость синтезатора как источника информации для потребителя зависит не только от качества синтетической речи, но и от правил, которыми руководствуется система, управляющая синтезатором, т. е. от того, что иногда называют этикетом общения [23]. Поскольку момент начала речевого сообщения выбирается управляющей системой безотносительно к внешней ситуации (слушатель может быть занят разговором или даже отсутствовать), то нужно использовать какие-то способы привлечения внимания и подтверждения приема сообщения. У синтезаторов низкого качества речевые сигналы настолько отличаются от естественной речи, что сами по себе привлекают внимание. Так, в одной системе аварийного оповещения на борту самолета сочли возможным отказаться от тонального сигнала, предваряющего аварийное сообщение, выиграв на этом около одной секунды, что очень важно в критических ситуациях. Однако и разборчивость таких синтезаторов довольно низкая, так что преимущество плохого синтеза представляется сомнительным. По мере повышения качества синтетической речи она становится все более похожа на естественную, и их различие перестает играть роль привлекающего внимание признака.

В качестве предупредительного сигнала могут использоваться различные неречевые звуки, как это делается, например, на вокзалах и в аэропортах. Во многих случаях может оказаться полезной некоторая нейтральная по содержанию фраза, предваряющая сообщение. Такая фраза обеспечивает и кратковременную тренировку, подготавливающую систему восприятия к особенностям синтетической речи, что особенно важно для слушателя, впервые или редко сталкивающегося с данным синтезатором.

Чувствительность к звукам и скорость восприятия у разных людей и у одного и того же человека в разных ситуациях сильно отличаются. Поэтому синтезаторы должны обладать способностью к регулировке скорости и громкости речи как со стороны потребителя (например, в персональной читающей машине для слепого), так и со стороны интеллектуальной системы, в которую включен синтезатор. Последнее необходимо для различения степени важности сообщения, так как

скороговорка и низкий уровень сигнала могут привести к пропуску важного сообщения слушателем, а чрезмерно артикулируемое и громкое тривиальное сообщение вызовут раздражение.

Должна быть предусмотрена и возможность переспроса, если по каким-либо причинам сообщение синтезатора было пропущено. Форма организации такого переспроса может быть различной — речевой запрос (в системе, распознающей речи), простое нажатие кнопки, или запрос через ЭВМ, например, о нескольких последних сообщениях. В таких системах акустические характеристики повторяемого сообщения должны несколько отличаться от первоначальных, так как точное повторение вызывает раздражение своей неестественностью — человек никогда не повторяет одну и ту же фразу в точности.

Структура речевого сообщения должна быть максимально простой в системах общего пользования, причем выбор слов, ключевых для понимания фразы, может производиться с учетом разборчивости фонетических элементов. Система искусственного интеллекта может учитывать индивидуальные особенности своего собеседника, но обсуждение тактик построения речевого диалога выходит за рамки анализа свойств синтезаторов речи.

§ 11.2. Оптимизация синтезатора

Оценка характеристик синтезатора начинается уже в процессе его разработки. Целью этой оценки является достижение наилучших возможных показателей. Тестирование законченной системы синтеза дает лишь оценку достигнутого качества, но количественное значение, например, разборчивости, может быть различным в зависимости от успеха оптимизации синтезатора. В процессе оптимизации должны использоваться не только методы тестирования готовой системы, но и специальные методы, разработанные в области психофизики восприятия речи. Достижение некоторого желаемого показателя качества синтезатора зависит от множества управляемых параметров, и никакая оптимизация не поможет, если отсутствуют какие-либо важные параметры. Но оптимизация позволяет достичь наилучшего качества, потенциально возможного при заданном наборе управляющих параметров. Математически задача оптимизации качества синтезатора формулируется так же, как и большинство задач оптимизации — варьируя значения n параметров, необходимо добиться экстремального (максимального или минимального) значения некоторого критерия при ограничениях на какие-то другие параметры. Например, критерием оптимальности может служить словесная разборчивость, а ограничением — время задержки решения об услышанном слове. Может быть поставлена и обратная задача — минимизировать время задержки при сохранении разборчивости не ниже заданного уровня.

Известно, что сегментная разборчивость и время задержки решения тесно связаны, но, тем не менее, результаты оптимизации по первому или второму критерию могут оказаться разными. И первая, и вторая постановки задачи вполне могут встретиться в реальных условиях даже в рамках одного и того же применения. Например, для пилота самолета, летящего на большой высоте, редкие сообщения о высоте оставляют время для восприятия, тогда как в критических ситуациях необходимо минимизировать время восприятия сообщения о параметрах полета.

Ясно, что критерий оптимальности для синтезатора, рассчитанного на применение в программах школьного обучения, должен отличаться от критерия оптимальности для синтезатора аварийного оповещения на атомной станции. Наиболее гибкие синтезаторы должны менять «стиль» произношения в зависимости от ситуации, точно так же, как это делает человек. Выбор критерия оптимальности и тактика смены критериев, как и всегда, являются наиболее ответственными и плохо формализуемыми факторами и зависят от намерений и возможностей разработчика синтезатора. В любом случае имеется более или менее осознаваемый этап оптимизации синтезатора на стадии его разработки, поэтому необходимо рассмотреть существующие методы оптимизации с целью их приспособления к весьма специфической задаче синтеза речи.

Как известно, методы поиска оптимума зависят от числа экстремумов и формы функции качества в многомерном пространстве параметров. Некоторые сведения об этих характеристиках можно получить из психофизических экспериментов, хотя в них обычно ставятся лишь одномерные задачи, т. е. определяется значение некоторого фактора, например, распознаваемости фонетического элемента в зависимости от одного параметра. Наиболее характерным признаком, проявляющимся в этих экспериментах, служит наличие плато в некоторой области значений независимой переменной. При этом встречаются две ситуации: ограничение зоны максимального отклика только с одной стороны и двустороннее ограничение. Для одностороннего ограничения чаще всего встречается плато, или зона насыщения, в которой величина отклика не изменяется при изменении величины стимула. Именно, таким образом, восприятие слога зависит от длительности гласного (рис. 11.2 по [66]). В других случаях может появляться более или менее выраженный максимум, как в зависимости разборчивости от громкости (см. § 11.3). В задачах с двусторонним ограничением появление максимума более вероятно, однако типичным все же оказывается плато (рис. 11.3 по [66]). Следовательно, специфика оптимизации синтезатора заключается не столько в поиске максимума (или минимума) некоторого критерия, сколько в определении границ области, в которой значения оптимизируемой величины предельно велики (малы) и мало отличаются друг от друга.

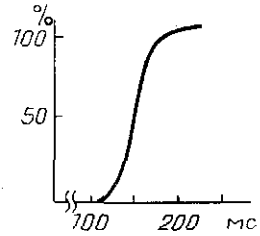


Рис. 11.2. Вероятность восприятия гласного как ударного в зависимости от его длительности (по [66])

Другое заключение, которое можно сделать из анализа психофизических экспериментов — монотонная зависимость отклика от величины стимула. Это очень важное свойство, оно указывает на высокую вероятность существования лишь одной зоны оптимума (унимодальность), хотя, конечно, свойства многомерного пространства могут сильно отличаться от свойств его проекций на одномерные сечения, и при многопараметрической оптимизации не исключено появление многих экстремумов.

В процессе разработки синтезатора проводится два вида оптимизации, которые условно можно назвать объективной и субъективной. Объективная оптимизация применяется в тех случаях, когда известна количественная характеристика, которой *нужно добиться. Тогда формулируется минимизация* некоторой меры (в частном случае — расстояния) между целевыми и реальными характеристиками синтезатора. Примером такого случая может служить оптимизация формантных частот для гласных в артикуляторном синтезаторе. Поскольку из непосредственных измерений можно установить множество целевых значений формантных частот для каждого гласного, то задача оптимизации сводится к поиску таких значений координат артикуляторных органов, которые обеспечили бы наибольшее приближение формантных частот к целевым. Эта ситуация рассматривалась в гл. 9 и было показано, что задача оптимизации относится к классу задач на поиск экстремума с ограничениями. Оптимизация по объективным характеристикам является лишь частью процесса оптимизации синтезатора.

Оптимизация по субъективным оценкам слушателей требуется во всех случаях, когда в качестве критерия выступают разборчивость, натуральность, трудность понимания или задержка восприятия. Вследствие необходимости проведения

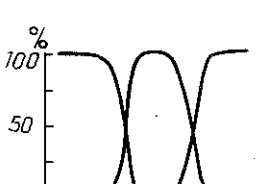


Рис. 11.3. Вероятность идентификации согласных в зависимости от направления переходов второй и третьей формант (по [66]). По оси абсцисс — номер стимула

многочисленных аудиторских испытаний на некотором множестве слушателей, задача оптимизации значительно усложняется, однако без такой оптимизации невозможно достичь наилучшего качества синтетической речи. Суждения разработчиков синтезатора субъективны, они не отражают свойств восприятия в достаточной мере. К тому же обычно возраст разработчиков превышает критический возраст в 30 лет, за которым начинается ухудшение слуха. По этим причинам нужно пользоваться специальной бригадой тренированных аудиторов.

Поскольку нельзя избежать оптимизации по субъективным оценкам, то следует пользоваться только такими методами, которые позволяют найти оптимум при минимальном количестве аудиторских экспериментов. Для получения непрерывных характеристик отклика, подобных показанным на рис. 11.2 и 11.3, перемешанные в случайном порядке стимулы многократно предъявляются аудиторами, и затем подсчитывается число положительных реакций, точнее, частота их появления. Такая методика хорошо работает в экспериментах по различению или распознаванию стимулов (например, звуков или слогов). Аналогичным образом можно найти количественную зависимость разборчивости от исследуемых факторов. Если аудитору последовательно предъявляется два стимула, то зачастую бывает очень трудно дать количественную оценку каждого из этих стимулов в соответствии с заданным критерием. Обычно эти оценки ненадежны, и можно получить лишь сравнительную оценку пары стимулов — является ли один из них лучше другого в смысле используемого критерия. Такая качественная оценка равносильна определению знака производной по оптимизируемой функции, но не самой производной. Это обстоятельство ограничивает применение методов, в которых оценка производной существенна, например, в методах градиентного спуска. Следует иметь в виду также наличие порога восприятия разницы в стимулах — для разных параметров этот порог различен, и в его пределах стимулы неразличимы.

Все эти обстоятельства затрудняют оптимизацию синтезатора. Имеются, однако, и такие свойства, которые ее облегчают. Так, практически всегда известны границы независимых переменных и примерное расположение области оптимума. Кроме того, оптимизация никогда не требует использования сразу всех независимых переменных (а их несколько десятков). В каждом конкретном случае, например, для каждого звука, существует весьма ограниченный набор параметров, по которым имеет смысл проводить оптимизацию. Это существенно уменьшает сложность оптимизации, поскольку число экспериментов в окрестности каждой точки оптимизируемой функции по крайней мере на единицу больше числа независимых переменных.

В ряде случаев оптимизация проводится всего по одному параметру или же задачу можно свести к последовательности оптимизации по каждому параметру. Тогда, предполагая наличие лишь одного экстремума (унимодальность) в диапазоне значений параметра, можно построить такую последовательность экспериментов, что их число будет близко к минимально необходимому. Один из таких методов называется правилом золотого сечения.

Пусть диапазон значений некоторого параметра находится между величинами a и b, $a<b$ и $b-a=l$. Располагая на интервале $[a, b]$ точки экстремума и сравнивая значения оптимизируемой функции F, будем искать положение экстремума. Очевидно, что в силу предположения унимодальности, после первой пары экспериментов поиск следует производить в окрестностях точки с наибольшим значением оптимизируемой функции. Для ситуации, показанной на рис. 11.4, это точка x_1 и интервал неопределенности, таким образом, сокращается с $l=b-a$ до $l_1=x_2-a$. Потребуем, чтобы каждая точка эксперимента делила интервал неопределенности l_i таким образом, чтобы $l_i=l_{i+1}+l_{i+2}$ и отношение длины наибольшего отрезка к длине всего интервала неопределенности равнялось отношению длины меньшего отрезка к длине наибольшего интервала:

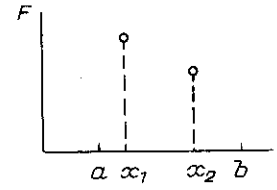


Рис. 11.4. Отсчеты оптимизируемой функции

$$l_{i+1}/l_i=l_{i+2}/l_{i+1}.$$

Это и есть правило золотого сечения. Из этих двух соотношений получаем квадратное уравнение $\tau^2+\tau+1=0$, где $\tau=l_{i+2}/l_{i+1}$. Решая его, найдем положительное значение корня $\tau=0{,}618$. Это означает, что внутри любого интервала неопределенности l_i точка эксперимента должна располагаться таким образом, чтобы разделить его на отрезки $l_{i+1}\approx 0{,}618 l_i$ и $l_{i+2}\approx 0{,}382 l_i$. В начале эксперимента проводятся измерения по двум точкам: $x_1=0{,}382l+a$, $x_2=0{,}618l+a$, а в дальнейших экспериментах проводится сравнение с одной из точек, оставшихся от предыдущего эксперимента. По мере увеличения числа экспериментов, интервал неопределенности сокращается и на k-м шаге он равен $l_k\approx 0{,}618^k$. Эксперименты прекращаются, когда отношение l_k/l станет меньше порога чувствительности или будет достигнута требуемая точность. Так, для $k=11$ относительная точность равна $\approx 0{,}008$, т. е. лучше 1%.

Поскольку точки эксперимента стягиваются к области экстремума, то их последовательность дает возможность восстановить форму оптимизируемой функции с неравномерными отсчетами — вблизи экстремума отсчеты расположены более часто, а вдали — редко. Метод золотого сечения дает

почти оптимальное число экспериментов, может быть, лишь на один эксперимент больше оптимального в интересующем нас диапазоне погрешностей.

В случае, когда оптимизируемая функция F не имеет ярко выраженного экстремума, а обладает плато, как на рис. 11.2,

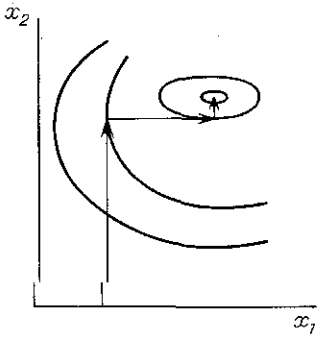


Рис. 11.5. Унимодальная функция с эллиптическими изолиниями

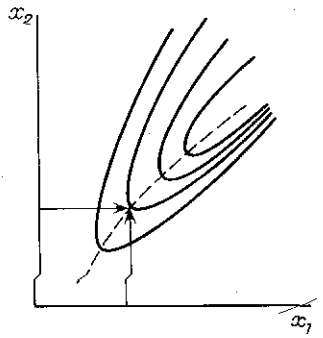


Рис. 11.6. Унимодальная функция с хребтами

то на некотором этапе появятся практически одинаковые значения $F(x_i)$ и $F(x_{i-1})$. Если нас не интересует граница плато, то можно удовлетвориться любым из этих значений, и эксперимент прекращается (конечно, при условии, что эти значения находятся вблизи максимума, если мы ищем наибольшую величину, или вблизи минимума, если ищется минимальное значение оптимизируемой функции F).

Существуют и другие методы одномерного поиска, отличающиеся сложностью и скоростью сходимости (см., например, [63]).

Рассмотренный выше метод поиска экстремума по одному параметру полностью применим и при поиске экстремума в многомерном пространстве, если используется, например, метод Гаусса — Зайделя, в котором многомерный поиск сводится к последовательности одномерных поисков по каждому из параметров при фиксированных остальных параметрах. Этот метод физически нагляден и его легко организовать при оптимизации с участием аудиторов. Число экспериментов, необходимых для нахождения оптимума с помощью правила золотого сечения пропорционально числу независимых параметров. Например, при требуемой точности в 3% и n параметрах максимальное количество экспериментов равно $7n$. К сожалению, метод Гаусса — Зайделя сходится к экстремуму лишь в ограниченном классе задач, а именно, только тогда, когда линии равного уровня оптимизируемой функции в многомерном пространстве находятся на поверхностях гиперэллипсоидов. Для двумерного случая это просто эллипсы

(рис. 11.5). Если же склоны целевой функции имеют гребни (хребты или овраги), то последовательный одномерный поиск может закончиться далеко от экстремума, поскольку никакое приращение вдоль любого из параметров не позволяет определить направление увеличения целевой функции (рис. 11.6). Такая ситуация встречается при сильной взаимной зависимости исследуемых параметров.

В методе золотого сечения значения целевой функции сравниваются в точках, достаточно удаленных друг от друга. При многомерной оптимизации, наоборот, на некоторых этапах стараются расположить точки эксперимента как можно ближе друг к другу с тем, чтобы оценить направление наибольшего возрастания или спада целевой функции. Это связано с возможностью разложения функции $F(x_1, x_2, ..., x_n)$ в ряд Тейлора в окрестности некоторой точки x^j с координатами $x_1^j, x_2^j, ..., x_n^j$, если это непрерывная функция. Разложение с точностью до членов первого порядка имеет вид

$$\Delta F = \sum_{i=1}^{n} \left(\frac{\partial F}{\partial x_i}\right) \Delta x_i,$$

где n — число переменных, Δx_i — приращение по каждой переменной. Представляя таким образом функцию F, осуществляем линейную аппроксимацию гиперповерхности. Поскольку в окрестности экстремума первые производные по всем параметрам близки к нулю, то для более точного определения положения экстремума иногда прибегают к нелинейной аппроксимации с разложением функции F в ряд Тейлора с точностью до членов второго порядка.

Когда производится поиск минимума F, то перспективными для дальнейших исследований являются только те направления, для которых $\Delta F < 0$, и в одном из этих направлений можно совершить довольно большой скачок. Существуют различные тактики выбора направлений и величины скачка для следующего эксперимента. Среди них следует упомянуть методы касательных и градиентные методы. В методе касательных из уравнения $\Delta F = 0$ находят положение гиперплоскости, касательной к поверхности F, определяют направление возрастания F, и следующая точка эксперимента размещается, например, в центре объема перспективной области. Эта процедура продолжается до тех пор, пока объем неопределенности не станет достаточно мал.

Метод касательных не зависит от масштабов по переменным, но он очень чувствителен к ошибкам эксперимента, и для его сходимости требуется строгая унимодальность функции F, т. е. необходимо, чтобы на любой прямой, проведенной из точки эксперимента к вершине F, значение F возрастало. Достоинством метода касательных является и то, что в нем не требуется оценки величины производной,

а достаточно определить лишь знак приращения ΔF. Это позволяет применять его в тех задачах оптимизации синтезатора, где ответы аудиторов строятся по типу «лучше — хуже».

Очень популярные методы градиентного подъема, требующие вычисления направления наискорейшего возрастания функции F, в задачах оптимизации синтезатора следует применять с большой осторожностью и не только из-за ошибок аудиторских оценок. Вследствие различной физической природы параметров синтезатора направление градиента может быть любым в зависимости от масштаба по каждому параметру, что по крайней мере сильно усложняет процедуру поиска.

Поиск максимума можно ускорить, если удается найти гребень, ведущий к вершине, и двигаться по нему, даже если этот гребень криволинейный. Это же относится и к поиску минимума, только там вместо гребня нужно следовать вдоль «оврага». Разработан ряд методов движения вдоль гребня, но их преимущества и недостатки нужно сопоставлять с конкретной задачей оптимизации синтезатора, а об этих особенностях пока известно очень мало.

Если в процессе оптимизации найден некий экстремум целевой функции F, то нужно позаботиться о проверке его единственности. Известно, что во многих случаях для человека равное качество некоторых стимулов может иметь место при отличающихся, иногда весьма значительно, значениях параметров. Поэтому возникает подозрение о существовании нескольких оптимумов, которое необходимо проверять в каждом конкретном случае, и при обнаружении полимодальности каким-то образом принимать решение о выборе того или иного экстремума.

Поиск оптимума качества синтезатора связан также с вопросом о существовании так называемых суперстимулов, т. е. таких сигналов, которые обеспечивали бы, например, разборчивость, более высокую, чем максимальная разборчивость, наблюдающаяся для естественной речи. Из анализа поведения животных, рыб и птиц известно о существовании таких суперстимулов, причем обычно суперстимул характеризуется преувеличенным значением какого-либо важного признака, но таким его значением, которое в естественных условиях не встречается. Система управления артикуляцией обладает ограниченной управляемостью, что связано с кинематикой и динамическими ограничениями артикуляторов. В синтезаторе часть этих ограничений может быть преодолена — например, можно имитировать мышечные усилия, значительно превышающие те, которые действуют в реальных процессах речеобразования. Таким образом, не исключается возможность достижения качества синтетической речи, превышающего качество естественной речи, хотя, конечно, этот вопрос требует тщательных исследований.

Концепция оптимизации синтезаторов речи с помощью математических методов пока является совершенно новой и нет никаких сведений о результатах применения той или иной процедуры, кроме экспериментов, описанных в § 9.3. Поэтому цель данного раздела состоит не в обсуждении конкретных результатов оптимизации, а в обосновании необходимости оптимизации и поиске возможных направлений в этой области.

§ 11.3. Разборчивость

Из первого раздела данной главы следует, что речь характеризуется большим числом показателей, так что для всесторонней оценки синтетической речи нужно уметь измерять наиболее важные параметры. Поскольку синтетическая речь предназначена для передачи информации человеку, то, в конечном счете, оценку качества синтезатора может дать только человек. Подобная ситуация существует уже давно в области оценки телефонных каналов, особенно в вокодерных системах, где ряд проблем аналогичен проблемам в оценке синтезаторов. Поэтому можно воспользоваться опытом, накопленным при тестировании телефонных каналов. К сожалению, этот опыт, с одной стороны, недостаточен, а, с другой стороны, может быть лишь частично использован при тестировании синтезаторов. Принципиальная разница между этими двумя задачами состоит в том, что при тестировании телефонных систем источником испытательных тестов является человек, что гарантирует естественные соотношения между всеми акустическими параметрами и лингвистическими уровнями, в то время как в синтетической речи такой гарантии не существует, и необходимо проверять адекватность всех основных свойств. Например, известно, что для любого конкретного языка существует однозначная зависимость между фонемной, слоговой и словесной разборчивостью — измерив один вид разборчивости, остальные можно рассчитать. Однако эта зависимость не обязательно справедлива для синтетической речи. По крайней мере, для существующих синтезаторов она не справедлива. Отсюда следует необходимость проведения специальных экспериментов для измерения каждого вида разборчивости в каждом типе синтезаторов.

Задача оценки качества синтетической речи настолько же сложнее задачи оценки качества телефонного канала, насколько задача синтеза речи сложнее задачи ее передачи — вместо одного вида разборчивости нужно оценивать целый ряд видов, вместо усредненных оценок нужно находить их распределения, требуется измерять новые характеристики — натуральность, задержку восприятия, трудность понимания и т. д.

Из аудиометрических измерений известно, что с возрастом острота и частотный диапазон слуха у человека ухудшаются. Поэтому в телефонии рекомендуется составлять аудиторские

бригады для тестирования из людей в возрасте 18—23 лет и, во всяком случае, не старше 30 лет. При оценке синтезаторов следует различать процедуры аудиторской экспертизы для оптимизации и тестирование законченной системы. В первом случае действительно нужно, чтобы аудиторы обладали максимальной чувствительно-

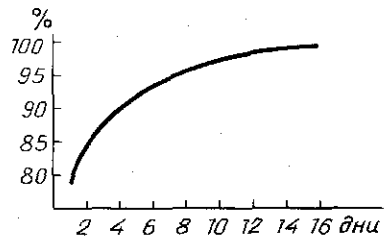


Рис. 11.7. Изменение слоговой разборчивости аудиторов в процессе тренировки (по [42])

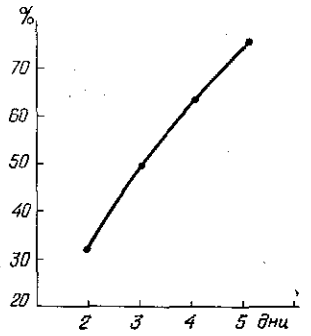


Рис. 11.8. Тренировка к синтетической речи. Словесная разборчивость (по [111])

стью к изменениям параметров с целью быстрейшего определения оптимума. Оценка же завершенной системы, в принципе, должна проводиться бригадой, отражающей возрастной состав потенциальных пользователей. Вообще говоря, особенности пользователей должны учитываться еще в процессе оптимизации для возможной компенсации одних признаков (например, не воспринимаемых пожилыми людьми) другими признаками. Можно ожидать, что оценки качества одного и того же синтезатора детьми (например, в системах школьного обучения) и пожилыми людьми (например, для читающих машин), окажутся сильно различающимися. Конечно, учет различия в свойствах восприятия разных групп людей осложняет задачу оценки качества синтезатора, но игнорирование этого фактора может заметно снизить эффективность систем синтеза речи.

Бригада аудиторов, помимо возрастного состава, характеризуется степенью тренированности, причем этот показатель меняется в течение довольно длительного времени. Так, из рис. 11.7 видно, что рост слоговой разборчивости от 78% до 98% происходит в течение двух недель. В этих экспериментах диктор и аудитор находились на расстоянии одного метра друг от друга, и передача речи осуществлялась просто по воздуху. Если речевой сигнал подвергается искажениям, то процесс тренировки требует еще большего времени [42]. Аналогичное нарастание разборчивости в процессе тренировки наблюдается и для синтетической речи (рис. 11.8). Установлено, что хотя слушатель, достигший асимптотического значения разборчивости в процессе тренировки, в значительной мере сохраняет ее на протяжении многих месяцев, наблюдаются

колебания разборчивости — в одной и той же серии экспериментов разборчивость нарастает, а после перерыва в 1—2 дня — падает. Поскольку в числе пользователей синтезатором могут оказаться и такие, кто слышит его впервые, как, например, в информационно-справочных системах с речевым выводом (адреса, телефоны, расписание поездов и т. д.), то, наряду с асимптотической оценкой разборчивости, синтезатор должен характеризоваться и начальной разборчивостью, присущей нетренированному аудитору. В [111] было найдено, что даже единственное появление искаженного стимула перед тестом улучшает его распознаваемость в процессе испытаний. Исходя из этого, в системах массового пользования с большой вероятностью появления нетренированного слушателя, целесообразно предварять информационное сообщение легко распознающимся вводным словом или фразой для проведения экспресс-тренировки.

Есть основания полагать, что процесс тренировки для синтетической речи отличается от процесса тренировки для естественной речи. Так, при испытаниях системы формантного синтеза *Votrax*, в которой фразы просто составляются из слов без каких-либо коартикуляционных изменений на их границах, отмечается большое различие в словесной разборчивости в зависимости от вида тренировки [111]. Если аудиторы тренировались на отдельных словах, то словесная разборчивость повышалась с течением времени также и для слов, не участвовавших в тренировке. Однако разборчивость тех же слов, предъявленных в слитной последовательности (во фразах) оказалась такой же, как и у нетренированных дикторов. Если же аудиторы тренировались на фразах, то их словесная разборчивость возрастала и для изолированных слов. В случае, когда аудиторы, тренировавшиеся на изолированных словах, получают информацию о начале слова, например, в виде артикля «the», то словесная разборчивость в слитной речи также возрастает.

Это свидетельствует о необходимости выработки у аудитора ключей сегментации слитной синтетической речи на слова. По-видимому, это свойство присуще не только синтетической речи — при обучении иностранному языку также имеются трудности при переходе от восприятия отдельных слов к восприятию слитной речи. В этом отношении синтетическая речь может рассматриваться как специфическая разновидность иностранного языка. Возможно, что для тренировки на восприятие синтетической речи плохого качества окажутся полезными методики обучения иностранному языку, но, очевидно, следует приложить максимум усилий для создания искусственной речи, неотличимой от естественной.

Разборчивость речи зависит не только от тренированности аудиторов, но и от громкости. На рис. 11.9 приведено семейство кривых, соответствующих слоговой разборчивости русской речи при различных уровнях шумов с равномерным спектром. Видно, что при малых шумах разборчивость обладает максимумом

в области 90—100 дБ. Дальнейшее повышение громкости приводит к понижению разборчивости вследствие самомаскировки речи, когда интенсивные гласные маскируют качество предшествующего согласного. Поскольку люди обладают разной

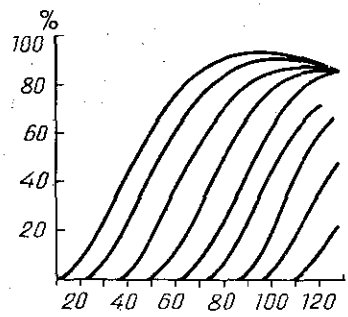


Рис. 11.9. Слоговая разборчивость как функция суммарной громкости речи и аддитивного белого шума. Уровень шумов возрастает на 10 дБ для каждой кривой при сдвиге слева направо (по [42])

остротой слуха, то необходимо предусмотреть ручную регулировку коэффициента усиления синтезатора. Изменение этого коэффициента в соответствии с уровнем шума в помещении можно осуществлять автоматически на основе измерения этого уровня. Можно ожидать, что при восприятии синтетической речи уровень громкости будет влиять на разборчивость подобно тому, как это происходит при восприятии иностранной речи при плохом знании языка. Поэтому кривая разборчивости синтетической речи при отсутствии шумов является одной из характеристик синтезатора, требующей аудиторской оценки. О восприятии синтетической речи в шумах будет говориться в § 11.5.

При повышении громкости естественной речи уровень верхних формант повышается относительно уровня первой форманты. Это вызвано повышением уровня высокочастотных компонент в импульсах голосового источника. Возможно, что этот признак может служить маркером важности сообщения. Поэтому необходимо исследовать различия в восприятии как при простом повышении уровня громкости путем изменения общего коэффициента усиления, так и при повышении внутрилегочного давления в модели голосового источника.

Разборчивостью называется доля (процент) правильно принятых элементов речи. Различают звуковую (фонемную), слоговую, словесную и фразовую разборчивости. Опираясь на элементарные соображения о вероятности правильного приема группы элементов при заданных вероятностях ошибок приема этих элементов, можно получить общий вид зависимости между разными видами разборчивости. Эксперименты на естественной речи, в общем, подтверждают теоретические выкладки. На рис. 11.10 показана зависимость слоговой, словесной и фразовой разборчивости от фонемной для естественной речи (по [42]). Обращает на себя внимание разная чувствительность различных видов разборчивости, т. е. скорость изменения кривых.

Так, при плохой фонемной разборчивости (ниже 20%), кривая словесной разборчивости обладает большей чувствительностью, чем кривые слоговой и фразовой разборчивости. Фразовая и слоговая разборчивость обладают максимальной чувствительностью в диапазоне 30—60% звуковой разборчивости. При высокой звуковой разборчивости (выше 60%) наибольшей чувствительностью обладает слоговая разборчивость. Поэтому, хотя и существует однозначная зависимость между всеми видами разборчивости, следует *пользоваться разными методами* измерений разборчивости в зависимости от качества речи.

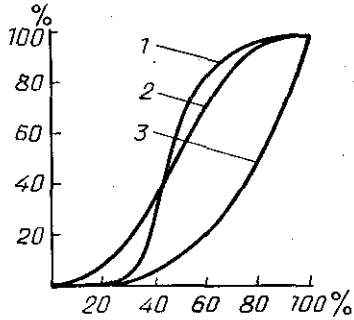


Рис. 11.10. Зависимость фразовой (*1*), словесной (*2*) и слоговой (*3*) разборчивости от фонемной разборчивости (по [42])

В синтетической речи также существуют определенные зависимости между разными видами разборчивости, однако они определяются типом синтезатора и отличаются от зависимостей в естественной речи. Поэтому нельзя пользоваться законами, полученными для естественной речи, например, для пересчета слоговой разборчивости в словесную при оценке качества синтезатора.

Фонемная разборчивость является базовой, поскольку все остальные вида разборчивости зависят от нее. Это справедливо и для синтетической речи. Установлено, что синтезаторы с худшей *фонемной разборчивостью обладают* и *худшей словесной* разборчивостью [41]. Фонемную разборчивость измеряют путем сопоставления осмысленных и бессмысленных слов, отличающихся лишь одним звуком. К числу первых относятся такие последовательности, как «вол, пол, гол, мол, тол», «кочка, почка, дочка, точка, мочка, ночка», «был, бил, бал» и т. д. Однако в осмысленных словах нельзя найти все возможные противопоставления и варианты положений звуков, поэтому пользуются и бессмысленными слогами типа ГСГ или СГС, отличающимися лишь одним звуком: «аба, ада, ага, ...», «бут, пут, дут, тут, ...» и т. д.

Тесты, использующие слоги СГС, в западной литературе называют рифмованными. При оптимизации синтезатора целесообразно начинать со слогов ГСГ и СГС, поскольку *согласный помещается в начальной, интервокальной и конечной* позициях. После проведения первого этапа оптимизации следует проверить фонемную разборчивость на списках слов с минимальными отличиями. Результаты этих экспериментов, представленные в виде матрицы ошибок, дают ясную информацию о том, какие звуки чаще всего перепутываются при восприятии. Следующий этап оптимизации проводится для улучшения различимости наиболее часто перепутываемых звуков.

Существует два способа организации рифмованных тестов: с ограниченным и неограниченным выбором. В первом способе аудитору предлагается список, например, из шести слов, в котором он должен подчеркнуть услышанный слог. В силу такого ограничения вероятность случайного угадывания повышается, но качественная структура ошибок сохраняется, а подсчет числа ошибок очень прост и легко автоматизируется. Результаты испытаний разных типов синтезаторов на подобном тесте показаны в табл. 11.2.

Таблица 11.2. Ошибки восприятия синтетических звуков, %

Позиция согласного	Речь	DECTalk Paul	DECTalk Betty	Prose	Infovox
Начальная	0,5	1,6	3,4	7,1	10
Конечная	0,55	4,9	7,9	4,3	15

В тестах с неограниченным выбором фонемная разборчивость, конечно оказывается хуже, но зависит при этом от вида речевого материала — *слогов или фраз* [112] (см. табл. 11.3).

Таблица 11.3. Ошибки восприятия синтетических звуков, %

Выбор	Речевой материал	DECTalk Paul	DECTalk Betty
Ограниченный	слоги	3,3	4,6
	фразы	4,7	13,2
Неограниченный	слоги	13,2	17,5
	фразы	9,5	24

Сопоставление матриц ошибок восприятия фонем для естественной и синтетической речи дает возможность оценить степень правильности соблюдения основных перцептивно важных свойств речи в синтезаторе. В формантном синтезаторе этого достичь довольно трудно — структура матрицы ошибок даже для лучших формантных синтезаторов сильно отличается от структуры матрицы ошибок для естественной речи [41, 148]. Такое различие свидетельствует о неадекватном моделировании процессов речеобразования в формантном синтезаторе и приводит к возрастанию числа ошибок, затруднению понимания и увеличению задержки распознавания. Отсюда следует принципиальная важность оценки фонемной разборчивости и анализа ошибок для оптимизации синтезатора.

Разборчивость фонем зависит и от контекста. В синтезе речи это обстоятельство дополняется возможной погрешностью правил коартикуляции. Поэтому следующий вид разборчивости,

требующий отдельной оценки — это слоговая разборчивость. В телефонии хорошо разработаны методы оценки слоговой разборчивости путем передачи и приема слоговых артикуляционных таблиц [6, 42]. Артикуляционные таблицы должны отражать закономерности языка, быть сбалансированными относительно частоты появления различных звукосочетаний, обладать минимальной избыточностью и обеспечивать экономность объема экспериментов. В соответствии с этими требованиями разработаны слоговые таблицы, содержащие по 50 звукосочетаний. При оценке качества телефонного канала обычно используется пять таких таблиц, т. е. всего 250 слогов. Для предотвращения запоминания слоговых таблиц их общее число должно быть около 100, охватывая, таким образом, 5000 слогов.

Оценка слоговой разборчивости на этапе оптимизации синтезатора преследует цель выявления таких звукосочетаний, разборчивость которых особенно низка. В этом заключается отличие от цели измерения слоговой разборчивости телефонных каналов, где единственное число — показатель разборчивости, служит базовым для расчета словесной и фразовой разборчивости и, в конечном счете, оценки качества телефонного канала.

Оценка словесной разборчивости является еще более трудной задачей, чем оценка слоговой разборчивости, так как для получения достоверного показателя нужно было бы включить в тесты по крайней мере наиболее часто встречающиеся слова и словоформы, а их количество исчисляется тысячами. Вследствие необходимости ограничения объема эксперимента ищут такие способы оценки словесной разборчивости, которые были бы представительны при малом числе используемых слов. Один из таких способов называется методом выбора. Он состоит в чтении подряд нескольких слов так, как если бы они составляли предположение. Аудитор располагает карточкой, на которой в том же порядке колонками написаны слова, похожие на слова в переданном сообщении. Задачей аудитора является подчеркнуть то слово, которое, по его мнению, наиболее похоже на переданное слово. Например, при передаче последовательности «бланк, помадка, мол» выбор производится из четырех слов, часто образующих минимальные пары с переданными словами:

бланк	помадка	мол
план	палатка	пол
бант	породка	гол
банк	повадка	вол

Как видно, это тест с ограниченным выбором. Его достоинством является быстрая обработка результатов и практически отсутствие необходимости тренировки аудиторов (для естественной речи). Однако оценка словесной разборчивости по методу выбора нелинейно связана с истинной словесной разборчивостью [6]. Для естественной речи эту зависимость можно определить один раз и пользоваться

во всех экспериментах. Однако для каждого типа синтезаторов и даже для одного и того же синтезатора на разных стадиях оптимизации такую зависимость нужно устанавливать каждый раз заново. Принимая во внимание эти соображения в Хаскинских лабораториях (США) был разработан специальный тест на словесную разборчивость для синтезаторов. Это тест с неограниченным выбором, состоящий в передаче грамматически правильных, но семантически аномальных фраз типа «бодрая дверь скачет взахлеб». Обработка результатов такого теста, однако, очень трудоемка.

Более удачным оказался тест на проверку истинности или ложности фраз, содержащих три — шесть слов, и отличающихся лишь одним ключевым словом, причем решение о ложности нельзя принять до тех пор, пока не будет прослушано последнее слово во фразе. Результаты этого теста коррелируются с результатами рифмованных тестов, но значительно более точны и устойчивы [176]. Обработка ответов аудиторов не представляет никаких трудностей, и легко может быть автоматизирована с помощью двух кнопок — «истина» и «ложь», сигналы от которых поступают в ЭВМ сразу же после генерирования машиной тестовой фразы.

Выше упоминалось о том, что тренировка на восприятие изолированных слов и слитных фраз дает разные результаты. Имеется по крайней мере еще один фактор, не позволяющий ограничиваться оценкой разборчивости изолированных слов. Так, в [18] найдено, что в изолированном произнесении слов лучше всего распознаются существительные, и хуже всего — наречия, тогда как в слитной речи лучше всего распознаются глаголы и наречия. Таким образом, чисто лингвистические характеристики слова влияют на вероятность правильного его приема. Частота встречаемости слов также оказывается одним из факторов, определяющих разборчивость слов.

Тест на истинность — ложность фразы может также использоваться и для оценки степени понимаемости. Другой способ этой оценки заключается в чтении диалога одним и тем же лицом и разбиении аудитором прослушиваемого текста на реплики [21]. В тестах на понимаемость, однако, многое зависит от знакомства аудитора с предметом, обсуждаемым в текстах, его развития, психолингвистических особенностей, мотивации. Поэтому такие оценки для широкого круга слушателей оказываются неустойчивыми.

§ 11.4. Натуральность

Синтетической речи обычно приписывается большая натуральность при повышении разборчивости, хотя эти факторы до некоторой степени независимы. Как уже обсуждалось в гл. 5, основной вклад в натуральность синтетической речи вносят характеристики голосового источника, а также просодические характеристики — длительность сегментов и изменение частоты

основного тона на фразе. Каждый из этих факторов можно оценивать по отдельности, но необходима и интегральная оценка натуральности. При оценке натуральности еще в меньшей степени, чем при оценке разборчивости, применимы количественные методы. Наиболее пригодными представляются метод парных сравнений и метод категориальных оценок.

В методе парных сравнений аудитору последовательно предъявляется два звукосочетания с инструкцией оценить, какое из этих звукосочетаний звучит более натурально. Манипулируя параметрами синтезатора, можно попытаться отыскать максимально возможную степень натуральности. Процедура поиска максимальной натуральности математически не отличается от ранее описанных процедур для поиска максимума разборчивости в случае оценок типа «лучше — хуже». Достоинством метода парных сравнений является возможность установления функциональной зависимости между параметрами звукового стимула и оценкой натуральности звучания. Его недостаток — большая трудоемкость.

Известно, что в паре стимулов последний стимул оценивается более положительно, чем первый, причем разница в числе положительных оценок тем больше, чем меньше объективное различие между стимулами. В этом проявляется один из законов психофизики, состоящий в увеличении разброса оценок в области неопределенности. Для того чтобы избавиться от влияния порядка предъявления стимулов, нужно каждую пару стимулов предъявлять дважды — с перестановкой стимулов. Модификация этого метода состоит в предъявлении не двух, а трех звукосочетаний. При этом каждое из звукосочетаний поочередно назначается эталоном, и аудитор должен ответить, какое из двух оставшихся звукосочетаний более похоже на эталон по натуральности звучания.

Метод категориальных оценок состоит в условном разбиении диапазона возможных ответов на ряд категорий. Число этих категорий обычно не превышает 9, согласно известному свойству восприятия. Чем меньше число категорий, тем более надежны оценки. На практике обычно пользуются пятибалльной системой: «очень мало», «мало», «средне», «много», «очень много». В применении к оценке натуральности, «мало» или «много» означает степень натуральности, субъективно воспринимаемой аудитором. Результирующая оценка определяется как среднее по всем реализациям и аудиторам. Более информативна не средняя оценка, а распределение этих оценок, но для получения более или менее достоверных распределений нужно увеличить число опытов. Разновидность метода категориальных оценок состоит в предъявлении двух стимулов: естественной речи и синтезированной, всегда в одном и том же порядке, с инструкцией оценить в баллах степень похожести по натуральности синтетической речи на естественную. Конечно, при этом необходимо соблюдать определенные правила: звукосочетания должны быть одни и те же, частота основного тока или интонация (для фраз) также должны соответствовать друг другу.

§ 11.5. Восприятие в помехах

Восприятие речи в помехах существенно отличается от восприятия в идеальных акустических условиях. В экспериментах по маскировке слогов ГСГ белым шумом обнаружено перераспределение вероятностей правильного приема и перепутывания согласных в зависимости от уровня шума [59]. Так, согласный /Б/ принимается лучше, чем /Н/ до отношений сигнал/шум +6 дБ (по отношению к среднему по множеству слогов максимальному уровню гласных), а при худших соотношениях эти согласные меняются местами, причем разница в разборчивости возрастает пропорционально отношению шум/сигнал. Аналогично, согласный /Ш/ воспринимается надежнее, чем /З/ и /Ф/ до отношений примерно +2 дБ, а при худших отношениях разборчивость /Ш/ становится значительно ниже, чем /З/ и /Ф/.

В [18] найдено, что с изменением уровня шума число и соотношение лингвистических факторов, влияющих на разборчивость, может меняться. Так, лингвистический признак «часть речи» наиболее важен при распознавании слов в условиях, когда отношение сигнал/шум равно +4 дБ. Этот же признак занимает лишь четвертое место при отношениях −6 дБ. Чем выше уровень шума, тем большую роль в распознавании слов играет их частота встречаемости в речи. По [163] для наиболее часто встречающихся слов отношение сигнал/шум может быть на 20 дБ хуже при равной разборчивости с менее частыми словами.

Как видно из рис. 11.9, приведенного выше при обсуждении роли уровня громкости, ухудшение отношения сигнал/шум на 10 дБ приводит к падению слоговой разборчивости в среднем на 20—25% в диапазоне ниже 70% слоговой разборчивости при отсутствии шумов. В области оптимальной громкости речи (около 90 дБ) влияние шумов нелинейно зависит от их уровня.

Разборчивость формантного синтеза падает гораздо быстрее, чем разборчивость естественной речи при возрастании уровня шумов. По данным [41] при отношении сигнал/шум +28 дБ словесная разборчивость естественной речи равна 96,6%, синтезатора DECTalk — 92,2%, Prose — 62,8%, Votrax — 28%. При отношении 10 дБ разборчивость естественной речи падает до 92,2%, т. е. на 4,4%, тогда как разборчивость наилучшего синтезатора DECTalk оказывается равной всего 29,1%, т. е. падает на 63%. Это, конечно, свидетельствует о малой избыточности акустических признаков, заложенных в систему формантного синтеза по правилам. Не менее важным следствием является необходимость измерения разборчивости синтетической речи при разных уровнях шумов, поскольку соотношение разборчивости синтетической и естественной речи при отсутствии шумов не сохраняется при наличии шумов.

Зависимость разборчивости от уровня шумов меняется при смене вида шума. Наиболее часто для тестирования используется белый шум (с равномерным спектром) и шум Хота

(с падением −3 дБ/окт в речевом диапазоне частот). Измерение этих характеристик необходимо для сравнения разных синтезаторов. Если предусматривается использование синтезатора в присутствии шумов специфического вида, нужно провести измерение разборчивости и для данного шума.

Зависимостью числа ошибок восприятия от уровня шумов целесообразно воспользоваться на конечных этапах оптимизации синтезатора, когда достигнута довольно высокая разборчивость, и для оценки тенденций нужно проводить все больше и больше экспериментов. Если, например, ошибки близки к 1%, то в среднем нужно не менее 100 опытов для их появления. Использование шумов подходящего уровня увеличивает число ошибок и смещает оптимизируемую функцию разборчивости в зону наибольшей чувствительности. Пользуясь этим приемом, нужно помнить и о качественных изменениях в восприятии речи при изменении отношения сигнал/шум, которые обсуждались выше.

Случайные шумы — не единственный вид помех. Восприятие речи (естественной или синтетической) на фоне другой речи (естественой или синтетической) также относится к случаю с помехами. Чем больше синтетическая речь отличается от естественной, тем лучше ее обнаружение на фоне других разговоров, хотя разборчивость при этом невысока. Разборчивость синтетической речи на фоне синтетической речи того же синтезатора хуже разборчивости естественной речи на фоне естественной речи [165]. Поэтому в число характеристик синтезатора должны входить оценки разборчивости на фоне других речевых сигналов (естественных и синтетических) определенного уровня.

Помимо помех, т. е. посторонних звуков, на разборчивость влияют также различные искажения — частотные, временные, амплитудные. Известно, что ограничение частотного диапазона сверху или снизу снижает разборчивость (см. рис. 11.1). Отсюда следует необходимость оценки разборчивости синтезатора в типичных условиях его применения — в телефонах и радиоканалах, в больших помещениях с длительной реверберацией и т. п.

§ 11.6. Сложность восприятия

Восприятие синтетической речи представляет бо́льшую сложность для мозга, чем восприятие естественной речи, требует больших умственных усилий, занимает больший объем кратковременной оперативной памяти, характеризуется большим временем задержки распознавания. Некоторые из этих свойств обсуждались ранее в § 11.1. Всякое отличие характеристик синтетической речи от характеристик естественной речи, по-видимому, затрудняет обработку сигнала мозгом и дольше задерживает его в оперативной памяти. Одним из проявлений перегрузки оперативной памяти является отключение восприятия при прослушивании слитной синтетической речи. Этот эффект состоит в том, что регулярно выпадают целые куски

речевого сообщения, хотя слушатель и осознает, что речь не прерывалась. Затем восприятие вновь включается, а через некоторое время опять наступает отключение. Этот эффект наблюдается не только для формантных синтезаторов, но и для компиляционных (слоговых) синтезаторов. По-видимому, сложность обработки синтетической речи приводит к тому, что кратковременная память не успевает освобождаться для приема новых речевых сообщений и однажды наступает такой момент, когда память занята полностью, и вновь пришедший речевой сигнал просто некуда записать. После того, как часть речевого сообщения понята, оперативная память освобождается от старого речевого сигнала и готова к приему нового.

Из всех упомянутых характеристик сложности восприятия речевого сообщения, легче всего измерять время задержки распознавания. Это время достаточно хорошо характеризует сложность обработки синтетической речи для мозга и может служить удобной интегральной оценкой качества синтеза. Одна из методик измерения времени задержки состоит в обнаружении заданного слога в последовательности других слогов и нажатии кнопки как можно быстрее. Другая методика состоит в обнаружении осмысленного слова в последовательности, включающей и бессмысленные слова. Обнаружено, что для высококачественного формантного синтезатора (*DECTalk*) задержки в решении «слово/не слово» на 140—150 мс больше, чем для естественной речи [41, 164]. Наиболее чувствительна методика, в которой требуется определить истинность или ложность фразы и отреагировать нажатием одной из двух кнопок. После прослушивания фразы ее требуется записать, и ответы «истина/ложь» принимаются к обработке лишь в том случае, если транскрипция фразы верна [156, 176]. Посредством этой методики было найдено, что задержка на истинные фразы для вокодера с линейным предсказанием была равна 1,4—1,8 с, а для дельта-модуляции — примерно 0,7 с, тогда как для ложных фраз задержки были 1,7—2,2 с и 0,9 с соответственно. Для фраз, состоящих из трех и шести слов, разница в задержке восприятия между синтетической (*DECTalk*) и естественной речью составляет примерно 200 мс [175].

На время задержки влияет целый ряд факторов. Так, фразы с монотонной интонацией дают задержку на 50 мс больше, чем фразы с естественной интонацией. Степень предсказуемости фразы может изменить задержку для естественной речи на 250 мс, для синтеза — до 350 мс [175]. Время задержки увеличивается для длинных фраз и ложных сообщений [176].

При восприятии естественной речи на фоне других разговоров, или синтетической речи на фоне естественной речи задержка слогов СГ увеличивается на 50—70 мс [165]. Все эти факторы нужно учитывать при формировании теста на задержки восприятия синтетической речи, однако главное влияние на время задержки оказывает фонемная разборчивость.

ПРИЛОЖЕНИЕ

ГЕОМЕТРИЧЕСКИЕ И АКУСТИЧЕСКИЕ ХАРАКТЕРИСТИКИ РЕЧЕВОГО ТРАКТА ДЛЯ ЗВУКОВ РУССКОЙ РЕЧИ

Наряду с математическим адекватным описанием физических процессов речеобразования, для успешного исследования методов синтеза речи необходимо располагать обширным набором количественных данных о геометрических и акустических параметрах речевого тракта. Накопление этих данных связано с трудоемкими и дорогостоящими экспериментами, поскольку сведения о форме и динамике артикуляторов могут быть получены только с помощью кинорентгеносъемки или микролучевого рентгеноскопа. Без этих сведений невозможно определить координаты и параметры артикуляторных органов, соответствующие звукам того или иного языка, и, следовательно, формировать команды управления синтезатором иначе, как методом проб и ошибок. Поэтому в настоящем Приложении приводятся основные количественные данные о параметрах речевого тракта, позволяющие разрабатывать артикуляторные синтезаторы речи.

Все эти данные получены путем измерения параметров лишь одного диктора, однако посредством непрерывной деформации разных участков речевого тракта можно получить бесконечное множество индивидуальностей, в том числе и такие, которые не встречаются в действительности. Манипуляция геометрическими и механическими параметрами речевого тракта, а также изменение формы и длительности управляющих команд дают возможность исследовать патологию речеобразования и особенности певческого голоса. Контроль за реализацией изменений формы речевого тракта возможен при условии, что ЭВМ, на которой ведутся эти преобразования, обладает достаточно богатыми средствами графического обеспечения.

Все приводимые ниже данные соответствуют геометрическим размерам речевого тракта в средне-сагиттальной плоскости XOY.

Форма поверхности языка определяется пятью собственными функциями, суммирующимися с разными коэффициентами. Четыре собственные функции для всего языка и одна функция для кончика языка (точнее, для передней половины языка) описываются через тригонометрические и гиперболические функции (см. Введение). Вычисление собственных функций языка требует двойной точности на ЭВМ, но в силу того, что сами собственные функции не меняются в процессе артикуляции, имеет смысл вычислить их только один раз, и хранить в табулированном виде. В табл. П1 собственные функции языка представлены своими отсчетами в 50 точках. Этого числа отсчетов вполне достаточно для описания формы языка, а следовательно, и площади поперечного речевого тракта.

Поскольку в нейтральном состоянии форма поверхности языка очень близка к дуге полуокружности, то вычисления собственных функций выполнены в полярной системе координат от $0°$ до $180°$, так что каждый отсчет соответствует $3{,}6°$. Отсчеты собственной функции для кончика языка начинаются от угла радиус-вектора, равного $90°$, или с 26-го отсчета.

Под действием внешних и внутренних мышц в процессе артикуляции кончик языка может смещаться далеко назад, в результате чего он занимает

Таблица П1. Собственные функции языка

N	ψ_1	ψ_2	ψ_3	ψ_4	ψ_5
1	0,0000	0,0000	0,0000	0,0000	0,0000
2	−0,0484	0,1440	−0,2074	0,2688	0,0000
3	−0,0964	0,2852	−0,4062	0,5188	0,0000
4	−0,1436	0,4207	−0,5883	0,7326	0,0000
5	−0,1895	0,5479	−0,7463	0,8951	0,0000
6	−0,2337	0,6642	−0,8736	0,9950	0,0000
7	−0,2759	0,7674	−0,9649	1,0254	0,0000
8	−0,3165	0,8553	−1,0166	0,9841	0,0000
9	−0,3524	0,9263	−1,0266	0,8740	0,0000
10	−0,3861	0,9790	−0,9943	0,7028	0,0000
11	−0,4163	1,0123	−0,9211	0,4825	0,0000
12	−0,4426	1,0256	−0,8101	0,2284	0,0000
13	−0,4648	1,0187	−0,6657	−0,0416	0,0000
14	−0,4826	0,9918	−0,4940	−0,3087	0,0000
15	−0,4957	0,9453	−0,3019	−0,5542	0,0000
16	−0,5040	0,8803	−0,0973	−0,7610	0,0000
17	−0,5072	0,7982	0,1113	−0,9146	0,0000
18	−0,5053	0,7005	0,3155	−1,0042	0,0000
19	−0,4980	0,5894	0,5067	−1,0235	0,0000
20	−0,4853	0,4671	0,6773	−0,9713	0,0000
21	−0,4672	0,3361	0,8202	−0,8512	0,0000
22	−0,4434	0,1991	0,9296	−0,6715	0,0000
23	−0,4142	0,0589	1,0010	−0,4448	0,0000
24	−0,3795	−0,0815	1,0315	−0,1869	0,0000
25	−0,3393	−0,2192	1,0199	0,0841	0,0000
26	−0,2937	−0,3514	0,9669	0,3494	0,0012
27	−0,2429	−0,4752	0,8746	0,5903	0,0110
28	−0,1869	−0,5880	0,7469	0,7903	0,0300
29	−0,1260	−0,6873	0,5892	0,9352	0,0577
30	−0,0603	−0,7710	0,4082	1,0150	0,0935
31	0,0099	−0,8370	0,2114	1,0244	0,1368
32	0,0846	−0,8838	0,0072	0,9627	0,1872
33	0,1633	−0,9101	−0,1958	0,8343	0,2440
34	0,2459	−0,9149	−0,3889	0,6486	0,3067
35	0,3321	−0,8978	−0,5638	0,4188	0,3748
36	0,4217	−0,8585	−0,7128	0,1612	0,4477
37	0,5142	−0,7973	−0,8292	−0,1057	0,5250
38	0,6096	−0,7147	−0,9075	−0,3625	0,6062
39	0,7074	−0,6117	−0,9437	−0,5907	0,6907
40	0,8074	−0,4895	−0,9353	−0,7732	0,7780
41	0,9092	−0,3494	−0,8812	−0,8960	0,8679
42	1,0128	−0,1934	−0,7824	−0,9489	0,9598
43	1,1177	−0,0232	−0,6408	−0,9260	1,0534
44	1,2237	0,1592	−0,4602	−0,8262	1,1484
45	1,3307	0,3517	−0,2452	−0,6528	1,2443
46	1,4383	0,5523	−0,0015	−0,4133	1,3410
47	1,5464	0,7590	0,2651	−0,1185	1,4381
48	1,6549	0,9699	0,5484	0,2188	1,5356
49	1,7636	1,1836	0,8427	0,5851	1,6332
50	1,8724	1,3988	1,1429	0,9677	1,7309

положение, соответствующее углу в полярной системе координат, меньшему 180°. В этом случае количество отсчетов для собственных функций сохраняется равным 50, но угол между соседними отсчетами уменьшается соответственно уменьшению значения угла, соответствующего положению кончика языка.

В таблицах П2, П3, П4 представлены координаты (x, y) в неподвижной системе координат XOY неподвижных поверхностей речевого тракта, и таких поверхностей, которые движутся и деформируются под воздействием смещения артикуляторных органов. В табл. П2 и П3 приводятся координаты, соответственно, задней и передней поверхностей фарингиальной области, начинающейся от голосовой щели. В нейтральном положении высота голосовой щели, т. е. вертикальная координата y, принята равной 1,6 см. С этой величины начинаются отсчеты в табл. П2 и П3. Следует обратить внимание на то, что все данные в этих таблицах о координатах поверхностей речевого тракта *приведены с масштабным коэффициентом 1,25, что было связано с условиями* измерения на кинорентгенограммах. Поэтому каждое значение в табл. П2 — П5 нужно делить на 1,25.

Таблица П2. Координаты задней поверхности фарингиальной области

x	y	x	y	x	y	x	y
2,7	1,6	2,3	2,8	1,9	4,0	1,0	4,4
2,7	1,8	2,2	3,0	1,8	4,2	0,8	4,2
2,6	2,0	2,2	3,2	1,7	4,4	0,7	4,0
2,5	2,2	2,1	3,4	1,6	4,6	0,6	3,8
2,45	2,4	2,05	3,6	1,4	4,8	0,3	3,6
2,4	2,6	2,00	3,8	1,1	4,6	0,0	3,6

Таблица П3. Координаты передней поверхности фарингиальной области

x	y	x	y	x	y	x	y
4,5	1,6	3,2	3,0	2,55	4,4	2,4	5,8
4,35	1,8	3,1	3,2	2,5	4,6	2,35	6,0
4,2	2,0	3,0	3,4	2,5	4,8	2,2	6,2
4,0	2,2	2,85	3,6	2,5	5,0	2,15	6,4
3,75	2,4	2,75	3,8	2,45	5,2	2,05	6,6
3,6	2,6	2,65	4,0	2,45	5,4	1,85	6,8
3,4	2,8	2,6	4,2	2,4	5,6	1,7	7,0

Таблица П4. Форма верхней поверхности речевого тракта

x	y	x	y	x	y	x	y
0,0	12,4	3,2	13,95	6,6	14,35	10,0	12,8
0,2	12,5	3,4	13,95	6,8	14,3	10,1	12,5
0,4	12,6	3,6	13,9	7,0	14,3	10,2	12,3
0,6	12,7	3,8	13,9	7,2	14,3	10,3	12,1
0,8	12,8	4,0	13,9	7,4	14,25	10,4	11,9
1,0	12,9	4,2	13,95	7,6	14,2	10,6	11,5
1,2	12,05	4,4	14,0	7,8	14,1	10,6	11,6
1,4	13,2	4,6	14,0	8,0	14,1	11,0	11,4
1,6	13,4	4,8	14,1	8,2	14,0	11,2	11,3
1,7	13,6	5,0	14,15	8,4	13,9	11,4	11,25
1,85	13,8	5,2	14,2	8,6	13,8	11,6	11,2
2,05	14,0	5,4	14,25	8,8	13,65	11,8	11,3
2,25	14,1	5,6	14,3	9,0	13,55	12,0	11,4
2,4	14,1	5,8	14,3	9,2	13,45	12,2	11,6
2,6	14,0	6,0	14,3	9,4	13,3	12,4	11,8
2,8	14,0	6,2	14,3	9,6	13,15		
3,0	13,95	6,4	14,35	9,8	13,0		

Таблица П5. Подъязычная поверхность

x	y	x	y	x	y
−2,7	3,1	−1,3	3,7	0,0	4,8
−2,5	3,15	−1,1	3,8	0,0	4,6
−2,3	3,25	−0,9	4,0	0,0	4,4
−2,1	3,3	−0,7	4,15	0,0	4,2
−1,9	3,4	−0,5	4,35	0,0	4,0
−1,7	3,5	−0,3	4,6		
−1,5	3,6	−0,1	4,7		

От фарингиальной области до небной занавески принимается, что задняя поверхность речевого тракта расположена строго на вертикальной оси Y, и последняя координата x задней поверхности фарингиальной области в табл. П2, *таким образом, равна нулю. Последний отсчет передней поверхности фарингиальной области согласуется с положением корня языка.* Как задняя, так и передняя поверхности фарингиальной области деформируются в соответствии с высотой голосовой щели и положением корня языка на плоскости XOY. Могут быть использованы различные алгоритмы такой непрерывной деформации, и простейший из них состоит в линейном преобразовании координат (x, y) *этих поверхностей в зависимости от положения* голосовой щели и корня языка.

Форма верхней поверхности речевого тракта — от небной занавески до верхней губы, представлена своими координатами x и y в табл. П4, причем по оси X используются равномерные отсчеты через 0,2/1,25 см. При необходимости моделирования опускания небной занавески *участок на протяжении* примерно 3 см поворачивается в локальной полярной системе координат с центром, находящимся на верхней поверхности тракта. При огублении, когда нижняя и верхняя губы вытягиваются вдоль оси X, участок, соответствующий верхней губе, линейно деформируется в этом направлении путем изменения расстояния между отсчетами координаты x.

Когда язык смещается назад, обнажается подъязычная поверхность, которая распространяется от нижнего зуба в направлении задней стенки речевого тракта. Эта поверхность представлена в табл. П5, причем координаты (x, y) определяются в подвижной системе координат $X_1O_1Y_1$, связанной с нижней челюстью, так что положение *подъязычной поверхности в неподвижной* системе координат XOY зависит от положения нижней челюсти.

Форма нижней губы представлена в табл. П6 *своими координатами* (x, y) также в подвижной системе координат $X_1O_1Y_1$. При огублении координаты вдоль оси X линейно смещаются в соответствии со степенью огубления синхронно с деформацией поверхности верхней губы.

Таблица П6. Форма нижней губы

x	y	x	y
0,0	0,0	1,2	0,4
0,25	0,15	1,4	0,35
0,5	0,25	1,6	0,3
0,76	0,35	1,8	0,25
1,0	*0,4*	*2,0*	*0,15*

Сведения о каждом гласном звуке сгруппированы в три таблицы, содержащие информацию об артикуляторных параметрах, площади поперечного сечения, частотах и затуханиях *резонансов речевого тракта.* Кроме того, на четырех рисунках показывается форма речевого тракта, его амплитудно-частотные характеристики, форма площади поперечного сечения и три первые собственные функции для акустического давления. Например, коор-

динаты артикуляторных органов для гласного /А/ показаны в табл. П8. Эти координаты служат целевыми значениями для команд управления артикуляцией. Соотвстствующая этим артикуляторным координатам площадь поперечного сечения табулирована в табл. П7 с шагом вдоль средней линии тракта, равным 4,4 мм, что соответствует тактовой частоте сдвига бегущих волн, примерно равной 80 кГц. В табл. П9 показана частота резонанса радиальных колебаний в речевом тракте, связанных с податливостью его стенок, а также шесть первых резонансных частот для каждого резонанса. Эти резонансные частоты, а также показанные на рис. П4 три первых собственных функции, рассчитаны методом Галеркина.

Таблица П7. Площадь поперечного сечения речевого тракта для гласного /А/

Координата вдоль оси тракта, см	Площадь поперечного сечения, $см^2$	Координата вдоль оси тракта, см	Площадь поперечного сечения, $см^2$
0,00	2,75	10,12	4,39
0,44	2,17	10,56	4,65
0,88	1,47	11,00	5,00
1,32	1,03	11,44	6,24
1,76	0,81	11,88	7,28
2,20	0,65	12,32	7,10
2,64	0,72	12,76	7,02
3,08	1,14	13,20	7,41
3,52	4,38	13,64	8,15
3,96	4,11	14,08	8,82
4,40	3,73	14,52	9,16
4,84	3,25	14,96	8,67
5,28	2,27	15,40	7,56
5,72	1,08	15,84	6,24
6,16	1,87	16,28	4,92
6,60	1,31	16,72	3,53
7,04	0,87	17,16	2,48
7,48	0,61	17,60	2,07
7,92	0,86	18,04	3,46
8,36	1,38	18,48	4,54
8,80	2,00	18,92	3,94
9,24	2,70	19,36	4,37
9,68	3,47		

Таблица П8. Артикуляторные параметры для гласного /А/

Ширина голосовой щели	0,0
Горизонтальный сдвиг нижней челюсти	0,0
Угол поворота нижней челюсти, рад	0,16
Горизонтальная координата корня языка	−7,2
Вертикальная координата корня языка	−1,0
Горизонтальная координата кончика языка	−1,4
Вертикальная координата кончика языка	4,0
Угол поворота небной занавески, рад	0,0
Длина губ	4,0
Вертикальная координата нижней губы	0,4
Вертикальная координата голосовой щели	2,0
Коэффициент при 1-й собственной функции языка	−0,425
Коэффициент при 2-й собственной функции языка	0,92
Коэффициент при 3-й собственной функции языка	0,05
Коэффициент при 4-й собственной функции языка	−0,1
Коэффициент при 5-й собственной функции языка	−0,3

Таблица П9. Резонансные частоты и ширина полос для гласного /А/, Гц

	F_p	F_1	F_2	F_3	F_4	F_5	F_6
Частота	273,5	574,6	994,1	2404,8	2711,4	3796,5	4735,3
Ширина полосы	72,4	78,1	48,3	77,7	102,5	145,6	221,8

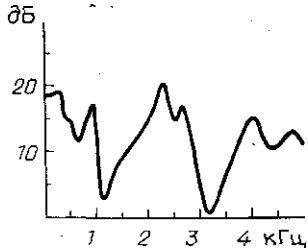


Рис. П1. Амплитудно-частотная характеристика гласного /А/

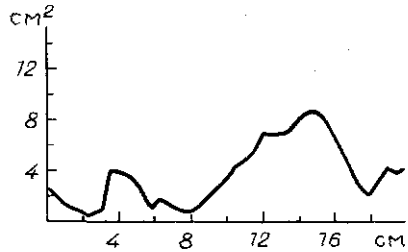


Рис. П2. Площадь поперечного сечения для гласного /А/

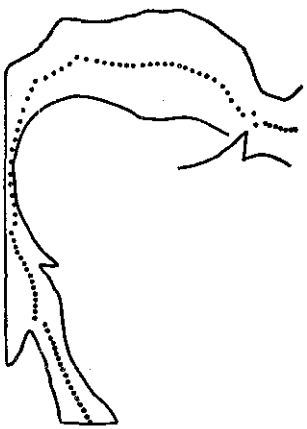


Рис. П3. Форма речевого тракта для гласного /А/

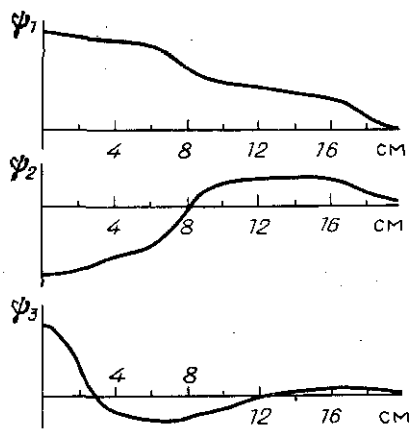


Рис. П4. Собственные функции акустического давления для гласного /А/

Таблица П10. *Площадь поперечного сечения речевого тракта для гласного /О/*

Координата вдоль оси тракта, см	Площадь поперечного сечения, $см^2$	Координата вдоль оси тракта, см	Площадь поперечного сечения, $см^2$
0,00	2,75	10,12	3,37
0,44	2,17	10,56	3,49
0,88	1,47	11,00	3,73
1,32	1,03	11,44	4,35
1,76	0,81	11,88	5,11
2,20	0,65	12,32	5,12
2,64	0,72	12,76	5,28
3,08	1,15	13,20	5,80
3,52	4,52	13,64	6,86
3,96	4,27	14,08	8,26
4,40	3,82	14,52	9,21
4,84	3,35	14,96	9,68
5,28	2,52	15,40	8,84
5,72	1,61	15,84	7,61
6,16	2,21	16,28	6,06
6,60	1,66	16,72	4,39
7,04	1,18	17,16	3,05
7,48	0,74	17,60	2,60
7,92	0,56	18,04	1,63
8,36	0,84	18,48	2,63
8,80	1,39	18,92	1,82
9,24	2,03	19,36	2,11
9,68	2,75		

Таблица П11. Артикуляторные параметры для гласного /О/

Ширина голосовой щели	0,0
Горизонтальный сдвиг нижней челюсти	0,0
Угол поворота нижней челюсти, рад	0,07
Горизонтальная координата корня языка	−8,0
Вертикальная координата корня языка	−1,0
Горизонтальная координата кончика языка	−2,2
Вертикальная координата кончика языка	3,4,
Угол поворота небной занавески, рад	0,0
Длина губ	4,0
Вертикальная координата нижней губы	0,4
Вертикальная координата голосовой щели	1,0
Коэффициент при 1-й собственной функции языка	−0,85
Коэффициент при 2-й собственной функции языка	0,9
Коэффициент при 3-й собственной функции языка	0,3
Коэффициент при 4-й собственной функции языка	−0,0875
Коэффициент при 5-й собственной функции языка	−0,4

На рис. П3 показана форма речевого тракта в средне-сагиттальной плоскости и (пунктиром) — средняя линия, определяющая длину речевого тракта. На рис. П1 показана собственная амплитудно-частотная характеристика речевого тракта (без влияния источника возбуждения), рассчитанная по резонансным частотам и затуханиям, а также относительным амплитудам резонансов, полученным методом Галеркина. Первый максимум на этой характеристике присутствует условно, он отображает резонанс радиальных колебаний. Во время звонкой смычки в пространство излучаются колебания, главным образом, на частоте радиального резонанса. На рис. П2 показана форма площади поперечного сечения тракта, а на рис. П4 приводятся собственные функции для акустического давления.

Таблица П12. Резонансные частоты и ширина полос для гласного /О/, Гц

	F_0	F_1	F_2	F_3	F_4	F_5	F_6
Частота	287,6	497,1	914,2	2316,4	2635,1	4030,9	4728,3
Ширина полосы	72,4	100,9	47,1	67,9	87,6	142,3	189,5

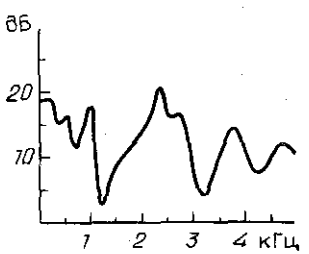


Рис. П5. Амплитудно-частотная характеристика гласного /О/

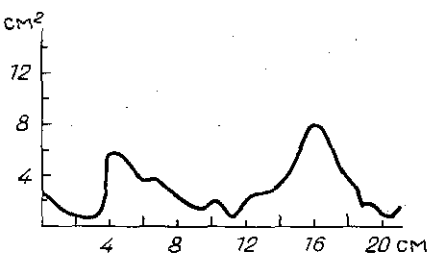


Рис. П6. Площадь поперечного сечения тракта для гласного /О/

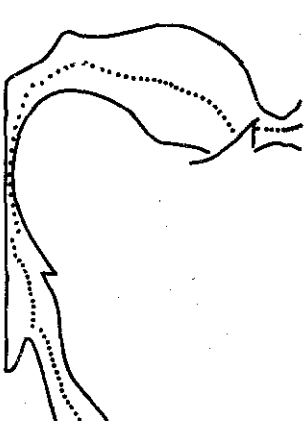


Рис. П7. Форма речевого тракта для гласного /О/

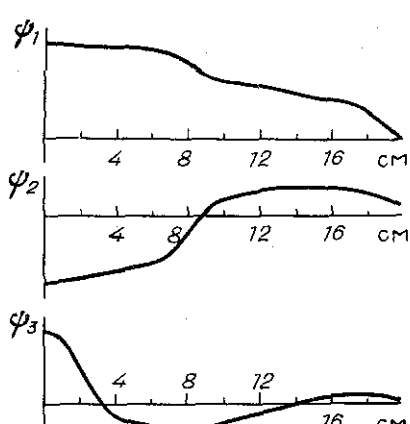


Рис. П8. Собственные функции акустического давления для гласного /О/

Таблица П13. Площадь поперечного сечения речевого тракта для гласного /У/

Координата вдоль оси тракта, см	Площадь поперечного сечения, $см^2$	Координата вдоль оси тракта, см	Площадь поперечного сечения, $см^2$
0,00	2,75	11,00	0,70
0,44	2,23	11,44	0,76
0,88	1,56	11,88	1,82
1,32	1,16	12,32	2,54
1,76	0,89	12,76	2,57
2,20	0,69	13,20	2,58
2,64	0,67	13,64	2,83
3,08	0,74	14,08	3,47
3,52	1,37	14,52	4,25
3,96	5,90	14,96	5,36
4,40	5,84	15,40	6,84
4,84	5,39	15,84	8,20
5,28	4,88	16,28	8,09
5,72	3,93	16,72	7,37
6,16	3,50	17,16	6,15
6,60	3,87	17,60	4,47
7,04	3,42	18,04	3,90
7,48	2,87	18,48	2,87
7,92	2,40	18,92	1,47
8,36	1,98	19,36	1,69
8,80	1,64	19,80	1,10
9,24	1,45	20,24	0,66
9,68	1,54	20,68	0,61
10,12	2,02	21,12	1,17
10,56	1,57		

Таблица П.14. Артикуляторные параметры для гласного /У/

Ширина голосовой щели	0,0
Горизонтальный сдвиг нижней челюсти	0.0
Угол поворота нижней челюсти, рад	0,055
Горизонтальная координата корня языка	−7,4
Вертикальная координата корня языка	−0,45
Горизонтальная координата кончика языка	−3,0
Вертикальная координата кончика языка	4,0
Угол поворота небной занавески, рад	0,0
Длина губ	3,6
Вертикальная координата нижней губы	0,0
Вертикальная координата голосовой щели	−0,5
Коэффициент при 1-й собственной функции языка	−1,8
Коэффициент при 2-й собственной функции языка	0,3
Коэффициент при 3-й собственной функции языка	0,6
Коэффициент при 4-й собственной функции языка	−0,12
Коэффициент при 5-й собственной функции языка	−0,025

Собственные функции используются при расчете коэффициентов затухания δ_k, а значения собственных функций $\psi_k(l)$ на излучающем конце речевого тракта определяют относительные амплитуды резонансных колебаний. Упрощенно для речевого сигнала можно записать

$$f(t)=\sum_k \psi_k(l)\, e^{-\delta_k t}\cos \omega_k t,$$

где ω_k — резонансные частоты. Кроме того, форму собственных функций, их

Таблица П15. Резонансные частоты и ширина полос для гласного /У/, Гц

	F_0	F_1	F_2	F_3	F_4	F_5	F_6
Частота	296,8	408,6	858,0	2042,8	2761,3	3612,3	4434,3
Ширина полосы	72,4	149,2	41,9	54,2	71,2	92,4	122,7

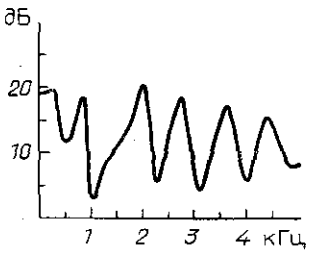


Рис. П9. Амплитудно-частотная характеристика гласного /У/

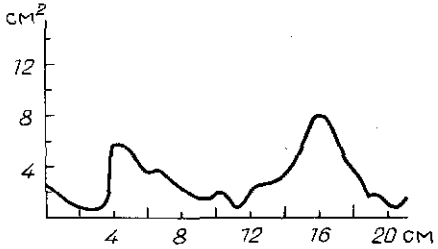


Рис. П10. Площадь поперечного сечения тракта для гласного /У/

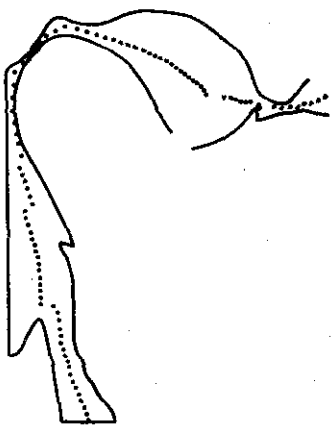


Рис. П11. Форма речевого тракта для гласного /У/

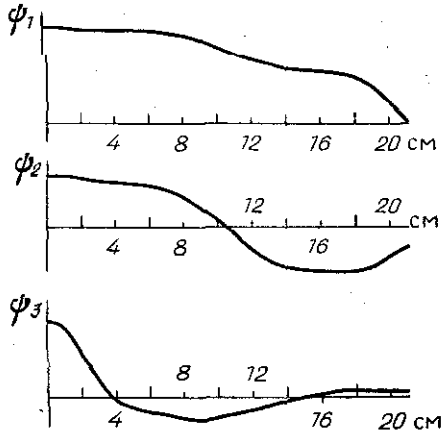


Рис. П12. Собственные функции акустического давления для гласного /У/

Таблица П16. Площадь поперечного сечения речевого тракта для гласного /И/

Координата вдоль оси тракта, см	Площадь поперечного сечения, $см^2$	Координата вдоль оси тракта, см	Площадь поперечного сечения, $см^2$
0,00	2,75	8,36	4,40
0,44	1,53	8,80	3,45
0,88	0,88	9,24	2,82
1,32	0,68	9,68	2,40
1,76	5,68	10,12	1,91
2,20	8,18	10,56	1,59
2,64	7,48	11,00	1,24
3,08	7,26	11,44	0,89
3,52	7,28	11,88	0,81
3,96	8,54	12,32	0,67
4,40	8,71	12,76	0,70
4,84	7,94	13,20	0,69
5,28	8,12	13,64	0,58
5,72	8,37	14,08	0,54
6,16	9,09	14,52	0,70
6,60	9,72	14,96	2,28
7,04	8,19	15,40	1,76
7,48	7,54	15,84	1,27
7,92	6,17	16,28	2,27

Таблица П17. Артикуляторные параметры для гласного /И/

Ширина голосовой щели	0,0
Горизонтальный сдвиг нижней челюсти	0,0
Угол поворота нижней челюсти, рад	0,07
Горизонтальная координата корня языка	−6,2
Вертикальная координата корня языка	1,0
Горизонтальная координата кончика языка	−0,4
Вертикальная координата кончика языка	4,3
Угол поворота небной занавески, рад	0,0
Длина губ	4,0
Вертикальная координата нижней губы	0,2
Вертикальная координата голосовой щели	2,0
Коэффициент при 1-й собственной функции языка	−0,2125
Коэффициент при 2-й собственной функции языка	−0,35
Коэффициент при 3-й собственной функции языка	0,15
Коэффициент при 4-й собственной функции языка	−0,05
Коэффициент при 5-й собственной функции языка	−0,65

нули и экстремумы нужно знать для того, чтобы определить место деформации речевого тракта при необходимости коррекции резонансных частот. Например, сужение в области пучности (максимума и минимума) собственной функции повышает частоту соответствующего резонанса, а сужение в области нуля собственной функции понижает соответствующий резонанс. При деформации речевого тракта меняются частоты всех резонансов, и знание формы собственных функций помогает отыскать оптимальное воздействие на форму речевого тракта.

Параметры гласных звуков /А, О, У, И, Ы, Э/ представлены в табл. П7—П24 и рис. П1—П24. Аналогично в табл. П25—П36 и рис. П25—П40 представлены параметры фрикативных звуков /С, Ш, Х, Ф/. В этих таблицах приведена та же информация, что и для гласных звуков: координаты артикуляторных органов, табулированная площадь поперечного сечения тракта,

Таблица П18. Резонансные частоты и ширина полос для гласного /И/, Гц

	F_0	F_1	F_2	F_3	F_4	F_5	F_6
Частота	287,7	393,5	2272,1	3094,6	4003,6	5047,3	6103,5
Ширина полосы	72,4	54,9	66,1	77,6	83,7	117,0	133,6

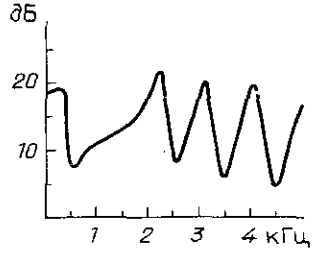


Рис. П13. Амплитудно-частотная характеристика гласного /И/

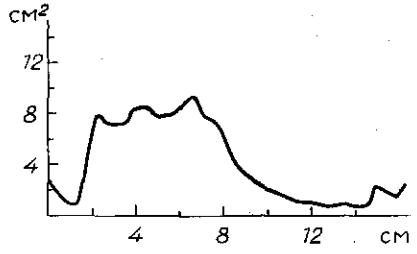


Рис. П14. Площадь поперечного сечения тракта для гласного /И/

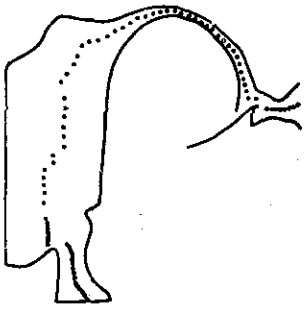


Рис. П15. Форма речевого тракта для гласного /И/

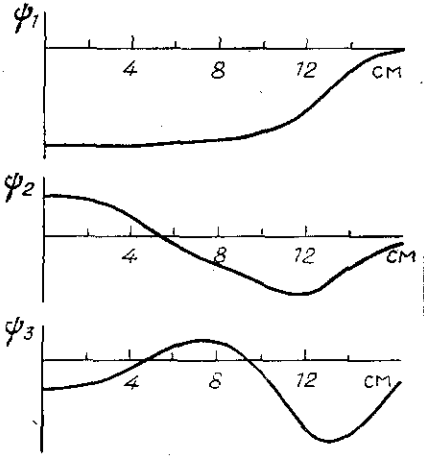


Рис. П16. Собственные функции акустического давления для гласного /И/

Таблица П19. Площадь поперечного сечения речевого тракта для гласного /Ы/

Координата вдоль оси тракта, см	Площадь поперечного сечения, $см^2$	Координата вдоль оси тракта, см	Площадь поперечного сечения, $см^2$
0,00	2,75	10,56	4,09
0,44	2,27	11,00	3,58
0,88	1,67	11,44	3,49
1,32	1,27	11,88	3,45
1,76	0,94	12,32	2,58
2,20	0,79	12,76	1,78
2,64	0,65	13,20	1,40
3,08	0,69	13,64	1,57
3,52	0,83	14,08	2,08
3,96	2,63	14,52	2,54
4,40	5,31	14,96	3,03
4,84	5,21	15,40	3,27
5,28	4,72	15,84	3,23
5,72	4,35	16,28	3,22
6,16	3,70	16,72	3,39
6,60	3,58	17,16	3,41
7,04	4,41	17,60	2,97
7,48	4,68	18,04	2,82
7,92	4,51	18,48	2,14
8,36	4,16	18,92	1,12
8,80	4,12	19,36	2,67
9,24	4,13	19,80	1,71
9,68	4,22	20,24	1,64
10,12	4,51	20,68	2,84

Таблица П20. Артикуляторные параметры для гласного /Ы/

Ширина голосовой щели	0,0
Горизонтальный сдвиг нижней челюсти	0,0
Угол поворота нижней челюсти, рад	0,06
Горизонтальная координата корня языка	−7,7
Вертикальная координата корня языка	0,0
Горизонтальная координата кончика языка	−3,0
Вертикальная координата кончика языка	4,0
Угол поворота небной занавески, рад	0,0
Длина губ	4,0
Вертикальная координата нижней губы	0,4
Вертикальная координата голосовой щели	−1,0
Коэффициент при 1-й собственной функции языка	−0,2625
Коэффициент при 2-й собственной функции языка	−0,3375
Коэффициент при 3-й собственной функции языка	0,1875
Коэффициент при 4-й собственной функции языка	−0,05
Коэффициент при 5-й собственной функции языка	−0,44

а также частоты и затухания резонансов. Кроме этого, для каждого фрикативного согласного указывается характерное значение числа Рейнольдса и частота максимума в спектре источника турбулентного шума. На рисунках показаны форма и площадь поперечного сечения речевого тракта, а также три собственные функции подобно тому, как это было сделано для гласных звуков. Отличие заключается в амплитудно-частотных характеристиках, где вместо резонансов тракта показаны спектр шумового источника возбуждения (1) и спектр фрикативного звука (2), синтезированного методом бегущих волн.

Таблица П21. Резонансные частоты и ширина полос для гласного /Ы/, Гц

	F_0	F_1	F_2	F_3	F_4	F_5	F_6
Частота	302,6	485,7	1378,4	1874,7	2574,5	3732,5	4421,9
Ширина полосы	72,4	85,5	47,0	46,3	63,3	97,7	124,8

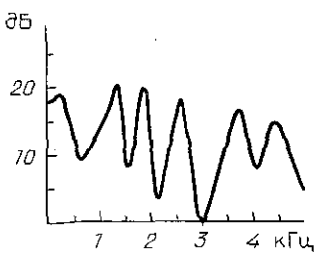


Рис. П17. Амплитудно-частотная характеристика гласного /Ы/

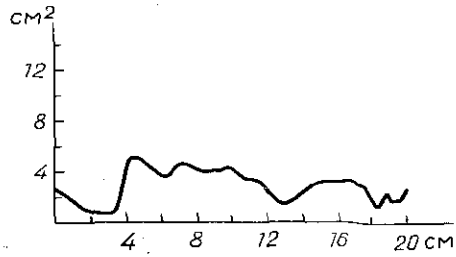


Рис. П18. Площадь поперечного сечения тракта для гласного /Ы/

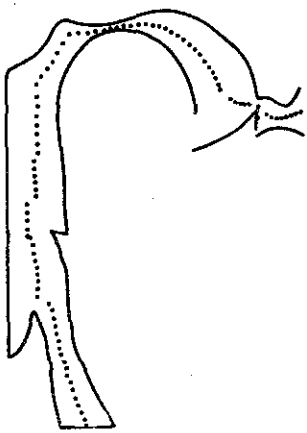


Рис. П19. Форма речевого тракта для гласного /Ы/

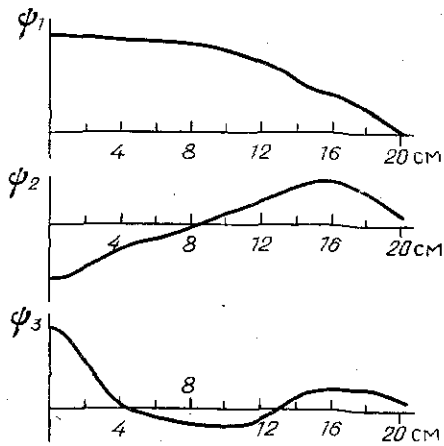


Рис. П20. Собственные функции акустического давления для гласного /Ы/

Таблица П22. Площадь поперечного сечения речевого тракта для гласного /Э/

Координата вдоль оси тракта, см	Площадь поперечного сечения, $см^2$	Координата вдоль оси тракта, см	Площадь поперечного сечения, $см^2$
0,00	2,75	10,12	6,08
0,44	2,17	10,56	5,46
0,88	1,47	11,00	5,62
1,32	1,03	11,44	6,19
1,76	0,81	11,88	5,62
2,20	0,65	12,32	4,83
2,64	0,72	12,76	4,63
3,08	1,16	13,20	4,69
3,52	4,86	13,64	5,01
3,96	4,65	14,08	5,25
4,40	4,06	14,52	5,48
4,84	3,59	14,96	5,17
5,28	3,27	15,40	4,59
5,72	3,13	15,84	3,90
6,16	3,95	16,28	3,15
6,60	3,87	16,72	2,38
7,04	3,81	17,16	1,60
7,48	3,80	17,60	1,22
7,92	3,84	18,04	0,85
8,36	3,96	18,48	2,79
8,80	4,18	18,92	2,99
9,24	4,51	19,36	2,45
9,68	5,49	19,80	3,17

Таблица П23. Артикуляторные параметры для гласного /Э/

Ширина голосовой щели	0,0
Горизонтальный сдвиг нижней челюсти	0,0
Угол поворота нижней челюсти, рад	0,1
Горизонтальная координата корня языка	−7,5
Вертикальная координата корня языка	−1,2
Горизонтальная координата кончика языка	−1,0
Вертикальная координата кончика языка	4,3
Угол поворота небной занавески, рад	0,0
Длина губ	4,0
Вертикальная координата нижней губы	0,4
Вертикальная координата голосовой щели	2,0
Коэффициент при 1-й собственной функции языка	0,3375
Коэффициент при 2-й собственной функции языка	0,0875
Коэффициент при 3-й собственной функции языка	0,225
Коэффициент при 4-й собственной функции языка	−0,05
Коэффициент при 5-й собственной функции языка	−0,3

Таблица П24. Резонансные частоты и ширина полос для гласного /Э/, Гц

	F_0	F_1	F_2	F_3	F_4	F_5	F_6
Частота	279,0	490,9	1353,0	2235,0	2775,0	3575,7	4226,4
Ширина полосы	72,4	73,1	41.4	60,8	78,5	109,4	141,3

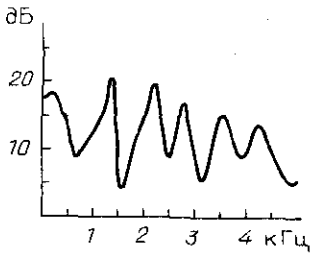


Рис. П21. Амплитудно-частотная характеристика гласного /Э/

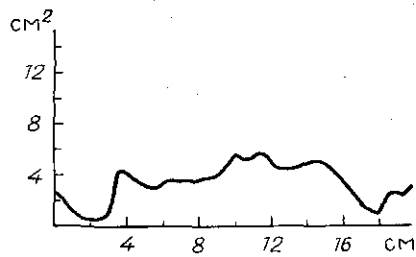


Рис. П22. Площадь поперечного сечения тракта для гласного /Э/

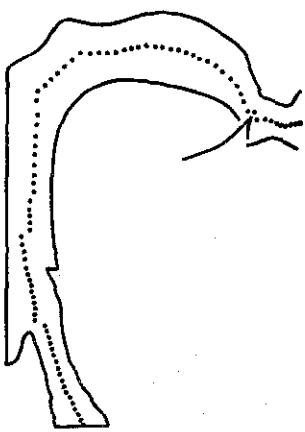


Рис. П23. Форма речевого тракта для гласного /Э/

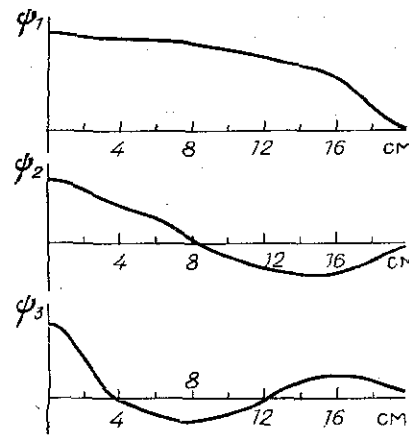


Рис. П24. Собственные функции акустического давления для гласного /Э/

Таблица П25. Площадь поперечного сечения речевого тракта для фрикативного /С/

Координата вдоль оси тракта, см	Площадь поперечного сечения, $см^2$	Координата вдоль оси тракта, см	Площадь поперечного сечения, $см^2$
0,00	2,75	9,24	4,54
0,44	1,53	9,68	4,15
0,88	0,88	10,12	3,91
1,32	0,68	10,56	3,94
1,76	3,38	11,00	4,03
2,20	4,06	11,44	4,00
2,64	3,32	11,88	3,92
3,08	3,23	12,32	3,52
3,52	3,09	12,76	2,80
3,96	3,70	13,20	2,21
4,40	3,81	13,64	1,64
4,84	3,72	14,08	1,18
5,28	3,66	14,52	0,78
5,72	3,67	14,96	0,42
6,16	3,77	15,40	0,25
6,60	3,99	15,84	0,17
7,04	4,34	16,28	1,56
7,48	4,64	16,72	1,79
7,92	4,34	17,16	1,69
8,36	4,45	17,60	2,89
8,80	4,79		

Таблица П26. Артикуляторные параметры для фрикативного /С/

Ширина голосовой щели	0,4
Горизонтальный сдвиг нижней челюсти	0,11
Угол поворота нижней челюсти, рад	0,08
Горизонтальная координата корня языка	−8,0
Вертикальная координата корня языка	−0,3
Горизонтальная координата кончика языка	−0,1
Вертикальная координата кончика языка	5,3
Угол поворота небной занавески, рад	0,0
Длина губ	4,0
Вертикальная координата нижней губы	0,2
Вертикальная координата голосовой щели	2,0
Коэффициент при 1-й собственной функции языка	0,3
Коэффициент при 2-й собственной функции языка	0,2
Коэффициент при 3-й собственной функции языка	0,125
Коэффициент при 4-й собственной функции языка	−0,182
Коэффициент при 5-й собственной функции языка	0,1

Таблица П27. Резонансные частоты и ширина полос для фрикативного /С/, Гц

	F_0	F_1	F_2	F_3	F_4	F_5	F_6
Частота	325,4	482,7	1619,4	2861,0	4029,8	4406,1	5290,6
Ширина полосы	72,4	72,7	45,7	72,7	106,3	115,9	153,9

Число Рейнольдса, Re = 3200
Частота первого максимума в спектре шумового источника $F_{1ш} = 2050$ Гц
Частота второго максимума в спектре шумового источника $F_{2ш} = 4100$ Гц

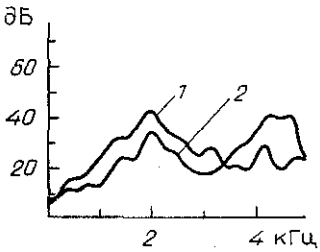


Рис. П25. Амплитудно-частотная характеристика: *1* — шумового источника, *2* — фрикативного /С/

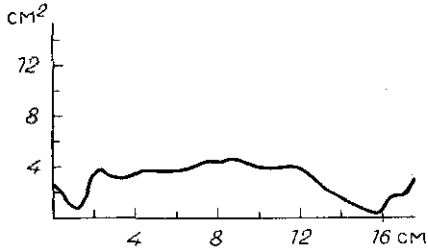


Рис. П26. Площадь поперечного сечения тракта для фрикативного /С/

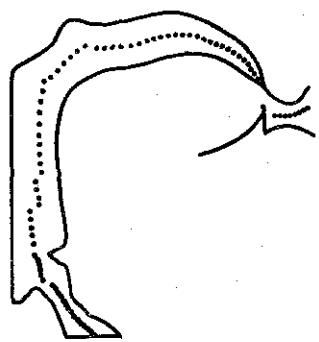


Рис. П27. Форма сечения тракта для фрикативного /С/

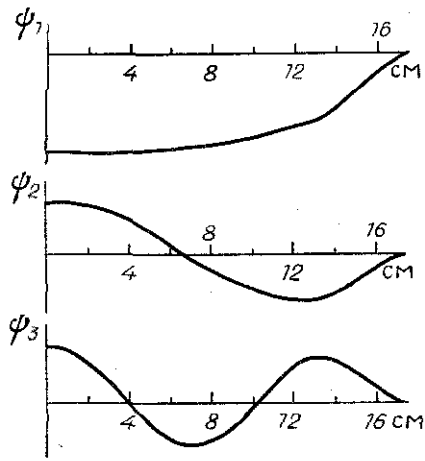


Рис. П28. Собственные функции акустического давления для фрикативного /С/

Таблица П28. Площадь поперечного сечения речевого тракта для фрикативного /Ш/

Координата вдоль оси тракта, см	Площадь поперечного сечения, $см^2$	Координата вдоль оси тракта, см	Площадь поперечного сечения, $см^2$
0,00	2,75	10,56	3,34
0,44	2,17	11,00	2,88
0,88	1,47	11,44	2,38
1,32	1,03	11,88	2,11
1,76	0,81	12,32	2,24
2,20	0,65	12,76	2,54
2,64	0,72	13,20	2,75
3,08	1,16	13,64	2,76
3,52	4,90	14,08	2,67
3,96	4,70	14,52	2,22
4,40	4,10	14,96	1,77
4,84	3,63	15,40	1,28
5,28	3,31	15,84	0,85
5,72	3,03	16,28	0,52
6,16	3,73	16,72	0,24
6,60	3,54	17,16	0,82
7,04	3,41	17,60	3,82
7,48	3,35	18,04	2,58
7,92	3,36	18,48	1,30
8,36	3,48	18,92	2,23
8,80	3,69	19,36	1,46
9,24	3,47	19,80	0,95
9,68	3,07	20,24	1,94
10,12	3,12		

Таблица П29. Артикуляторные параметры для фрикативного /Ш/

Ширина голосовой щели	0,4
Горизонтальный сдвиг нижней челюсти	0,06
Угол поворота нижней челюсти, рад	0,056
Горизонтальная координата корня языка	−8,0
Вертикальная координата корня языка	0,4
Горизонтальная координата кончика языка	−1,2
Вертикальная координата кончика языка	5,6
Угол поворота небной занавески, рад	0,0
Длина губ	4,0
Вертикальная координата нижней губы	0,2
Вертикальная координата голосовой щели	2,0
Коэффициент при 1-й собственной функции языка	0,0
Коэффициент при 2-й собственной функции языка	0,3
Коэффициент при 3-й собственной функции языка	0,2
Коэффициент при 4-й собственной функции языка	−0,1
Коэффициент при 5-й собственной функции языка	0,1

Таблица П30. Резонансные частоты и ширина полос для фрикативного /Ш/, Гц

	$F_р$	F_1	F_2	F_3	F_4	F_5	F_6
Частота	335,1	473,4	1439,9	2101,6	2528,8	3159,8	4516,8
Ширина полосы	72,4	97,5	53,7	57,1	62,8	72,9	117,3

Число Рейнольдса, Re = 3080
Частота первого максимума в спектре шумового источника, $F_{1ш} = 1460$ Гц
Частота второго максимума в спектре шумового источника, $F_{2ш} = 2920$ Гц

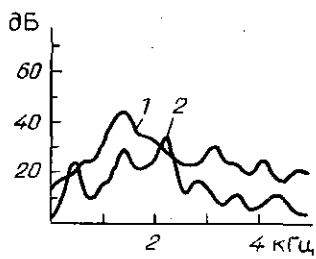


Рис. П29. Амплитудно-частотная характеристика: *1* — шумового источника, *2* — фрикативного /Щ/

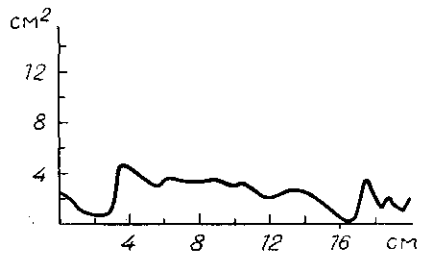


Рис. П30. Площадь поперечного сечения тракта для фрикативного /Ш/

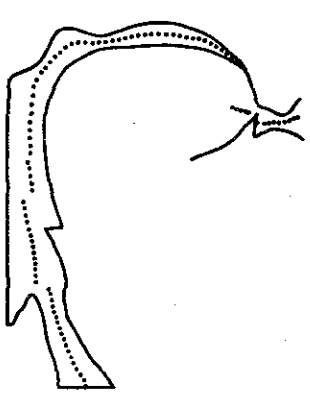


Рис. П31. Форма речевого тракта для фрикативного /Ш/

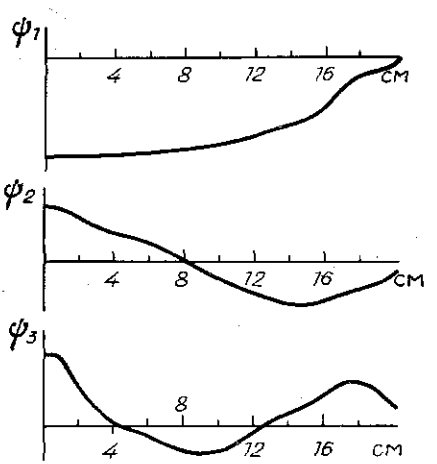


Рис. П32. Собственные функции акустического давления для фрикативного /Ш/

Таблица П31. Площадь поперечного сечения речевого тракта для фрикативного /Х/

Координата вдоль оси тракта, см	Площадь поперечного сечения, $см^2$	Координата вдоль оси тракта, см	Площадь поперечного сечения, $см^2$
0,00	2,75	10,12	1,74
0,44	2,17	10,56	1,87
0,88	1,47	11,00	2,07
1,32	1,03	11,44	1,41
1,76	0,81	11,88	0,69
2,20	0,65	12,32	0,29
2,64	0,72	12,76	1,20
3,08	1,14	13,20	1,97
3,52	4,23	13,64	2,59
3,96	3,98	14,08	3,00
4,40	3,64	14,52	3,29
4,84	3,11	14,96	3,14
5,28	2,54	15,40	2,84
5,72	2,55	15,84	2,42
6,16	3,52	16,28	2,17
6,60	3,46	16,72	1,87
7,04	3,35	17,16	1,41
7,48	3,21	17,60	1,22
7,92	3,10	18,04	3,09
8,36	3,06	18,48	2,99
8,80	3,13	18,92	2,35
9,24	3,13	19,36	2,19
9,68	2,28		

Таблица П32. Артикуляторные параметры для фрикативного /Х/

Ширина голосовой щели	0,4
Горизонтальный сдвиг нижней челюсти	−0,2
Угол поворота нижней челюсти, рад	0,1
Горизонтальная координата корня языка	−8,0
Вертикальная координата корня языка	0,25
Горизонтальная координата кончика языка	−1,6
Вертикальная координата кончика языка	4,6
Угол поворота небной занавески, рад	0,0
Длина губ	4,0
Вертикальная координата нижней губы	0,2
Вертикальная координата голосовой щели	0,0
Коэффициент при 1-й собственной функции языка	−1,2
Коэффициент при 2-й собственной функции языка	−0,275
Коэффициент при 3-й собственной функции языка	0,4125
Коэффициент при 4-й собственной функции языка	−0,1
Коэффициент при 5-й собственной функции языка	1,35

Таблица П33. Резонансные частоты и ширина полос для фрикативного /Х/, Гц

	F_0	F_1	F_2	F_3	F_4	F_5	F_6
Частота	349,9	543,8	1459,7	2035,0	2915,1	3699,1	4540,6
Ширина полосы	72,4	91,9	54,8	53,5	78,5	93,5	120,5

Число Рейнольдса, Re=3100
Частота первого максимума в спектре шумового источника $F_{1ш}$=985 Гц
Частота второго максимума в спектре шумового источника $F_{2ш}$=1970 Гц

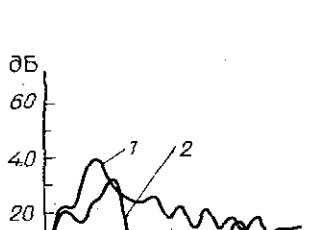


Рис. П33. Амплитудно-частотная характеристика: *1* шумового источника, *2* — фрикативного /Х/

Рис. П34. Площадь поперечного сечения тракта для фрикативного /Х/

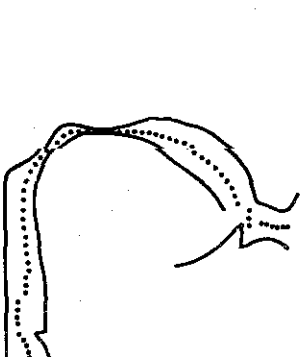


Рис. П35. Форма речевого тракта для фрикативного /Х/

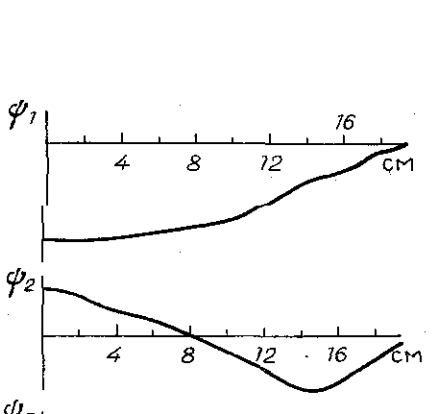


Рис. П36. Собственные функции акустического давления для фрикативного /Х/

Таблица П34. Площадь поперечного сечения речевого тракта для фрикативного /Ф/

Координата вдоль оси тракта, см	Площадь поперечного сечения, $см^2$	Координата вдоль оси тракта, см	Площадь поперечного сечения, $см^2$
0,00	2,75	10,12	6,11
0,44	2,17	10,56	5,49
0,88	1,47	11,00	5,65
1,32	1,03	11,44	6,17
1,76	0,81	11,88	5,54
2,20	0,65	12,32	4,84
2,64	0,72	12,76	4,73
3,08	1,17	13,20	4,98
3,52	5,13	13,64	5,77
3,96	4,95	14,08	6,36
4,40	4,38	14,52	6,90
4,84	3,78	14,96	6,77
5,28	3,46	15,40	6,31
5,72	3,31	15,84	5,53
6,16	4,20	16,28	4,59
6,60	4,03	16,72	3,39
7,04	3,95	17,16	2,90
7,48	3,91	17,60	2,56
7,92	3,95	18,04	1,16
8,36	4,06	18,48	0,30
8,80	4,27	18,92	0,38
9,24	4,58	19,36	1,68
9,68	5,55		

Таблица П35. Артикуляторные параметры для фрикативного /Ф/

Ширина голосовой щели	0,4
Горизонтальный сдвиг нижней челюсти	0,0
Угол поворота нижней челюсти, рад	0,085
Горизонтальная координата корня языка	−7,5
Вертикальная координата корня языка	−1,2
Горизонтальная координата кончика языка	−1,0
Вертикальная координата кончика языка	4,3
Угол поворота небной занавески, рад	0,0
Длина губ	4,0
Вертикальная координата нижней губы	−0,55
Вертикальная координата голосовой щели	2,0
Коэффициент при 1-й собственной функции языка	0,3375
Коэффициент при 2-й собственной функции языка	0,0875
Коэффициент при 3-й собственной функции языка	0,225
Коэффициент при 4-й собственной функции языка	−0,05
Коэффициент при 5-й собственной функции языка	−1,2

Таблица П36. Резонансные частоты и ширина полос для фрикативного /Ф/, Гц

	F_p	F_1	F_2	F_3	F_4	F_5	F_6
Частота	274,9	338,9	1204,6	2110,2	2694,5	3872,9	4798,0
Ширина полосы	72,4	83,2	37,4	43,2	53,5	78,0	104,9

Число Рейнольдса, Re = 3000
Частота первого максимума в спектре шумового источника $F_{1m} = 1150$ Гц
Частота второго максимума в спектре шумового источника $F_{2m} = 2300$ Гц

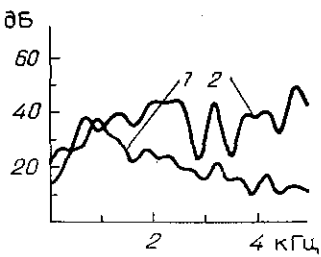


Рис. П37. Амплитудно-частотная характеристика: *1* — шумового источника, *2* — фрикативного /Ф/

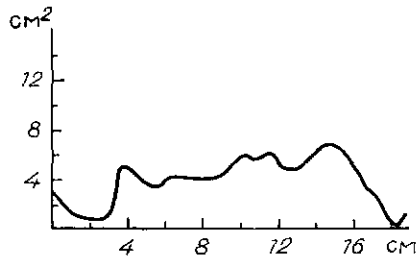


Рис. П38. Площадь поперечного сечения тракта для фрикативного /Ф/

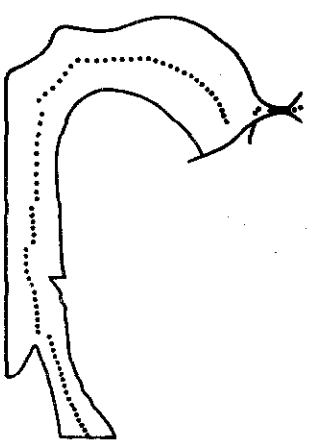


Рис. П39. Форма речевого тракта для фрикативного /Ф/

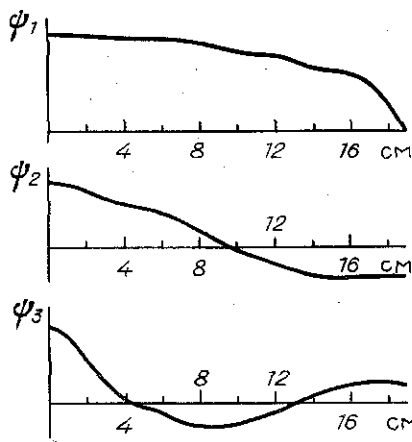


Рис. П40. Собственные функции акустического давления для фрикативного /Ф/

Таблица П37. Параметры мягких фрикативных

Фрикативный	С'	Щ	Ф'	Х'
Ширина голосовой щели	0,4	0,4	0,4	0,4
Горизонтальный сдвиг нижней челюсти	0,3	0,0	0,0	−0,2
Угол поворота нижней челюсти	0,07	0,056	0,085	0,1
Горизонтальная координата корня языка	−8,0	−8,0	−6,2	−6,0
Вертикальная координата корня языка	0,5	0,6	1,0	0,24
Горизонтальная координата кончика языка	−0,6	−1,2	−0,4	−1,6
Вертикальная координата кончика языка	5,3	5,6	4,3	4,6
Угол поворота небной занавески, рад	0,0	0,0	0,0	0,0
Длина губ	4,0	4,0	4,0	4,0
Вертикальная координата нижней губы	0,2	0,2	−0,5	0,2
Вертикальная координата голосовой щели	2,0	2,0	1,0	1,0
Коэффициент при 1-й собственной функции языка	0,24	0,0	−0,2125	−1,29
Коэффициент при 2-й собственной функции языка	−0,1875	0,3	−0,35	−0,32
Коэффициент при 3-й собственной функции языка	0,125	−0,11	0,15	0,4135
Коэффициент при 4-й собственной функции языка	−0,182	−0,25	−0,05	0,1
Коэффициент при 5-й собственной функции языка	0,1	0,1	−0,65	0,8
Число Рейнольдса	3060	3145	3170	3150
Частота первого максимума в спектре источника	1840	1590	1095	1050
Частота второго максимума в спектре источника	3680	3180	2190	2100

Таблица П38. Артикуляторные параметры для согласного /Б/ в коартикуляции с гласными

Гласный	А	О	У	И	Э	Ы
Ширина голосовой щели	0,0	0,0	0,0	0,0	0,0	0,0
Горизонтальный сдвиг нижней челюсти	0,0	0,0	0,0	0,0	0,0	0,0
Угол поворота нижней челюсти, рад	0,12	0,07	0,055	0,07	0,1	0,07
Горизонтальная координата корня языка	−7,12	−8,0	−7,4	−6,2	−7,5	−7,7
Вертикальная координата корня языка	−1,0	−1,0	−0,45	1,0	−1,2	0,0
Горизонтальная координата кончика языка	−1,4	−2,2	−3,0	−0,4	−1,0	−3,0
Вертикальная координата кончика языка	4,0	3,4	4,0	4,3	4,3	4,0
Угол поворота небной занавески, рад	0,0	0,0	0,0	0,0	0,0	0,0
Длина губ	4,0	4,0	3,6	4,0	4,0	4,0
Вертикальная координата нижней губы	−1,2	−0,9	−0,8	−1,0	−1,0	−1,0
Вертикальная координата голосовой щели	2,0	1,0	−0,5	2,0	2,0	−1,0
Коэффициент при 1-й собственной функции языка	−0,425	−0,85	−1,6	−0,2125	0,3375	−0,2625
Коэффициент при 2-й собственной функции языка	0,92	0,9	0,3	−0,35	0,0875	−0,3375
Коэффициент при 3-й собственной функции языка	0,05	0,3	0,6	0,15	0,225	0,1875
Коэффициент при 4-й собственной функции языка	−0,1	−0,0875	−0,12	−0,05	−0,05	−0,05
Коэффициент при 5-й собственной функции языка	−0,3	−0,4	−0,025	−0,65	−0,3	−0,44

Таблица П39. Артикуляторные параметры для согласного /Д/ в коартикуляции с гласными

Гласный	А	О	У	И	Э	Ы
Ширина голосовой щели	0,0	0,0	0,0	0,0	0,0	0,0
Горизонтальный сдвиг нижней челюсти	0,0	0,0	0,0	0,0	0,0	0,0
Угол поворота нижней челюсти, рад	0,04	0,05	0,04	0,07	0,05	0,06
Горизонтальная координата корня языка	−8,4	−7,5	−8,4	−6,5	−7,5	−7,7
Вертикальная координата корня языка	0,4	0,5	0,4	0,8	−1,0	0,5
Горизонтальная координата кончика языка	−1,4	−1,4	−1,4	−0,4	−1,4	−1,4
Вертикальная координата кончика языка	5,0	5,0	5,0	4,3	5,0	5,5
Угол поворота небной занавески, рад	0,0	0,0	0,0	0,0	0,0	0,0
Длина губ	4,0	4,0	3,6	4,0	4,0	4,0
Вертикальная координата нижней губы	0,4	0,4	0,4	0,2	0,4	0,4
Вертикальная координата голосовой щели	2,0	1,0	−0,5	2,0	2,0	−1,0
Коэффициент при 1-й собственной функции языка	−0,4	−0,4	−0,4	−0,2125	0,45	−0,3
Коэффициент при 2-й собственной функции языка	0,3	0,9	0,3	−0,35	0,0875	0,4
Коэффициент при 3-й собственной функции языка	0,05	0,3	0,05	0,15	0,225	0,1875
Коэффициент при 4-й собственной функции языка	−0,4	−0,2	−0,4	−0,25	−0,25	−0,05
Коэффициент при 5-й собственной функции языка	0,8	0,2	0,8	−0,05	0,3	0,4

Таблица П40. Артикуляторные параметры для согласного /Г/ в коартикуляции с гласными

Гласный	А	О	У	И	Э	Ы
Ширина голосовой щели	0,0	0,0	0,0	0,0	0,0	0,0
Горизонтальный сдвиг нижней челюсти	−0,2	−0,2	−0,2	−0,2	−0,2	−0,2
Угол поворота нижней челюсти, рад	0,08	0,07	0,055	0,07	0,08	0,06
Горизонтальная координата корня языка	−8,0	−8,0	−8,0	−8,0	−8,0	−8,0
Вертикальная координата корня языка	0,24	0,24	0,24	0,24	0,24	0,24
Горизонтальная координата кончика языка	−1,6	−1,6	−1,6	−1,6	−1,6	−1,6
Вертикальная координата кончика языка						
Угол поворота небной занавески, рад	0,0	0,0	0,0	0,0	0,0	0,0
Длина губ	4,0	4,0	3,6	4,0	4,0	4,0
Вертикальная координата нижней губы	0,2	0,2	0,0	0,2	0,2	0,4
Вертикальная координата голосовой щели	2,0	1,0	−0,5	2,0	2,0	−1,0
Коэффициент при 1-й собственной функции языка	−1,3	−1,3	−1,3	−1,3	−1,3	−1,3
Коэффициент при 2-й собственной функции языка	−0,28	−0,28	−0,28	−0,28	−0,28	−0,28
Коэффициент при 3-й собственной функции языка	0,4135	0,4135	0,4135	0,4135	0,4135	0,4135
Коэффициент при 4-й собственной функции языка	−0,1	−0,1	−0,1	−0,1	−0,1	−0,1
Коэффициент при 5-й собственной функции языка	1,35	1,35	1,35	1,35	1,35	1,35

Таблица П41. Артикуляторные параметры для назального (М) в коартикуляции с гласными

Гласный	А	О	У	И	Э	Ы
Ширина голосовой щели	0,0	0,0	0,0	0,0	0,0	0,0
Горизонтальный сдвиг нижней челюсти	0,0	0,0	0,0	0,0	0,0	0,0
Угол поворота нижней челюсти, рад	0,12	0,07	0,055	0,07	0,1	0,07
Горизонтальная координата корня языка	−7,2	−8,0	−7,0	−6,2	−7,5	−7,7
Вертикальная координата корня языка	−1,0	−1,0	−1,45	−1,0	−1,2	0,0
Горизонтальная координата кончика языка	−1,4	−2,2	−3,0	−0,4	−1,0	−3,0
Вертикальная координата кончика языка	4,0	3,4	4,0	4,3	4,3	4,0
Угол поворота небной занавески, рад	0,07	0,07	0,07	0,7	0,1	0,08
Длина губ	4,0	4,0	3,6	4,0	4,0	4,0
Вертикальная координата нижней губы	−1,2	−0,9	−0,8	−1,0	−1,0	−1,0
Вертикальная координата голосовой щели	2,0	1,0	−0,5	2,0	2,0	−1,0
Коэффициент при 1-й собственной функции языка	−0,425	−0,85	−1,3	−0,2125	0,3375	−0,2625
Коэффициент при 2-й собственной функции языка	0,92	0,9	0,3	−0,35	0,0875	−0,3375
Коэффициент при 3-й собственной функции языка	0,05	0,3	0,6	0,15	0,225	0,1875
Коэффициент при 4-й собственной функции языка	−0,1	−0,0875	−0,12	−0,05	−0,05	−0,05
Коэффициент при 5-й собственной функции языка	−0,3	−0,4	−0,025	−0,65	−0,3	−0,44

Таблица П42. Артикуляторные параметры для назального /Н/ в коартикуляции с гласными

Гласный	А	О	У	И	Э	Ы
Ширина голосовой щели	0,0	0,0	0,0	0,0	0,0	0,0
Горизонтальный сдвиг нижней челюсти	0,0	0,0	0,0	0,0	0,0	0,0
Угол поворота нижней челюсти, рад	0,04	0,05	0,04	0,07	0,05	0,06
Горизонтальная координата корня языка	−8,4	−7,5	−1,4	−6,5	−7,5	−7,7
Вертикальная координата корня языка	0,4	0,5	−0,4	0,8	−1,0	0,5
Горизонтальная координата кончика языка	−1,4	−1,4	−1,4	−0,4	−1,4	−1,4
Вертикальная координата кончика языка	5,0	5,0	5,0	4,3	5,0	5,5
Угол поворота небной занавески, рад	0,07	0,07	0,07	0,07	0,07	0,07
Длина губ	4,0	4,0	3,6	4,0	4,0	4,0
Вертикальная координата нижней губы	0,4	0,4	0,4	0,2	0,4	0,4
Вертикальная координата голосовой щели	2,0	1,0	−0,5	2,0	2,0	−1,0
Коэффициент при 1-й собственной функции языка	−0,4	−0,4	−0,4	−0,2125	0,45	−0,3
Коэффициент при 2-й собственной функции языка	0,3	0,9	0,3	−0,35	0,0875	0,4
Коэффициент при 3-й собственной функции языка	0,05	0,3	0,05	0,15	0,225	0,1875
Коэффициент при 4-й собственной функции языка	−0,4	−0,2	−0,4	−0,25	−0,25	−0,05
Коэффициент при 5-й собственной функции языка	0,8	0,2	0,8	−0,05	0,3	0,4

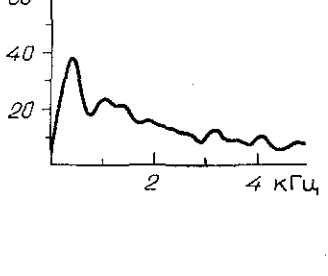


Рис. П41. Амплитудно-частотная характеристика источника шума в голосовой щели. Число Рейнольдса Re=3000, частота первого максимума $F_1=400$ Гц, частота второго максимума $F_2=800$ Гц

В табл. П37 приведены координаты артикуляторных органов, числа Рейнольдса и частоты спектрального максимума источника шумового возбуждения для мягких фрикативных согласных /С', Щ, Ф', Х'/. На рис. П41 показан спектр шума голосовой щели при ее максимальном раскрытии в процессе фонации.

Коартикуляция гласных и согласных звуков требует изменения координат артикуляторов таким образом, чтобы не возникали ложные места артикуляции. В табл. П38—П42 приведены параметры для взрывных /Б, Д, Г/ и назальных /М, Н/, артикулируемых на фоне всех гласных звуков. Отличие от глухих взрывных /П, Т, К/ заключается только в расстоянии между задними концами голосовых складок, а остальные артикуляторные параметры — те же, что и для звонких взрывных.

СПИСОК ЛИТЕРАТУРЫ

1. Березин И. С., Жидков Н. П. Методы вычислений.— М.: Физматгиз, 1962. 446 с.
2. Блохинцев Д. И. Акустика неоднородной движущейся среды.— М.: Наука, 1981.— 206 с.
3. Бондарко Л. В. Звуковой строй современного русского языка.- М.: Просвещение, 1977.— 174 с.
4. Боголюбов Н. Н., Митропольский Ю. А. Асимптотические методы в теории нелинейных колебаний.— М.: Физматгиз, 1958.— 408 с.
5. Брызгунова Е. А. Звуки и интонация русской речи.— М.: Русская литература, 1977.— 279 с.
6. Вемян Г. В. Передача речи по сетям электросвязи.— М.: Радио и связь, 1985.— 272 с.
7. Венедиктов М. Д., Женевский Ю. П., Марков В. В., Эйдус Г. С. Дельта-модуляция.— М.: Связь, 1976.— 271 с.
8. Вербицкая Л. А., Алексеева И. Н. Использование машинных методов обработки сигнала в исследовании произносительной нормы русского литературного языка // Труды АРСО-15.— Таллинн: Изд-во института кибернетики АН ЭССР, 1984, С. 267.
9. Винарская Е. Н. Психологический аспект тестирования // Речевые тесты и их применение.— М.: Изд-во МГУ, 1986.— С. 20—26.
10. Герц С. Р., Кейдин Д., Карплус К. Д. Система Дельта для разработки правил синтеза речи по тексту // ТИИЭР.— 1985.— Т. 73, № 11.— С. 62—75.
11. Годунов С. К., Рябенький В. С. Разностные схемы.— М.: Наука, 1973.— 400 с.
12. Голдстейн М. Э. Аэроакустика.— М.: Машиностроение, 1981.— 294 с.
13. Гурфинкель В. С., Арутюнян Г. А., Мирский М. П. Организация движений при выполнении человеком точностной позной задачи // Биофизика.— 1969.— Т. 14, № 6.— С. 1103—1107.
14. Гурфинкель В. С., Левик Ю. С. Сенсорные комплексы и сенсомоторная интеграция // Физиология человека.— 1979.— № 3.— С. 399—414.
15. Демидович Б. П., Марон И. А. Основы вычислительной математики.— М.: Наука, 1970.— 664 с.
16. Деркач М. Ф., Гумецкий Р. Я., Гура Б. М., Чабан М. Е. Динамические спектры речевых сигналов.— Львов.: Вища школа, 1983.— 167 с.
17. Елкина В. Н., Юдина Л. С. Алгоритм автоматического транскрибирования текста // Вычислительные системы.— Новосибирск: Наука, 1973.— Вып. 55. С. 127—133.
18. Зиндер Л. Р., Штерн А. С. Лингвистический аспект тестирования // Речевые тесты и их применение.— М.: Изд-во МГУ, 1986.— С. 6—20.
19. Златоустова Л. В. Фонетическая структура слова в потоке речи.— Казань: Изд-во Казанского университета, 1962.
20. Златоустова Л. В., Хитина М. В., Колесников Б. М. и др. Принципы составления и структура речевых тестов // Речевые тесты и их применение.— М.: Изд-во МГУ, 1986.— С. 27—36.
21. Златоустова Л. В., Кодзасов С. В., Кривнова О. Ф., Фролова И. Г. Алгоритмы преобразования русских орфографических текстов в фонетическую запись.— М.: Изд-во МГУ, 1970.— 130 с.

22. Канторович Л. В., Крылов В. И. Приближенные методы высшего анализа.—М.—Л.: Физматгиз, 1962.—708 с.
23. Кейтер Д. Компьютеры — синтезаторы речи.— М.: Мир, 1985.—237 с.
24. Книппер А. В., Махонин В. А. К описанию речевых сигналов // Речевое общение в автоматизированных системах.—М.: Наука, 1975.—С. 46—59.
25. Книппер А. В., Краевский В. И., Савельев В. П., Сорокин В. Н. Чудновский Л. С. Артикуляторно-ориентированный первичный анализ речи // Труды АРСО-14.—Каунас, 1986.—С. 61—62.
26. Книппер А. В. Индивидуальные вариации длительности элементов речи // Речевая информатика.—М.: Наука, 1989.—С. 34—48.
27. Кодзасов С. В., Кривнова О. Ф. Фонетические возможности гортани и их использование в русской речи // Проблемы теоретической и экспериментальной лингвистики.—М.: Изд-во МГУ, 1977.—С. 180—209.
28. Коренев Г. В. Цель и приспособляемое движение.—М.: Наука, 1974.—528 с.
29. Кучеров В. Я., Лобанов Б. М. *Синтезированная речь в системах* массового обслуживания.— М.: Радио и связь, 1983.—130 с.
30. Кюннап Э. Взаимное влияние фонем эстонского языка на параметры друг друга // Труды АРСО-15.—Таллинн: Изд-во института кибернетики АН ЭССР, 1989.—С. 164—167.
31. Левин Л. С., Плоткин М. А. Цифровые системы *передачи информации*.—М.: Радио и связь, 1982.—216 с.
32. Ложкин В. Н., Савинский В. Г. К *вопросу о различительных* признаках взрывных согласных // Речевая информатика.—М.: Наука, 1989.—С. 55—62.
33. Лойцянский Л. Г. Механика жидкости и газа.—М.: Наука, 1978.—736 с.
34. Макул Д. Векторное квантование при кодировании речи // ТИИЭР.—1985. Т. 73, № 11.—С. 19—61.
35. Маркелл Д. Д., Грей А. Х. *Линейное предсказание речи*.—М.: Связь, 1980.—308 с.
36. Михайлов В. Г., Златоустова Л. В. Измерение параметров речи.—М.: Радио и связь, 1987.—168 с.
37. Морз Ф. Колебания и звук.—М.—Л.: ГИТТЛ, 1949.—496 с.
38. Назаров М. В., Прохоров Ю. Н. Методы цифровой обработки и передачи речевых сигналов.—М.: Радио и связь, 1985.—176 с.
39. Николаева Т. М. Фразовая интонация славянских языков.—М.: Наука, 1977.—278 с.
40. Оппенгейм А. В., Шафер Р. В. Цифровая обработка сигналов.—М.: Связь, 1979.—416 с.
41. Пизони Д. Б., Нусбаум Г. С., Грин Б. Г. Восприятие синтезированной речи, генерируемой по правилам // ТИИЭР.—Т. 73, № 11.—С. 146—160.
42. Покровский Н. Б. Расчет и измерение разборчивости речи.—М.: Связьиздат, 1976.—391 с.
43. Понтрягин Л. С., Болтянский В. Г., Гамкрелидзе Р. В., Мищенко Е. Ф. Математическая *теория оптимальных* процессов.—М.: Наука, 1961.—391 с.
44. Потапова Р. К. Слоговая фонетика германских языков.—М.: Высшая школа, 1986.—144 с.
45. *Потапова Р. К. Подготовка и использование тестового материала* в целях идентификации языковой принадлежности говорящего // Речевые тесты и их применение.—М.: Изд-во МГУ, 1986.—С. 40—52.
46. Прохоров Ю. Н. Статистические модели и рекуррентное предсказание речевых сигналов.—М.: *Радио и связь*, 1984.—239 с.
47. Рабинер Л., Голд Б. Теория и применение цифровой обработки сигналов.—М.: Мир, 1978.—848 с.
48. *Рабинер Л. Р., Шафер Р. В. Цифровая обработка речевых сигналов*.—М.: Радио и связь, 1981.—495 с.
49. Речь. Артикуляция и восприятие.—Л.: Наука, 1965.—241 с.
50. Ржевкин С. Н. Курс лекций по теории звука.—М.: Изд-во МГУ, 1960.—335 с.

51. Рихтмайер Р., Мортон К. Разностные методы решения краевых задач.— М.: Мир, 1972.—418 с.
52. Розанова Н. Н. Суперсегментная фонетика // Русская разговорная речь.— М.: Наука, 1983.— С. 5—79.
53. Романов В. Г. Некоторые обратные задачи для уравнений гиперболического типа.— Новосибирск: Наука, 1972.—172 с.
54. Русская разговорная речь.— М.: Наука, 1973.—485 с.
55. Сапожков М. А., Михайлов В. Г. Вокодерная связь.— М.: Радио и связь, 1983.—248 с.
56. Светозарова Н. Д. Интонационная система русского языка.— Л.: Изд-во ЛГУ, 1982.—175 с.
57. Сорокин В. Н. Быстрые процедуры анализа речи // Труды АРСО-13.— Новосибирск: Изд-во Новосибирского университета, 1984.— С. 81—82.
58. Сорокин В. Н. О роли подглоточной области в процессе речеобразования // Проблемы построения систем понимания речи.— М.: Наука, 1980.— С. 125—135.
59. Сорокин В. Н. Теория речеобразования.— М.: Радио и связь, 1985.— 312 с.
60. Сорокин В. Н. Бегущие волны в речевом тракте // Акуст. ж.— 1986.— № 4.— С. 506—510.
61. Сорокин В. Н. Временные параметры элементов русской речи // Речевая информатика.— М.: Наука, 1989.— С. 5—33.
62. Тихонов А. Н., Арсенин В. Я. Методы решения некорректных задач.— М.: Наука, 1974.—223 с.
63. Уайльд Д. Д. Методы поиска экстремума.— М.: Наука, 1967.—268 с.
64. Фант Г. Акустическая теория речеобразования.— М.: Наука, 1964.—284 с.
65. Фланаган Д. Анализ, синтез и восприятие речи.— М.: Связь, 1968.— 392 с.
66. Физиология речи. Восприятие речи человеком / Чистович Л. А. и др.— Л.: Наука, 1976.—386 с.
67. Adams J. A. Issues for a closed-loop theory of motor learning // Motor control.— London: Academic Press, 1976.— P. 87—107.
68. Allen D. R., Strong W. J. A model for the synthesis of natural sounding vowels // JASA.— 1985.— Part 1.— V. 78, № 1.— P. 58—69.
69. Allen G. D. Segmental timing in speech production // J. Phonetics.— 1973.— V. 1, № 3.— P. 219—237.
70. Ananthapadmanabha T. V. Acoustic analysis of voice source dynamics // STL QPSR.— 1984.— № 2, 3.— P. 1—24.
71. Atal B. S., Remde J. R. A new model of LPC excitation for producing natural—sounding speech at low bit rates // ICASSP-82.— 1982.— P. 614—617.
72. Auloge J. Y., Guerin B., Descout R. Modelisation élastique des cordes vocales -aspects modaux // 12 Journe d'etude sur la parole, Montreal.— 1981.— P. 13—25.
73. Avesani C. Declination and sentence intonation in Italian // XI ICPhS, Tallinn.— 1987.— V. 3.— P. 153—156.
74. Berg van Den J., Zantena J. T., Doornenball J. On the air resistance and the Bernoulli effect in the human larynx // JASA.— 1957.— V. 27, № 5.— P. 626—631.
75. Berkovits R. Duration and fundamental frequency in sentence final intonation // J. Phonetics.— 1984.— № 12.— P. 255—265.
76. Binh N., Gauffin J. Aerodynamic measurements in an enlarged static laryngeal model // STL QPSR.— 1983.— № 2, 3.— P. 36—60.
77. Bjork L. Velopharyngeal function in connected speech // Acta Radologica.— 1961.— Suppl. 1.
78. Bladon A., Carlson R., Granström B., Hunnicat S., Karlsson I. A text—to—speech system for British English and issues of dialect and style // STL QPSR.— 1987.— № 2, 3.— P. 1—5.
79. Blumstein S. E., Stevens K. N. Perceptial invariance and onset spectra for stop consonants in different vowel enviroment // JASA.— 1980.— V. 67, № 2.— P. 648—662.

80. Bo S., Jialy Z. Vowel intrinsic pitch in standart Chinese // XI ICPhS, *Tallinn.— 1987.— V. 1.— P. 142—145.*
81. Bocchieri E. L. An articulatory speech synthesizer: Ph.D. Thesis, Univ. of Florida, 1983.
82. Bowman J. R. The muscle spindle and neural control of the tongue.— Springfield, 1971.
83. Bruce G. Experiments with Swedish intonation model.— Preprint / Working group on intonation.— Tokyo, 1982.— P. 35— 46.
84. Carlsson R., Granström B., Klatt D. H. Some notes on the perception of temporal pattern in speech // *Frontiers of speech communication research.*— New York: Academic Press, 1979.— *P. 233 243.*
85. Carlsson R., Granström B. Swedish durational rules derived from a sentence data base // *STL QPSR.— 1986.— № 2, 3.— P. 13—25.*
86. Childers D. G., Naik J. M., Larar J. N. et al. Electroglottography, speech, and ultrahigh speed cinematography // *Vocal fold Physiology.*— 1983.— P. 202— 220.
87. Childers D. G., Paige A., Moore A. Laryngeal vibration patterns // Archives of Otolaryngology.— 1976.— V. 102.— P. 407—410.
88. Cooper W. E. *Syntactic control of speech timing:* Ph. D. Thesis, MIT, 1975.
89. Cooper W. E. Danly M. Segmental and temporal aspects of utterance—final *lengthening // Phonetica.— 1981.— V. 38, № 1—3.— P. 106—115.*
90. Dedina M. J., Nusbaum H. S. PRONOUNCE: A program for pronunciation by analogy // Research on Speech Perception, Indiana Univ.— 1986.— № 12.— P. 335—348.
91. Delattre P., Liberman A. M., Cooper F. S. Acoustic loci and transitional cues for consonants // JASA.— 1955.— V. 27.— P. 769—774.
92. Delattre P. Des dix intonations de base // The French Review.— 1966. — V. 40, № 1. — P. 1 —14.
93. Dunn H. K. The calculation of vowel resonances, and an electrical vocal tract // JASA.— 1950.— V. 22.— P. 740—753.
94. Delattre P. A. A comparison of syllable length conditioning among languages // *International review of applied linguistics.— 1966.— № 4.— P. 183—198.*
95. Fant G. Glottal source and excitation analysis // STL QPSR. — 1979.— № 1.— P. 85 — 107.
96. Ferrero E. E., Palamati G. M., Vogges K. Perceptional category shift of *voiceless Italian fricatives as a function* of duration shortening // Frontiers of speech communication research.— New York: Academic Press, 1979.— P. 159—165.
97. Flanagan J. L., Landgraf L. L. Self—oscillating source for vocal tract *synthesizer* // IEEE Trans. on Audio and Electroacoustics. — 1968.— V. AU— 16.— P. 57 —64.
98. Flanagan J., Coker C., Rabiner I. et al. Synthetic voices for computers // IEEE Spectrum.— 1970.— V. 7, № 10.— P. 22—45.
99. Folkins J. W., Abbs J. H. Lip and jaw motor control during speech: Responses to resistive loading of the jaw // JSHR.— 1975.— № 18.— P. 207— 220.
100. Fonagy I. Semantic diversity in intonation // XI ICPhS, Tallinn.— 1987. — V. 2.— P. 468— 471.
101. Fowler C. A. A relationship between coarticulation and compensatory shortening // Phonetica.— 1981.— V. 38, № 1—3.— P. 35—50.
102. Fromkin V. A. The non—anomalous nature of anomalous utterances // Language.— 1971.— № 47. — P. 27 —52.
103. Fujimura O. Stereo fiberscope // Dynamic aspects of speech production.— Tokyo: Univ. of Tokyo Press, 1976.— P. 133—138.
104. Fujimura O., Lovins J. Syllables as concatenative phonemic elements // Syllables and segments. — New York: North Holland, 1978.— P. 107—120.
105. Fujisaki H., Hirose K. Modeling the dynamic characteristics of voice *fundamental frequency* with application to analysis and synthesis of intonation.— Preprint / Working group on intonation.— Tokyo, 1982.— P. 57—70.
106. Fujisaki H., Lehiste I. Some temporal and tonal characteristics of declarative sentenses in Estonian.— Preprint / Working group on intonation.— Tokyo, 1982.— P. 121 —130.

107. Fujisaki H., Kawai H. Realization of linguistic information in the voice fundamental frequency contour of the spoken language // ICASSP-88.— 1988.— V. 1.— P. 663—666.
108. Gay T. Mechanisms in the control of speech rate // Phonetica.— 1981.— V. 38, № 1.— P. 148—158.
109. Gay T., Lindblom B., Lubker J. Production of bite—block vowels: Acoustic equivalence by selective compensation // JASA.— 1981.— V. 69, № 3.— P. 802—810.
110. Greene B. G., Pisoni D. B. Perception of synthetic speech by nonnative speakers of English // Research on speech perception, Indiana Univ.— 1985.— № 11.— P. 419—428.
111. Greenspan S. L., Nusbaum H. C., Pisoni D. B. Perceptial learning of synthetic speech produced by rule // Research on speech perception, Indiana Univ.— 1986.— № 12.— P. 43—86.
112. Greene B. G., Manou L. M., Pisoni D. B. Perceptial evaluation of DEC-Talk: a final report on version 1.8 // Research on speech perception, Indiana Univ.— 1984.— № 10.— P. 77—127.
113. Guerin B., Boe L. J. Etude de l'influence du couplage acoustic source—conduit vocal sur F_0 des voyelles orales // Phonetica.— 1980.— V. 37.— P. 169—192.
114. Harshman R., Ladefoged P., Goldstein L. Factor analysis of tongue shapes // JASA.— 1977.— V. 62, № 3.— P. 693—707.
115. Hiki S. Effect of the context on the duration of phonetic segment // Report of Research Institute of Electr. Communication.— Tohoku, 1968.— P. 14—17.
116. Hiki S., Imaizumi S. Observation of symmetry of tongue movement by use of dynamic palatography // Ann. Bull. RILP — Tokyo Univ.— 1974.— № 8.— P. 69—74.
117. Holmes J. N. Speech synthesis // Frontiers of speech communication research.— New York: Academic Press, 1979.— P. 275—285.
118. Holmes J. N. Formant synthesizer: cascade or parallel? // Speech Communication.— 1983.— № 2.— P. 251—273.
119. House J., Johnson M. Enlivening the intonation in text-to-speech synthesis: an «accent unit» model // XI ICPhS, Tallinn.— 1987.— V. 3.— P. 134—137.
120. Hunnicat S. Grapheme-to-phoneme rules, a review // STL QPSR.— 1980.— № 2, 3.— P. 38—60.
121. Imagava H., Kiritani S., Masaki S., Shirai K. Comparison of velocity and duration between open to close vowel transition and close to open vowel transition // Ann. Bull. RILP, Tokyo Univ.— 1983.— № 17.— P. 33—76.
122. Ishizaka K., Matsudaira M. Fluid mechanical consideration of vocal cords vibration // SCRL Monograph.— 1972.— № 8.
123. Ishizaka K., Flanagan J. L. Synthesis of voiced sounds from a two-mass model of the vocal cords // Bell. Syst. Techn. J.— 1972.— № 51.— P. 1233—1268.
124. Ishizaka K., French J. C., Flanagan J. L. Direct determination of vocal tract wall impedance // IEEE on ASSP.— 1975.— V. 23, № 4.— P. 370—373.
125. Ishizaka K., Matsudaira M., Kaneko T. Input acoustic impedance measurement of the subglottal system // JASA.— 1976.— V. 60.— P. 190—197.
126. Jospa P. Conséquences acoustiques des déformation dynamiques du conduit vocale // Articulatory modeling and phonetics.— Grenoble: 1977.— P. 49—64.
127. Kelly J. L., Lochbaum C. C. Speech synthesis // Proc. 4 Intern. Congr. Acoust., 1962, Paper G42.— P. 1—4.
128. Kelso S., Stelmach G. E. Central and peripheral mechanisms in motor control // Motor Control.— London: Academic Press, 1976.— P. 33—40.
129. Kiritani S., Mijanaki K., Fujimura O. A computantional model of the tongue // Ann. Bull., RILP, Tokyo Univ.— 1976.— № 10.— P. 243—252.
130. Kiritani S., Takenuka E., Sawashima M. Computer tomography of the vocal tract // Ann. Bull., RILP, Tokyo Univ.— 1978.— № 12.— P. 1—4.
131. Klatt D. H. Discrimination of fundamental frequency contours in synthetic speech: implications for models of pitch perception // JASA.— 1973.— V. 53.— P. 8—16.

132. Klatt D. H. Vowel lengthening is syntactically determined in a connected discourse // J. Phonetics.— 1975.— № 3.— P. 129—140.
133. Klatt D. H. Linguistic uses of segmental durations in English: acoustic and perceptual evidence // JASA.— 1976.— № 5.— P. 1208—1221.
134. Klatt D. H. Synthesis by rule of segmental durations in English sentences // Frontiers of speech communication research. — N. Y.: Academic Press, 1979.— P. 287—299.
135. Klatt D. H. Software for a cascade/parallel formant synthesizer // JASA.— 1979.— V. 67, № 3.— P. 971—995.
136. Klatt D. H. A strategy for the perceptial interpretation of durational cues in English sentences // Working Papers, MIT, Speech Communication Group.— 1982.— V. 1.— P. 83—91.
137. Klatt D. H. Review of text—to—speech conversion for English // JASA.— 1987.— V. 82, № 3.— P. 737—793.
138. Koizumi T., Taniguchi S., Hiromitsu S. Glottal source vocal tract interaction // JASA.— 1985.— V. 78, № 5.— P. 1541—1547.
139. Ladd D. R. A model of intonational phonology for use in speech synthesis by rule // Speech Technology Conf., Edinburgh.— 1987.— V. 2.— P. 21—24.
140. Ladefoged P., Harshman R., Goldstein L., Rice L. Generating vocal tract shapes from formant frequencies // JASA.— 1978.— V. 64.— P. 1027—1035.
141. Laine U. K. Modeling of lip radiation impedance in the Z—domain // ICASSP—82.— 1982.
142. Lienard J. S. et al. Diphone synthesis of French. Vocal response unit and automatic prosody from text // ICASSP—77, Hartford.— 1977.
143. Lin Q. Nonlinear interaction in voice production // STL QPSR.— 1987.— № 1.— P. 1—12.
144. Lindblom B., Lyberg B., Holmgren K. Durational patterns of Swedish phonology: do they reflect short—term motor memory process? // Rep. Stockholm Univ.— 1977.
145. Lilijenkrants J. A Fourier series description of the tongue profiles // STL QPSR, 1974.— № 4.— P. 9—18.
146. Lilijenkrants J. Speech synthesis with a reflection—type line analog: Ph. D. Thesis, Royal Institute of Technology, Stockholm, 1985.
147. Lisker L. et al. Transillumination of the larynx in running speech // JASA.— 1969.— V. 45.— P. 1544—1547.
148. Logan J. S., Pisoni D. B., Greene B. G. Measuring the segmental intelligibility of synthetic speech: results from eight text-to-speech systems // Research on speech perception, Indiana Univ.— 1985.— № 11.— P. 3—31.
149. Luce P. A., Feustel T. C., Pisoni D. B. Capacity demands in short—term memory for synthetic and natural speech // Human Factors.— 1983.— № 25.— P. 17—32.
150. Mackey D. G. Aspects of the syntax behaviour // Quart. J. Experim. Psychology.— 1974.— № 26.— P. 642—657.
151. Maeda S. On a simulation method of dynamically varying vocal tract. Reconsideration of the Kelly—Lochbaum model // Articulatory modeling and phonetics, Grenoble, 1977.— P. 281—288.
152. Maeda S. On the F_0 control mechanisms of the larynx // Seminaire Larynx et Parole, Grenoble.— 1979.— P. 245—258.
153. Maeda S. The role of the sinus cavity in the production of nasal vowels // ICASSP—82.— 1982.
154. Maeda S. A digital simulation method of the vocal—tract system // Speech Communication.— 1982.— № 1.— P. 199—229.
155. Manous L. M., Pisoni D. B., Dedina M. J., Nusbaum H. C. Comprehension of natural and synthetic speech using a sentence verification task // Research on speech perception, Indiana Univ.— 1985.— № 11.— P. 33—57.
156. Martin J. G., Bunnet H. T. Perception of anticipatory coarticulation effects in vowel—stop consonant—vowel sequencies // J. Experim. Psychol.— 1982.— № 8.— P. 473—488.

157. Mitleb F. M. Voicing effect on vowel duration is not an absolute universal // J. Phonetics.— 1984.— № 12.— P. 23—27.
158. Möbius B., Zimmerman A., Hess W. Microprosodic fundamental frequency variation in German // XI ICPhS, Tallinn.— 1987.— V. 1.— P. 146—149.
159. Muller E. M. Brown W. S. Variations in the supraglottal air pressure waveform and their articulatory interpretation // Speech and Language — New York: Academic Press, 1980.— P. 317—389.
160. Nakatani L. H., O'Connor K. D., Aston C. H. Prosodic aspects of American English speech rhythm // Phonetica.— 1981.— V. 38, № 1—3.— P. 84—106.
161. Nishinuma J. Prediction of phoneme duration by a distinctive feature matrix // J. Phonetics.— 1984.— № 12.— P. 169—173.
162. Nord L., Ananthapadmanabha T. V., Fant G. Signal analysis and perceptial tests of vowel responses with an interactive source filter model // STL QPSR;— 1984.— № 2, 3.— P. 25—52.
163. Nusbaum H. C., Pisoni D. B., Davis C. K. Sizing up the Hoosier mental lexicon: measuring the familiarity of 20000 words // Research on speech perception, Indiana Univ.— 1984.— № 10.— P. 357—376.
164. Nusbaum H. C., Pisoni D. B. Constraints on the perception of synthetic speech generated by rule // Research on speech perception, Indiana Univ.— 1984.— № 10.— P. 153—168.
165. Nusbaum H. C., Greenspan S. L., Pisoni D. B. Perceptial attention in monitoring natural and synthetic speech // Research on speech perception, Indiana Univ.— 1986.— № 12.— P. 307—318.
166. Nushikyan E. The typological analysis of emotional speech prosody // XI ICPhS, Tallinn.— 1987.— V. 3.— P. 210—213.
167. Odé C. A perceptional analysis of Russian intonation: some aspects // XI ICPhS, Tallinn.— 1987.— V. 3.— P. 194—197.
168. Ohala J. J. Physiological mechanisms underlying tone and intonation.— Preprint/Working group on intonation.— Tokyo, 1982.— P. 1—12.
169. Olive J. P., Liberman M. J. Text—to—speech—an overview // JASA.— 1985.— Suppl. 1.— V. 78.— P. 6.
170. Öhman S.E.G. Coarticulation in VCV utterances: spectrographic measurements // JASA.— 1966.— V. 39.— P. 151—168.
171. Pardo J. M., Martinez M., Quillis A., Minoz E. Improving text—to—speech conversion in Spanish: liguistic analysis and prosody // Speech Technology Conference, Edinburgh.— 1987.— V. 2.— P. 173—176.
172. Peterson G., Wang W., Sivertsen E. Segmentation techniques in speech synthesis // JASA.— 1958.— V. 30.— P. 739—742.
173. Pettorino M. Intrinsic pitch of vowels: an experimental study in Italian // XI ICPhS, Tallinn.— 1987.— V. 1.
174. Pierrehumbert J. B., Stede Sh. How many rise-fall contours? // XI ICPhS, Tallinn.— 1987.— V. 3.— P. 145—148.
175. Pisoni D. B., Manous L. M., Dedina M. J. Comprehension of natural and synthetic speech: II. Effect of predictability on the verification of sentences controlled for intelligibility // Research on speech perception, Indiana Univ.— 1986.— № 12.— P. 19—42.
176. Pisoni D. B., Dedina M. J. Comprehension of digitally encoded natural speech using a sentence verification task // Research on speech perception, Indiana Univ.— 1986.— № 12.— P. 3—18.
177. Pols L.C.W., Olive J. P. Intelligibility of consonants in CVC utterances produced by diadic rule synthesis // Speech Communication.— 1983.— № 2. P. 3—13.
178. Port R. F., Rotunno R. Relation between voice—onset time and vowel duration // JASA.— 1976.— V. 66, № 3.— P. 654—662.
179. Raphael L. J. Duration and context as cues to word final cognate opposition in English // Phonetica.— 1981.— V. 38, № 1—3.— P. 126—147.
180. Rothenberg M. A new inverse filtering technique for deriving the glottal air flow waveform during voicing // JASA.— 1973.— V. 53.— P. 1632—1645.
181. Rozsypal A. T., Millar B. F. Perception of jitter and shimmer in synthetic vowels // J. Phonetics.— 1979.— № 7.— P. 275—285.

182. Russel D. G. Spatial location cues and movement production // Motor Control.— London: Academic Press, 1976.— P. 67—85.
183. Sappok Ch. Irregular periodicity as a boundary cue between phrases // XI ICPhS, Tallinn.— 1987.— V. 3.— P. 157—160.
184. Sawashima M., Cooper F. S. Fiberoptic ovservation of the larynx and other speech organs // Dynamic aspects of speech production.— Tokyo: Univ. of Tokyo Press.— 1977.— P. 31—46.
185. Schiefer L. F. Perturbations in Hindi // XI ICPhS. Tallinn.— 1987.— V. 1.— P. 150—153.
186. Scully C. *Linguistic units and units of speech production* // Speech Communication.— 1987.— № 7.— P. 77—142.
187. Schmidt R. A. The schema as a solution to some persistent problems in motor learning theory // Motor Control.— London: Academic Press, 1976.— P. 41—66.
188. Schmidt R. A. The schema concept // Human Motor Behaviour.— New York: Academic Press, 1982.— P. 219—238.
189. Shirai K. Computer models for speech production // Auditory Signals.— 1983.— V. 2.— P. 102—141.
190. Shirai K., Honda M. An articulatory model and the estimation of articulatory parameters by nonlinear regression method // Electronics and Communications in Japan.— 1976.— V. 59 A, № 8.— P. 35—43.
191. Sorokin V. N. Coordination of muscles and articulators // XI ICPhS, Tallinn.— 1987.— V. 3.— P. 382—384.
192. Stevens K. N., Kasowski S., Fant G. An electrical analog of the vocal tract // JASA.— 1953.— V. 25.— P. 734—742.
193. Stevens K. N., House A. Development of a quantitative description of vowel articulation // JASA.— 1955.— V. 27.— P. 484—493.
194. Strube H. W. Time-varying wave digital filters for modeling analog systems // IEEE Trans.— 1982.— V. ASSP-30, № 6.— P. 864—868.
195. Tarnoszy T., Vicsi K. Decay characteristics of the vowel cavities and radiation properties of the mouth // 8—th Int. Congr. Acoust.— London, 1974.— P. 231.
196. t'Hart J. The stylization method applied to British English intonation.— Preprint/Working group on intonation.— Tokyo, 1982.— P. 23—34.
197. Thorsen N. Sentence intonation in Danish.— Preprint/Working group on intonation.— Tokyo, 1982.— P. 47—56.
198. Titze J. R. The human vocal cords: a mathematical model. Part 1 // Phonetica.— 1973.— V. 28.— P. 129—170.
199. Titze J. R. The human vocal cords: a mathematical model. Part 2 // Phonetica.— 1974.— V. 29.— P. 1—21.
200. Titze J. R. Synthesis of sang vowels using a time domain approach // Transcripts of the XI Symposium on Care and the Professional voice, 1982.— P. 90—98.
201. Umeda N. Vowel duration in American English // JASA.— 1975.— V. 58.— P. 434—445.
202. Viitanen J., Karjalainen M., Laine U. On the development of a reading machine for the blind // IV Nordic Meeting on medical and biological engineering, Copenhagen, 1977.— P. 1—31.
203. Vivaldi E., Sandri S., Miotti C. Real—time text processing for Italian speech synthesis // ICASSP-79, Washington.— 1979.— P. 880—883.
204. Williams C., Stevens K. N. Emotions and speech: some acoustical correlates // JASA.— 1972.— V. 52.— P. 1238—1250.
205. Wu Z. Rules intonation in standart Chinese.— Preprint/Working group on intonation.— Tokyo, 1982.— P. 95—108.
206. Wu H. Y., Badin P., Cheng J. M., Guerin B. Simulation du conduit vocale: réalisation de la variation continue de longuer dans un modele Kelly—Lochbaum.— Effects de l'echan tillonage spatial de la fonction d'aire // *Bull. du Laboratoire de la Communication Parlée.*— 1987.— P. 1—27.