ТЕОРЕТИЧЕСКИЕ ОСНОВЫ РЕЧЕВЫХ ТЕХНОЛОГИЙ

ТЕЗИСЫ КРАТКОГО КУРСА ЛЕКЦИЙ

Сорокин Виктор Николаевич
Институт проблем передачи информации РАН
vns@iitp.ru

Лекция 1

ВВЕДЕНИЕ

Способность к речи отличает человека от других представителей животного мира.
Речь не только является способом передачи информации от человека к человеку, но также служит своеобразным механизмом хранения знаний.

Посредством речи можно управлять поведением человека, вводить в гипноз и даже изменять биохимию крови. Долгое время люди верили, что с помощью заклинаний можно влиять на природные процессы, например, вызывать дождь.

Известная сказка о пещере, которая открывается при произнесении пароля, в принципе, может оказаться реализацией механизма распознавания с помощью резонаторов. Такой механизм, как кажется, вполне мог быть создан на основе знаний механики в средние века.

На первых этапах исследования свойств речи применялась методика анализа физических явлений, в которой теоретические построения проверялись экспериментами.

Еще в 18-м веке предпринимались попытки механического синтеза речи путем управления формой акустической системы или с помощью набора резонаторов, повторяющих форму речевого тракта.

Разработка теории звука в 19-м веке позволила Гельмгольцу определить частоту резонанса речевого тракта с наибольшей энергией для различных гласных. Он назвал эту частоту "формант". В своих экспериментах он создавал резонаторы с известной частотой и слушал, как они откликаются на произнесения разных гласных.

Тот факт, что речевой сигнал представляет собой изменение акустического давления, был использован при изобретении фонографа и телефона.

Исследование свойств периферического отдела слуховой системы привело к гипотезе о том, что взаимодействие колебаний базилярной мембраны во внутреннем ухе с волосковыми клетками создает возможность спектрального анализа звуков. Эта гипотеза привела к созданию спектрографа – устройства, которое показывает изменение амплитудно-спектрального состава речевого сигнала во времени (сонограмму). Такое представление получило название "видимой речи", и существенно продвинуло речевые исследования.

Осознание модуляционной природы речевого сигнала позволило снять противоречие между медленными движениями артикуляторных органов и широкой полосой частот, необходимой для достаточной разборчивости речи, что привело к разработке разнообразных систем сжатия речи в каналах связи. Речевой сигнал является результатом

волновых процессов в речевом тракте, где артикуляторные движения изменяют его форму и, соответственно, частоты и амплитуды его резонансов. Скорость изменения этих параметров примерно соответствует скорости движения артикуляторов. Интенсивные исследования привели к созданию практически приемлемых методов определения медленно меняющихся параметров спектра речевого сигнала, передачи их по каналу связи, и последующим синтезом речевого сигнала по этим параметрам. Так появились системы мобильной связи, которые принципиально изменили условия общения людей.

В первой половине 20-го века была поставлена задача автоматического распознавания речи. Эта задача формулировалась как задача создания фонетической пишущей машинки, переводящей речь в текст. При этом исходили из предположения, что звучащая речь состоит из последовательности фонем, соответствующих буквам текста.
В 60-х годах прошлого века начался бум исследований в области автоматического распознавания речи, главным образом, в интересах военных ведомств. Предполагалось, что управление объектами с помощью речевых команд облегчит деятельность пилотов и космонавтов. Однако вскоре выяснилось, что уровень ошибок распознавания настолько высок, что создает недопустимый риск в исполнительных системах. Поэтому финансирование от военных источников резко сократилось.

Однако за это время были получены фундаментальные результаты относительно свойств речи. Оказалось, что речь не представляет собой последовательность фонем, в виде, например разноцветных пасхальных яиц в лотке, а, скорее, яичницу-болтушку. Спектрально-временные свойства отрезка речевого сигнала, который воспринимается как определенный звук, не описываются никакими инвариантами, а зависят от множества условий. Эти условия будут рассматриваться в следующей лекции.

Также было установлено, что информации на акустическом уровне недостаточно для распознавания со сколько-нибудь приемлемой ошибкой, и необходимо использовать лексические, грамматические и семантические модели конкретного языка.

Так выяснилась чрезвычайно сложная информационная структура речевого сигнала, которая с трудом поддается математическому описанию.

В задаче распознавания команд проблему глубинного анализа речевого сигнала пытались обойти за счет компенсации различия в скорости произнесения одного и того же слова разными людьми. В нашей стране одновременно двумя учеными был предложен метод, который за рубежом был назван "dynamic time warping". Этот метод состоял в оценке меры сходства произнесенного слова с эталоном с использованием динамического программирования. Однако оказалось, что этот метод не работает в задаче распознавания независимо от диктора, в задаче с настройкой на диктора необходимо повторять обучение чуть ли не каждую неделю, и вероятность ошибки сильно зависит от внешних условий. К тому же, этот метод принципиально не применим к распознаванию слитной речи.

В результате примерно до середины 70-х годов прошлого века в области автоматического распознавания образовался кризис. Этот кризис был частично преодолен путем использования статистического подхода, который получил название метода скрытых Марковских моделей и практически полного отказа от анализа структуры речевого сигнала. В начале 70-х годов прошлого века был опубликован алгоритм оценки характеристик Марковской цепи, искаженной случайным шумом. В области распознавания речи за элементы марковской цепи принимаются спектральные характеристики сегмента речевого сигнала на коротком интервале, и на материале

обширной обучающей выборки оценивается вероятность появления последовательности сегментов для каждого слова из заданного словаря.

Оказалось, что такой подход обеспечивает расширение словаря до десятков и сотен тысяч словоформ. Наиболее успешная реализация этого подхода была выполнена супругами Baker, которые создали компанию Dragon. Все существующие коммерческие системы распознавания речи фактически являются клонами системы Baker.

Этот подход относится к классу "ignorance based approach", в котором отказываются от исследования специфических свойств объекта, полагаясь исключительно на статистические характеристики. Адепты такого подхода используют аргументы типа "самолеты не машут крыльями, а летают дальше и быстрее птиц". Однако при этом упускают различие в доступной мощности и то, что теория подъемной силы была создана, в том числе, при исследовании крыльев птиц. К тому же так и не создан махолет, хотя доказано, что мышечной силы тренированного человека достаточно для полета.

Все же нужно признать, что метод скрытых марковских моделей описывает определенные свойства речевого сигнала, хотя и не позволяет снизить ошибку распознавания для любого диктора до приемлемых величин. Исследования этого метода показали его критическую зависимость от различия в условиях обучения и эксплуатации, в результате чего ошибка распознавания слов вместо декларируемых 5 – 8% возрастает до неприемлемых величин в десятки процентов.

Относительный успех коммерческих систем распознавания речи на английском языке в значительной степени связан со строгой грамматической структурой литературного английского языка, диктующей определенную последовательность грамматических элементов. Для других языков, в частности, для русского, эти ограничения значительно слабее.

Искушение чисто технического подхода сказалось и при создании систем синтеза речи по произвольному тексту. На первых этапах исследования этой проблемы пытались использовать математические модели речеобразования. Наиболее успешным оказался так называемый формантный синтезатор, в котором выполнялось управление резонансными частотами и их амплитудами. Эта модель оказалась чувствительной к шумам во внешней среде. К тому же она требовала формирования команд управления вручную, что препятствовало созданию разных голосов.

Тогда решили обойтись без моделей речеобразования, и синтезировали речь из заранее нарезанных сегментов реальных речевых сигналов с помощью различных способов сшивки. Этот метод получил название компиляционного. Такой подход также не позволяет синтезировать произвольные голоса.

Главный недостаток формантного и компиляционного синтеза состоит в недостаточно адекватном формировании речевого сигнала, в результате чего создается когнитивная нагрузка на слушателя, и в условиях распределенного внимания разборчивость такой синтетической речи катастрофически падает.

Общепризнано, что окончательное решение проблемы синтеза речи состоит в использовании полной модели речеобразования, так называемого артикуляторного синтезатора.

Неудачная попытка игнорирования структуры речевого сигнала при распознавании человека по его голосу была предпринята в начале 60-х годов прошлого века. Этот подход получил название "отпечатков голоса" по аналогии с отпечатками пальцев. Было обнаружено, что линии равного уровня на трехмерной сонограмме речевого сигнала в координатах (время, частота, амплитуда) выглядят как папиллярные узоры, что и дало название этому методу. Метод оказался абсолютно неработоспособным в силу изменчивости спектра речевого сигнала в зависимости от внешних условий, типа, направления и расстояния до микрофона, громкости произнесения.

Несколько утрируя современное положение в области речевых технологий, можно сравнить его с эрой паровых автомобилей. Ехать можно, но медленно и не далеко.

Формальные и чисто технические приемы в области синтеза и распознавания речи, а также распознавания диктора, достигли пределов своих возможностей, не удовлетворяя требованиям большинства задач речевого общения человека к автоматом.

Необходимо вновь вернуться к традиционному подходу исследования механики процессов речеобразования и информационной структуры речевого сигнала. При этом нельзя игнорировать тот факт, что все задачи анализа речевого сигнала являются некорректными обратными задачами.

Более того, и задача распознавания слов, и задача понимания речи также являются некорректными обратными задачами.

Задачи определения формы речевого тракта, его резонансных частот, параметров артикуляторных движений и команд управления артикуляторами есть некорректные обратные задачи в силу свойств волнового уравнения, кинематики артикуляторов и многомерной системы управления артикуляцией.

Решение таких задач с более или менее приемлемой погрешностью требует использования дополнительной информации, накладывающей ограничения на возможные решения. На акустическом уровне такая информация состоит в математических моделях кинематики и акустических процессов речевом тракте, а также модели системы управления артикуляцией.

Речевой сигнал предназначен для передачи информации от человека к человеку, и в нем присутствует сложная кодовая структура, позволяющая понимать речь в условиях шумов и различия произношений людей.

Только используя свойства речи и слуха можно достичь уровня речевых технологий сопоставимого с характеристиками речевого общения людей.

Лекция 2

Виды речевых технологий и области их применения

1. Сжатие речевого сигнала в каналах связи.
2. Синтез речи по произвольному тексту.
3. Автоматическое распознавание и понимание речи.
4. Распознавание диктора.
5. Распознавание эмоционального и физического состояния.
6. Диагностика заболеваний.
7. Коррекция нарушений слуха.
8. Обучение иностранному языку.

Сжатие речевого сигнала в цифровых каналах связи (speech compression)

Полоса частот в мобильных системах связи в канале «абонент - станция» ограничена. Поэтому уменьшение скорости передачи речевого сигнала позволяет увеличить количество одновременных разговоров.

В проводных каналах связи уменьшение скорости передачи речевого сигнала снижает стоимость разговора.

Установлено, что для высококачественной передачи речевого сигнала требуется полоса частот до 5 кГц. Тогда в системах импульсно-кодового кодирования (ИКМ), согласно теореме Найквиста-Котельникова, частота передачи импульсов должна быть не меньше 10 кГц. При квантовании амплитуды сигнала на 16 бит получаем минимальную скорость передачи 160 кбит/с.

В современных системах сжатия скорость передачи речевого сигнала снижена больше, чем на порядок – до 10 кбит/с и даже 2 кбит/с.

Сжатие речевого сигнала выполняется по схеме «анализ - синтез».

На передающем конце вычисляются некие параметры сигнала, эти параметры передаются по каналу связи, и на приемном конце по этим параметрам восстанавливается (синтезируется) речевой сигнал.

Синтезированный сигнал отличается от исходного. Поэтому в системах сжатия речи существенную роль играет способность слуха к компенсации искажений речевого сигнала.

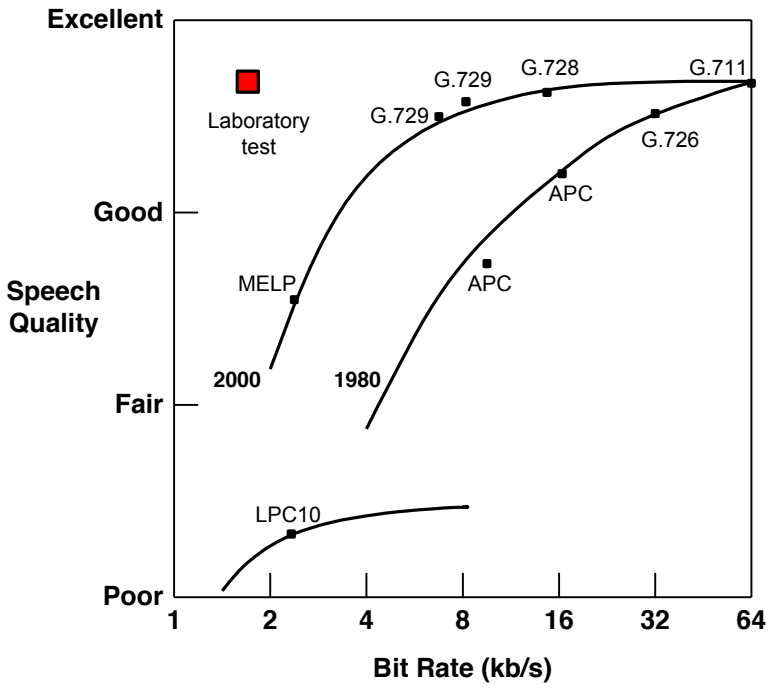


Рис. 2.1. Субъективная оценка качества восстановленного речевого сигнала в зависимости от скорости передачи и метода сжатия.

Снижение скорости передачи достигается за счет использования моделей речевого сигнала.

Синтез речи по произвольному тексту (TTS – Text-To-Speech Synthesis)

Значительную долю социальной и интеллектуальной информации человек получает путем речевого общения. В некоторых случаях речевой сигнал является предпочтительным или даже единственным источником информации, например, в темном помещении или на удалении, в частности, по телефонному каналу.

Области применения:

- читающие машины –
  для людей с ослабленным зрением;
  для слепых;
  при обучении;
- чтение электронных писем, SMS;
- речевой терминал (говорящая машина) для инвалидов (церебральный паралич, тяжелая патология речи, удаление гортани);
- получение инструкций в речевой форме, например, при сборке агрегатов или в навигационной системе автомобиля;
- доступ к информационно-справочным системам на удалении;
- вывод информации в речевой форме, например, о параметрах полета;
- в автоматизированных системах перевода (карманные переводчики, перевод с одного языка на другой с последующим речевым выводом).

Система синтеза речи по тексту состоит из трех основных блоков:

- перевод текста из буквенной формы в фонетическую;

- просодический анализ текста, т.е. определение длительности каждого речевого сегмента и частоты основного тона;
- формирование речевого сигнала.

Существует три основных метода формирования речевого сигнала:

- компиляционный (concatinative);
- формантный (formant);
- артикуляторный (articulatory).

Синтез речи по правилам характеризуется следующими параметрами:

- разборчивость;
- натуральность;
- помехоустойчивость относительно разных типов помех при разных отношениях сигнал-шум;
- возможность синтезировать текст голосами различных дикторов;
- задержка распознавания синтезированной речи человеком.

По этим параметрам формантный и компиляционный синтез заметно уступают естественной речи, что существенно ограничивает область их применения.

Автоматическое распознавание и понимание речи (speech recognition, speech understanding)

Автоматическое распознавание подразумевает перевод в текстовую форму слов в речевом сигнале, например, при поиске ключевых слов. Понимание речи не требует распознавания всех слов потоке речи, но должно отображать смысл речевого сообщения.

Области применения:

- диктофоны;
- запрос на получение информации по телефонному каналу (справки о расписании транспорта, навигация по адресам или типам объектов, заказ билетов);
- запрос на получение информации о состоянии компьютеризованных систем, например, о параметрах полета;
- голосовое управление браузером, компьютером, играми;
- управление исполнительными системами (набор телефонного номера с голоса, управление приборами автомобиля, бытовыми приборами);
- ввод информации (числовые данные, заполнение граф в формализованных документах, диагноз, описание рентгеновских снимков);
- перевод с одного языка на другой, когда входная информация представлена в речевой форме.

В зависимости от условий обучения существуют три типа систем автоматического распознавания речи:

- с настройкой на диктора;
- с адаптацией к диктору;
- независимо от диктора.

В зависимости от стиля произношения системы распознавания также разделяются на три типа:

- с изолированным произнесением слов, как, например, при распознавании команд, где между командами имеется относительно длительная пауза;
- с раздельным произнесением слов, т.е. небольшими паузами между словами;
- системы, допускающие слитное произнесение.

Подавляющее большинство известных систем распознавания речи основано на применении метода скрытых марковских моделей (HMM – Hidden Markov Model). Скрытой марковской моделью называется марковский процесс, который не наблюдается непосредственно, а искажен некоторым случайным процессом. Решение о том, какая последовательность элементарных символов Марковской цепи реализована в речевом потоке, принимается методом максимального правдоподобия.

Параметрами скрытой марковской модели являются:
а. возможные состояния процесса,
б. вероятность перехода из одного состояния в другое,
в. вероятность искажения наблюдаемого состояния.

Развитие метода скрытых марковских моделей в применении к распознаванию речи началось после того, как была доказана сходимость простого итеративного алгоритма к параметрам модели (Baum, 1972).

Синхронный метод скрытых марковских моделей состоит в оценке принадлежности спектрального разреза речевого сигнала к какому-либо классу через равноотстоящие моменты времени. Отсев одинаковых состояний и согласование неизвестной последовательности с эталоном производится дискретным вероятностным вариантом метода динамического программирования - обычно алгоритмом Витерби. Метод скрытых марковских моделей относится к так называемому "математическому подходу" (часто характеризуемому как "ignorance based approach"), в котором практически не принимаются во внимание свойства речевого сигнала, как переносчика информации. Это позволяет использовать большие объемы речевых сигналов для обучения системы распознавания практически без участия человека. Математически, преимущество этого метода состоит в возможности реконструкции вероятностной меры сходства с эталоном.

Вместе с тем, этому методу присущи принципиальные недостатки: предполагается, что последовательность состояний описывается марковской цепью первого порядка, процессы считаются независимыми, а исходные вероятностные распределения полагаются либо многомерными гауссовскими, либо смесью гауссовских распределений. Кроме того, используется ряд математических ограничений, относительно которых неизвестно, удовлетворяются ли они при распознавании речи.

Современные системы распознавания речи неустойчивы относительно характеристик канала связи и помех.

Вероятность правильного распознавания слов небольших словарей может доходить до 98 - 99%. Производители некоторых коммерческих систем декларируют, что вероятность распознавания слов в слитной или раздельной произвольной речи составляет 93 - 95%. Однако эти оценки справедливы только если условия обучения системы

совпадают с условиями эксплуатации. В противном случае вероятность правильного распознавания может упасть до 40 – 60%.

Как синтез произвольного текста, так и распознавание произвольной речи, требуют использования синтаксической модели языка и анализа смыслового содержания для того, чтобы соответствовать ожиданиям человека.

Если количество ошибок при автоматическом распознавании или синтезе речи превышает некоторый порог, то человек отказывается использовать эти системы. Весьма ограниченное распространение коммерческих систем указывает на то, что они не удовлетворяют ожиданиям большинства потенциальных пользователей.

Распознавание диктора (speaker recognition)

Два вида распознавания диктора: идентификации (identification) и верификация (verification, authentication).

Идентификация диктора состоит в определение вероятности того, что два речевых сообщения принадлежат одному и тому же человеку.

Области применения:

- криминалистическая экспертиза;
- антитеррористические мероприятия;
- контрразведка;
- слежка за определенными категориями граждан;
- идентификация диктора на фонограммах или в реальном времени во время конференций или интервью с несколькими участниками.

В силу того, что содержание сообщений или условия их регистрации могут различаться, вероятность правильной идентификации диктора невысока, и не принимается как доказательство в судах.

Верификация диктора состоит в подтверждении личности диктора на основе биометрии его голоса. Поскольку в этом случае возможно обучение системы верификации на голос диктора, а сам диктор заинтересован в том, чтобы система верификации подтвердила его личность, то ошибки (допуск самозванца и отказ законному пользователю) могут быть доведены до малых величин.

Поэтому определена весьма широкая область применения систем верификации:

- санкционирование распоряжения финансовыми процессами по электронным или телефонным каналам (управление банковским счетом, электронная коммерция, подтверждение права пользования кредитной картой);
- разрешение на смену пароля или PIN-кода;
- доступ к компьютеру или отдельным программам компьютера (вход в Интернет, доступ к конфиденциальным документам, базам данных и т.д.);
- разрешение на вход в помещение, открывание сейфа;
- управление механизмами и системами (например, запуск двигателя автомобиля);
- мониторинг того, кто, когда и к каким компьютерным ресурсам имел доступ.

Распознавание эмоционального и физического состояния

Темп речи, интонация и громкость речевого сигнала несут информацию о психическом состоянии человека, особенно в случае, когда имеются образцы его речи в нормальном состоянии.

Диагностика заболеваний

Темп речи, интонация, длительность пауз также могут дать информацию психических заболеваниях.

Ведутся интенсивные исследования по ранней диагностике заболеваний гортани – рак, полипы, паралич, анемия.

Установлена возможность диагностики асфиксии, мозговых повреждений, гипербиллирубина и синдрома Дауна по крику новорожденных.

Коррекция нарушений слуха

Для восстановления слуха используются имплантируемые электроды. С целью повышения разборчивости воспринимаемой речи применяется предобработка речевого сигнала (усилители звука с автоматическим управлением коэффициента усиления, подавление шумов на спектральном уровне). В некоторых случаях используются направленные микрофоны или система микрофонов с последующей цифровой обработкой для улучшения отношения сигнал/шум.

Обучение иностранному языку

Обучающемуся проигрывают слова и фразы иностранного языка, и показывают требуемые контур основного тона и сонограмму («видимую речь») вместе с этими же функциями, полученными путем анализа произнесенных слов и фраз.

Изменчивость речевого сигнала

Трудности в автоматическом распознавании речи связаны с изменчивостью акустического образа, приписываемого одному и тому же речевому элементу, например, слову.

Идея последовательности фонем в речи (like a string of Easter eggs in a groove), навеянная аналогией с буквенной записью в письме, не соответствует действительности. Информация о некотором звуке, субъективно идентифицируемом как фонетический элемент, распределена на значительном интервале времени (more like to scrambed eggs) и подвержена разнообразным изменениям.

Существует несколько видов изменчивости, каждая со своими закономерностями. Условно можно различать изменчивость, связанную с внешними условиями, дикторскую изменчивость и контекстную изменчивость. Ниже перечислены наиболее часто встречающиеся виды изменчивости.

- Акустические помехи внешней среды, среди которых наиболее часто встречаются нестационарные помехи в виде речи посторонних дикторов. Борьба с такими помехами, получившими название «coctail party effect», пока не увенчалась успехом.
- Искажение характеристик речевого сигнала в тракте между микрофоном и аналого-цифровым преобразователем. Сюда входят наводки электрических линий и шумы электронных цепей, разные коэффициенты усиления. Особенно велики помехи и замирания, характерные для радиоканалов с аналоговой передачей сигнала.
- Искажения амплитудно-частотных и временных характеристик речевого сигнала в результате реверберации замкнутых помещений. В частности, реверберация

приводит к длительному присутствию резонансных колебаний на смычках после гласных звуков.

- Искажение амплитудно-частотных характеристик речевого сигнала, связанное с различием типов микрофонов, расстояния от рта диктора до микрофона и направления микрофона. Близко расположенные микрофоны улучшают отношение «речевой сигнал - акустические шумы среды», однако при этом возникает эффект ближнего акустического поля, при котором амплитудно-частотные характеристики сигнала в низкочастотной области сильно зависят от расстояния до микрофона. К тому же, использование головных гарнитур с близко расположенным микрофоном неприемлемо для большинства пользователей.
- Изменчивость амплитудно-частотных характеристик стационарных сегментов речевого сигнала, связанная с различием размеров и формы речевого тракта дикторов.
- Различие в темпе речи дикторов, которая при прочих фиксированных условиях может доходить до 300%. Изменчивость длительности фонетических элементов в зависимости от стиля речи, эмоционального и физического состояния диктора.
- Изменчивость громкости речи диктора и связанная с этим изменчивость амплитудно-частотных характеристик речевого сигнала. В частности, известен так называемый эффект Ломбарда, состоящий в повышении уровня высокочастотных компонент речевого сигнала при непроизвольном повышении громкости при разговоре в присутствии помех.
- Разнообразие динамических характеристик речи, связанное с различием масс артикуляторных органов и особенностями артикуляции дикторов, стилем речи, эмоциональным и физическим состоянием дикторов.
- Изменчивость длительности и акустических характеристик фонетических элементов в зависимости от длительности фразы, положения относительно начала фразы и положения относительно логического ударения во фразе.
- Изменчивость граничных фонетических элементов слов в слитном потоке речи - слияния конечных и начальных фонетических элементов, оглушение, озвончение, назализация и прочие эффекты коартикуляции.

Лекция 3

Теория речеобразования

Речевой сигнал содержит информацию о том

кто говорит,

что говорит,

как говорит,

какова окружающая **среда**.

Эта информация определяется свойствами процессов речеобразования и восприятия речи.

Строение речевого тракта

В процессе управления артикуляцией принимают участие 27 мышц, но, некоторые мышцы являются антагонистами друг друга, а другие мышцы создают усилия в совпадающих направлениях.

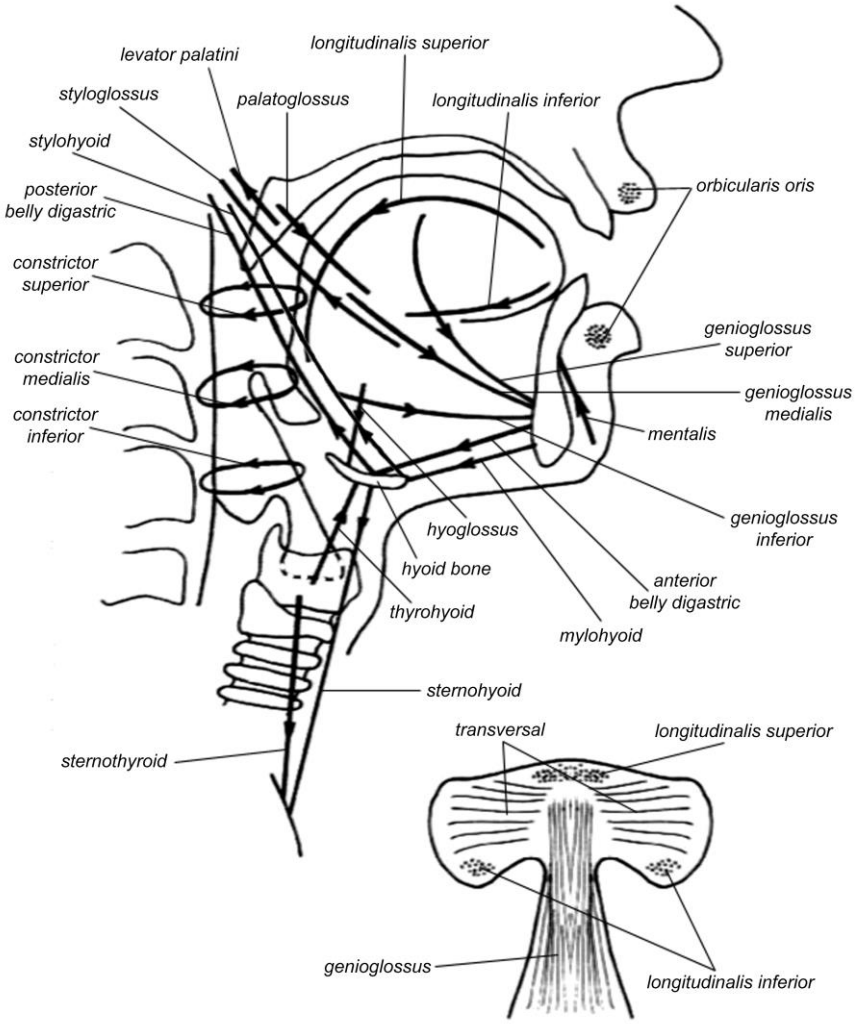


Рис. 3.1. Строение мышечного аппарата речевого тракта.

Упругие деформации языка

Упругие деформации языка под воздействием сокращения внешних и внутренних мышц описываются дифференциальным уравнением в частных производных в полярной системе координат, поскольку в нейтральном состоянии форма языка близка к полуокружности с радиусом R_0.

Уравнение упругих деформаций языка

$$\frac{1}{R_0^2}\frac{\partial^2}{\partial\varphi^2}\left(EJ_z\frac{\partial^2 u}{\partial\varphi^2}+u\right)+cu+r\frac{\partial u}{\partial t}+\rho\frac{\partial^2 u}{\partial t^2}=-\frac{\partial^2 M}{\partial\varphi^2} \qquad (3.1)$$

где φ - угол в полярной системе, $\varphi = 0$ в нижней точке корня языка, $\varphi = \pi$ на кончике языка; $M(\varphi)$ - изгибающий момент, создаваемый внешними и внутренними мышцами; E - модуль упругости тканей языка; J_z — момент инерции поперечного сечения; ρ -погонная плотность тканей языка; $\rho = \rho_T S(\varphi)$, r_T - показатель вязкого трения; $S(\varphi)$ - площадь поперечного сечения; $r(\varphi)$ - погонный коэффициент вязкости; $c(\varphi)$ – погонная упругость подстилающих тканей. Примерные оценки параметров уравнений таковы: площадь поперечного сечения языка S = 2 см^2; $\rho_T = 1.1 г/см^3$; момент инерции J_z = 0,4 см^4; модуль упругости E=10^5 Па (10^6 г/см·с^2); погонная упругость c =25 г/см·с^2; погонное сопротивление r = 170 г/см·с; $R_0 \approx 3 – 4$ см.

Решение этого уравнения можно представить в виде произведения двух функций, одна из которых зависит только от полярного угла, а другая – только от времени.

Тогда поверхность языка описывается как

$$u(x,t)=\sum_k c_k U_k(\varphi)T_k(t)+R_{0,}\quad k=1,...,5$$

где U_k - собственные функции упругих деформаций. Подгонка решения уравнения упругих деформаций языка к измерениям поверхности языка на рентгенограммах показала, что для тела языка достаточно использовать 4 собственные функции. Кончик языка часто выступает как самостоятельных артикуляторный орган, и для него достаточно использовать только одну собственную функцию. Таким образом, форма языка описывается пятью собственными функциями, показанными на Рис. 3.2.

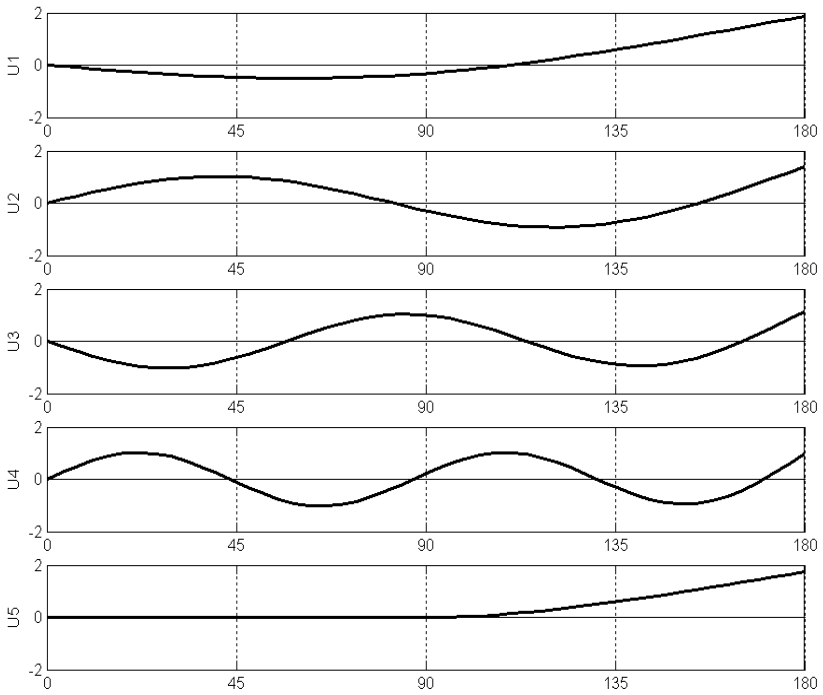


Рис. 3.2. Собственные функции упругих деформаций языка.

Соответствующие временные моды описываются обыкновенными дифференциальными уравнениями:

$$T_k'' + \frac{r}{\rho_t} T_k' + \omega_k^2 T_k = f_k(t),$$

r и ρ_t – коэффициент вязкого сопротивления и плотность тканей языка,

f_k - сила, приложенная мышцей.

Описание формы языка в средне-сагиттальной плоскости и динамики ее изменения уравнениями (1) и (2) гораздо нагляднее и удобнее, чем использование метода конечных элементов.

Уравнение упругих деформаций губы

$$\frac{\partial^2}{\partial x^2}\left[J_y (E+N) \frac{\partial^2 u}{\partial x^2} \right] - NS \frac{\partial^2 u}{\partial x^2} + cu + r \frac{\partial u}{\partial t} + \rho \frac{\partial^2 u}{\partial t^2} = F(x,t) \qquad (3.2)$$

где $u(x,t)$ — отклонение средней линии губы от нейтрального положения; J_y — момент инерции сечения относительно оси y; E — модуль упругости губы при расслабленном состоянии кольцевой мышцы; N — удельное натяжение, создаваемое кольцевой мышцей; S — площадь поперечного сечения; c — упругость подстилающего слоя; r — коэффициент вязкого трения; ρ — погонная плотность тканей; $F(x, t)$ — сила, создаваемая остальными мышцами.

Управляемые параметры в модели речевого тракта

Число управляемых параметров в кинематической модели артикуляции равно 17. Это минимальный набор параметров, обеспечивающий воспроизведение потенциально возможных форм речевого тракта для большинства языков:

- угол поворота нижней челюсти
- горизонтальное смещение нижней челюсти
- угол поворота языка
- вертикальная координата корня языка
- горизонтальная координата корня языка
- коэффициенты при собственных функциях, описывающих форму языка в сагиттальной плоскости
- коэффициент при собственной функции для кончика языка
- амплитуда аксиального прогиба языка
- высота гортани
- длина губ
- высота нижней губы
- угол поворота небной занавески
- 2 коэффициента при собственных функциях ширины глотки

Амплитудно-частотные характеристики кончика языка и небной занавески были определены в экспериментах с одновременным измерением электро-миограмм управляющих мышц и движений артикуляторов. Характеристика нижней челюсти была найдена в экспериментах с использованием электрической стимуляции жевательной мышцы, а для губ – с использованием внезапного растяжения углов губ.

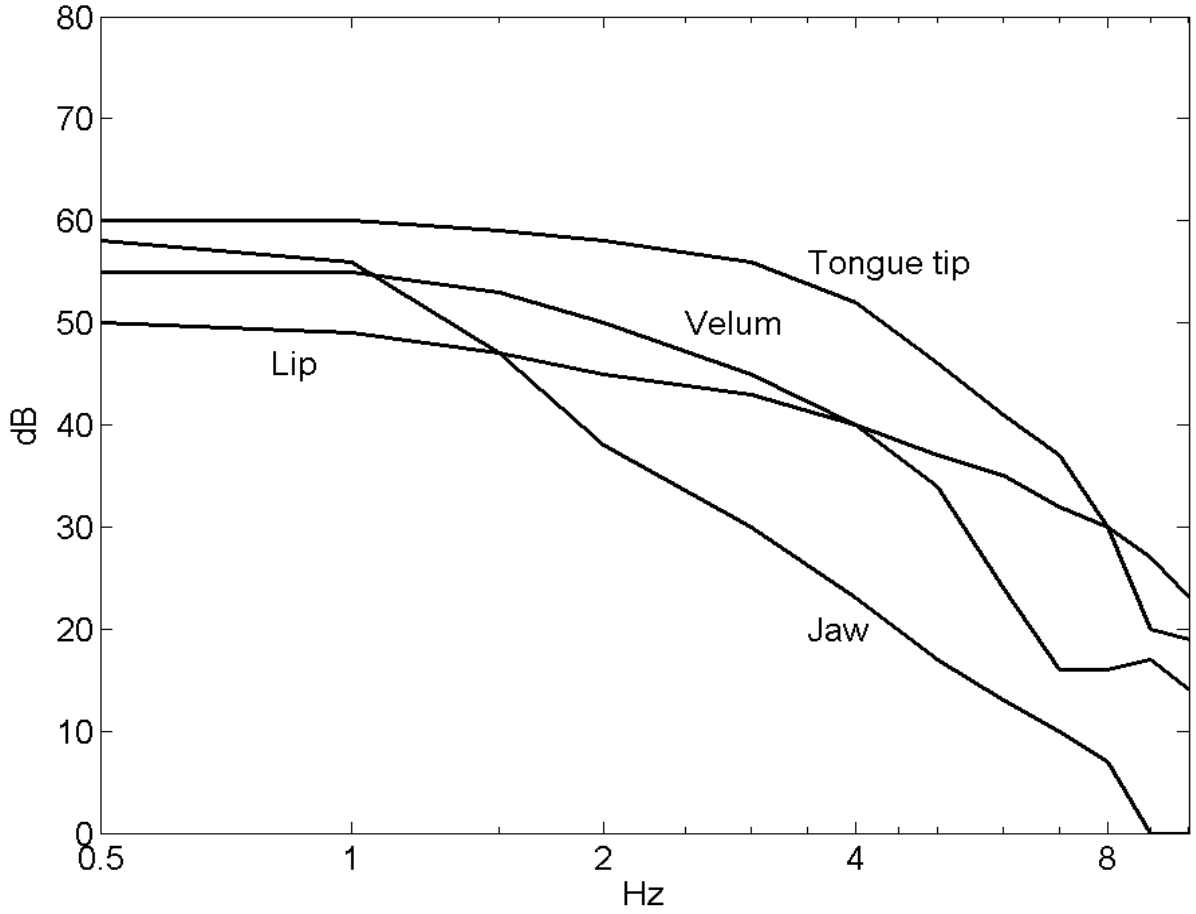


Рис. 3.3. Амплитудно-частотные характеристики артикуляторных органов

Методы измерения формы речевого тракта.

Ультразвуковое сканирование
Аналоговая ренгенография
Микролучевая ренгенография.
Электро-магнитная регистрация.
Магнитно-резонансная томография (MRI – Magneto Resonance Imaging).
Электропалатография
Глотография
Стробоскопия
Эндоскопия

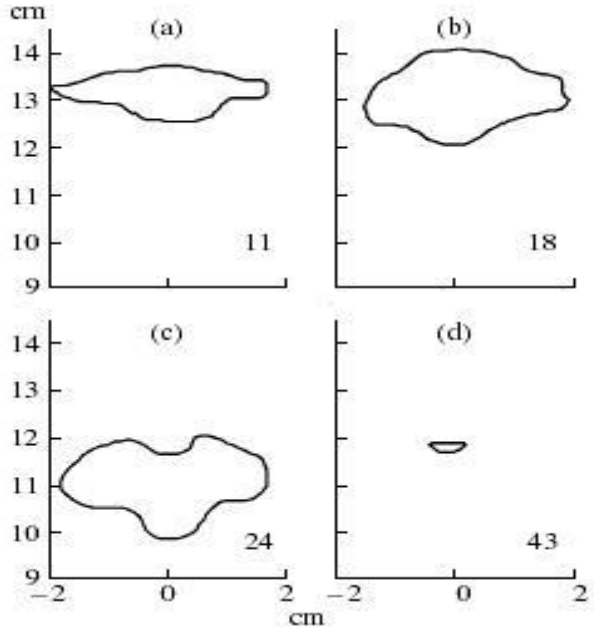


Рис. 3.4. Форма поперечного сечения речевого тракта на уровне входа в пищевод (a) и непосредственно перед небной занавеской (b), в области небной занавески (c) и в области твердого неба для гласного /*u*/ (d) по измерениям MRI/

Рис. 3.5. Электромагнитная система.

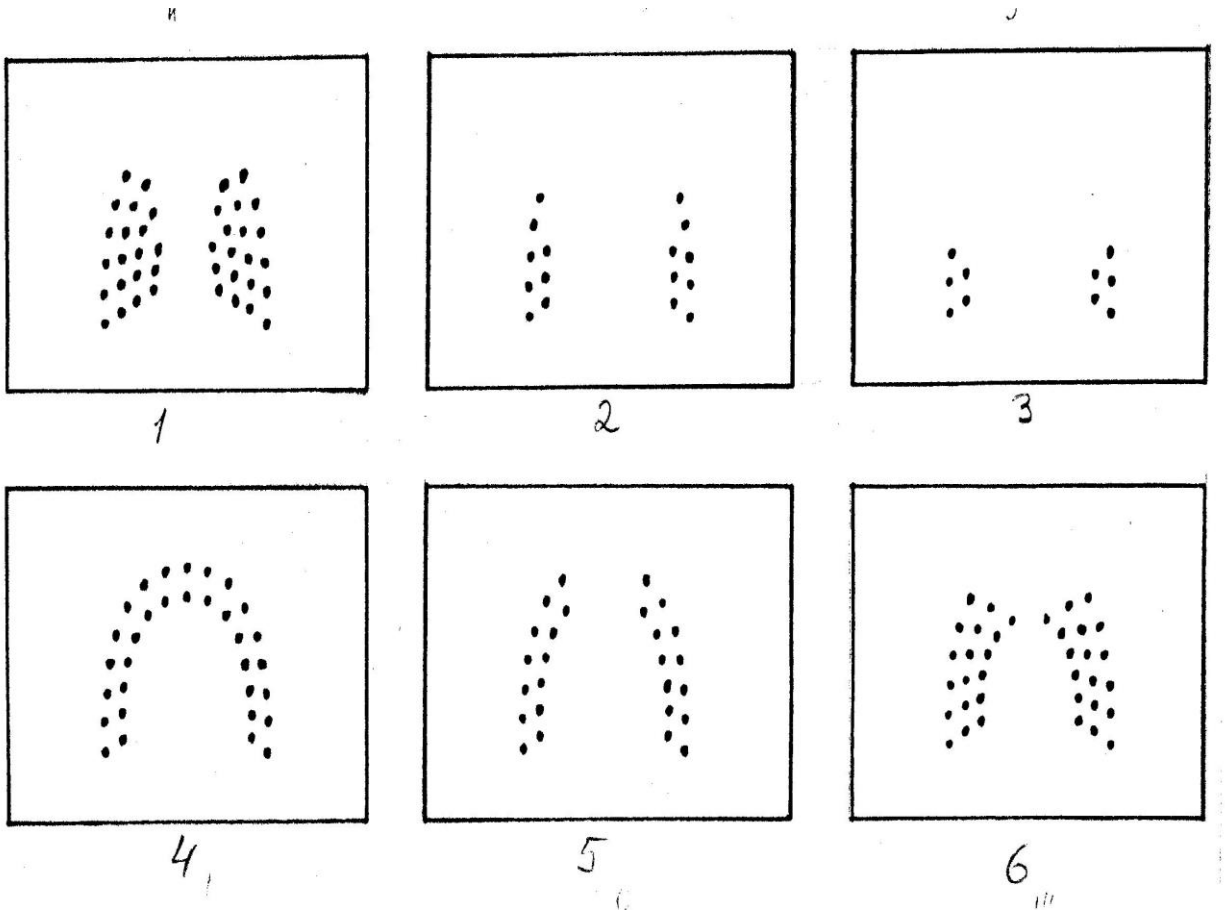


Рис. 3.6. Области контакта палатографа для разных типов артикуляции.

Голосовой источник

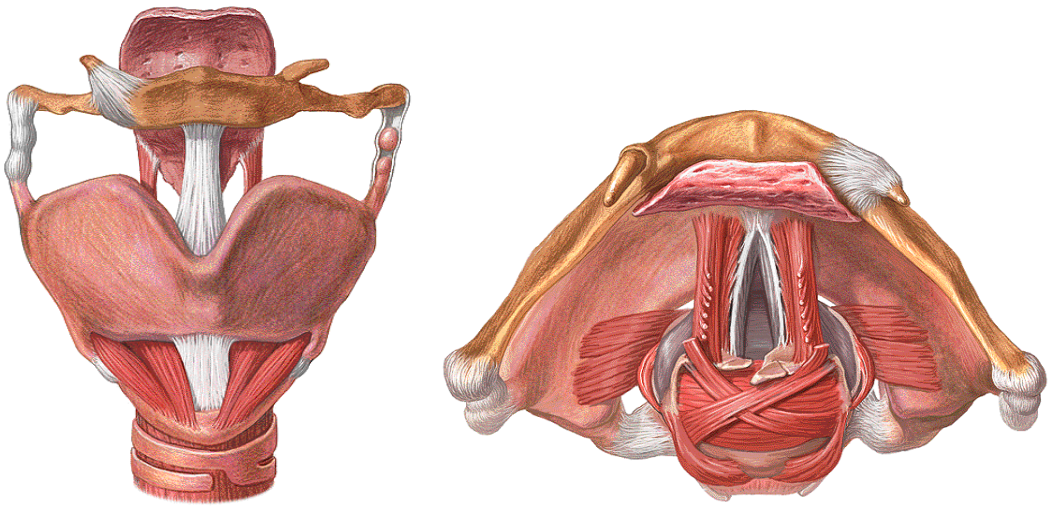


Рис. 3.7. Гортань. Слева – вид спереди во фронтальной плоскости. Справа – вид сверху в латеральной плоскости.

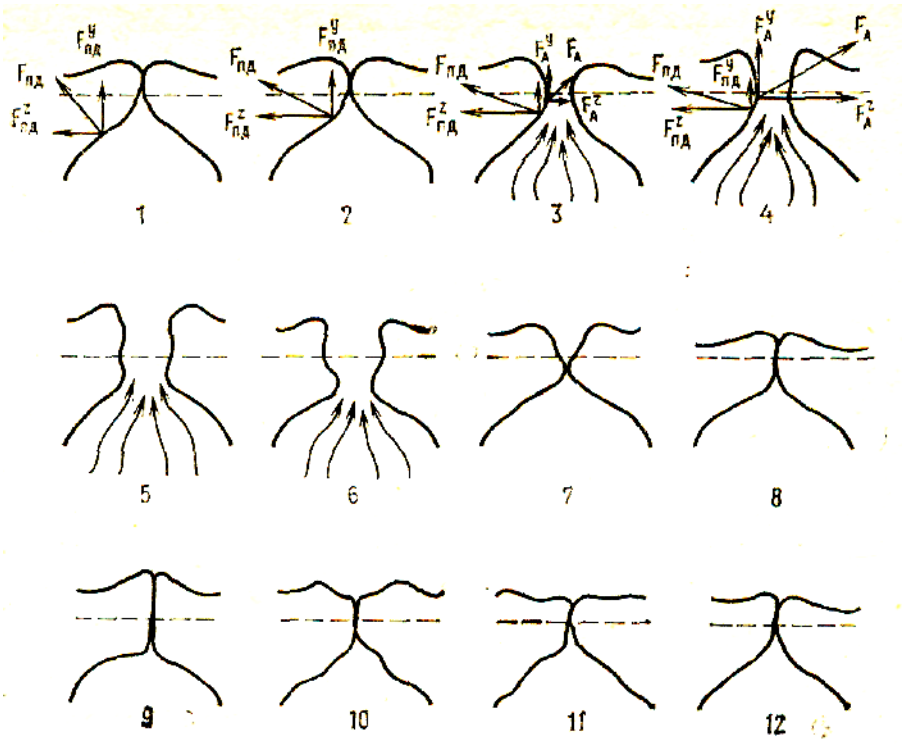


Рис. 3.8. Фазы колебаний голосовых складок.

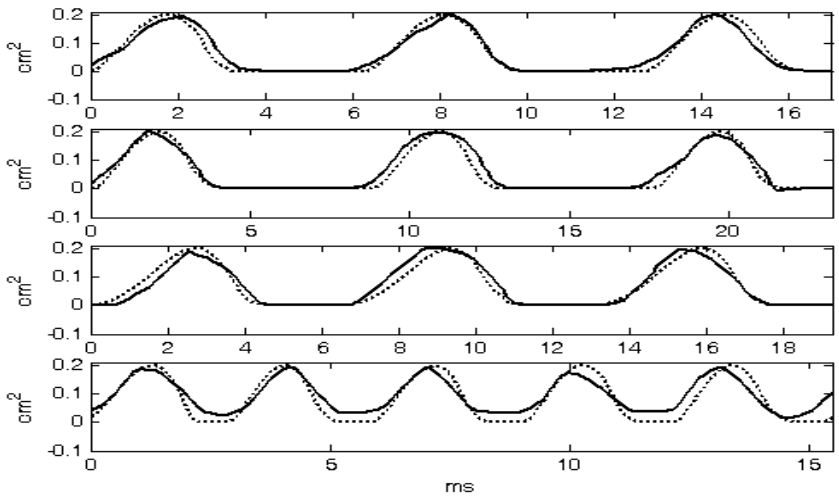


Рис. 3.9. Измеренные (——) и вычисленные (- - -) площади голосовой щели.

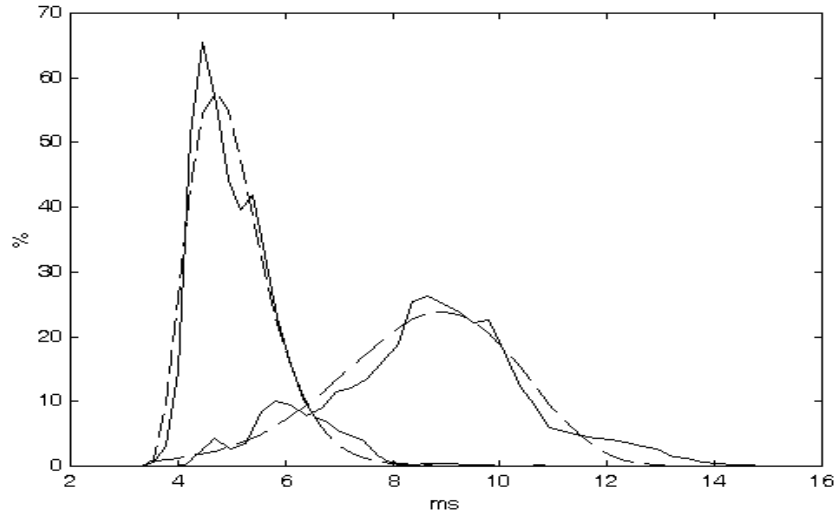


Рис. 3.10. Распределение периодов основного тона T_0 женских (слева) и мужских (справа) голосов на ударных гласных числительных русского языка (—) и их аппроксимация гамма-распределением (---).

Диапазон значений частоты основного тона F_0 у мужчин составляет примерно 65 ÷ 250 Гц со средним значением 125 Гц, а у женщин – 125 ÷ 350 Гц со средним значением 250 Гц.

Методы измерения колебаний голосовых складок

Эндоскопия
Трансиллюминация
Глоттография
Рентгенография аналоговая
Рентгенография металлических меток

Лекция 4

Теория речеобразования

Линейная акустическая теория речеобразования

Волновое уравнение (уравнение Вебстера)

Акустическая система может рассматриваться как система с сосредоточенными или распределенными параметрами в зависимости от соотношения ее геометрических размеров с длинами волн в интересующем диапазоне. Если некоторый характерный геометрический размер системы ***l*** гораздо меньше определенной длины волны λ, т.е. $l << \lambda$ то для всех волн, длина которых больше λ, акустическая система может считаться сосредоточенной. Конкретизируя это условие, например, как $10l = \lambda$, найдем граничную частоту $f_0 = c_0 / \lambda$, выше которой речевой тракт должен рассматриваться как система с рапспределенными параметрами, c_0 - скорость звука, $c_0 \approx 350$ м/сек. Длина речевого тракта около 20 см, поэтому граничная частота $f_0 = c_0 / 10\lambda$ примерно равна 175 Гц. Это означает, что во всем речевом диапазоне частот речевой тракт должен рассматриваться как система с распределенными параметрами.

Вместе с тем, в ряде случаев речевой тракт может рассматриваться как последовательное соединение резонаторов Гельмгольца, каждый из которых описывается как система с сосредоточенными параметрами. Обычно рассматриваются два резонатора Гельмгольца. Это позволяет сравнительно легко получить не слишком грубые оценки акустических параметров речевого тракта.

В акустической системе могут распространяться волны во всех трех измерениях. Если радиус цилиндрической трубы есть ***r***, то нижняя граничная частоты возникновения поперечных волн равна $f_{01} = 0.293 c_0 / r$. Принимая среднюю площадь поперечного сечения речевого тракта равной 5 см^2, получим нижнюю граничную частоту $f_{01} = 8 \quad кГц$. Более детальное рассмотрение этого вопроса показывает, что ниже частоты в 4.5 кГц в речевом тракте поперечные волны отсутствуют. Если $r << c_0 / f_{01}$, то такие трубы называются "очень узкими", и ее изгибы не влияют на распространение волн. Поэтому в речевом тракте распространяются только плоские волны, и для него справедливо одномерное волновое уравнение:

$$c_e^2(x,t)\frac{\partial}{\partial x}[A(x,t)\frac{\partial}{\partial x}P(x,t)] = \frac{\partial}{\partial t}[A(x,t)\frac{\partial}{\partial t}P(x,t)] \qquad (4.1)$$
$$c_e^2 = c_0^2(1 - Y_w c_0 / j\omega r)$$

P- акустическое давление, ***A -*** площадь поперечного сечения речевого тракта, c_0 - скорость звука в трубе с абсолютно жесткими стенками, c_e **-** скорость звука в трубе с податливыми стенками, Y_w - импеданс стенок, ω - круговая частота, ***x*** координата вдоль оси речевого тракта, ***t*** - время ***r*** - радиус или наименьшая сторона поперечного сечения.
Здесь гармонические колебания представляются как $e^{-j\omega t}$.

Граничные условия

Если не учитывать сопротивление излучения на губах, то акустическое давление $P(\boldsymbol{l}) = 0$, где $\boldsymbol{l}$ - длина речевого тракта. Сопротивление излучения приводит как бы к удлинению тракта, и давление равно нулю на некотором расстоянии от губ $l + \Delta l$, $\Delta l / l \approx 5 \div 7\%$. Акустический импеданс описывается как отношение акустического давления $\boldsymbol{P}$ к скорости $\boldsymbol{V}$ акустических колебаний частиц воздуха $\boldsymbol{Z=P/V}$, причем между давлением и скоростью имеется следующее соотношение:

$$-A\frac{\partial P}{\partial x} = \rho\frac{\partial}{\partial t}(AV)$$

$\rho_0 = 1.14 \cdot 10^{-3}\ g/cm^3$ - плотность воздуха при температуре $37^\circ C$.

Обозначим $\lambda = \omega / c_0$. Тогда импеданс излучения Z_l рассчитывается как для круглого поршня с эквивалентным радиусом a в сфере радиуса a_h (поправка Морза) при $\lambda a < 0.25$:

$$Z_l = 1 - \frac{J_1(\lambda a)}{\lambda a} - j\frac{K_1(2\lambda a)}{2(\lambda a)^2}, \qquad (4.2)$$

где J_1 - функция Бесселя первого рода, K_1 - функция Бесселя второго рода. Отношение эквивалентного радиуса ротового отверстия к радиусу головы $a / a_h \approx 0.06 \div 0.14$, и с погрешностью менее 10% импеданс излучения представляется как

$$\boldsymbol{Z_l} = \frac{\boldsymbol{P}}{\boldsymbol{V}} = \frac{(\lambda\, \boldsymbol{a})^2}{2} - \boldsymbol{j}\frac{8\lambda\, \boldsymbol{a}}{3\pi},$$

где a - эквивалентный радиус ротового отверстия, $a = \sqrt{A(l)/\pi}$.

При закрытой голосовой щели, акустическое сопротивление тканей голосовых складок близко к сопротивлению абсолютно жесткой стенки, на которой скорость колебаний частиц воздуха равна нулю, $\boldsymbol{V}$=0, и производная давления по пространственной координате также равна нулю $\partial \boldsymbol{P} / \partial \boldsymbol{x} = 0$, $\boldsymbol{x}$=0.

На открытой голосовой щели

$$\frac{\partial P}{\partial x}\Big|_{x=0} = -\rho\frac{\partial}{\partial t}\left(\frac{W}{A_0} + \frac{\partial y_{vs}}{\partial t}\right), \qquad (4.3)$$

ρ_0 - плотность воздуха, A_0 - площадь речевого тракта на уровне голосовой щели, y_{vs} - вертикальное смещение голосовых складок, W - объемная скорость воздушного потока, протекающего через голосовую щель из легких в речевой тракт.

Импеданс стенок речевого тракта в общем виде можно записать как

$$Z_w = R_w + j\left(\frac{C_w}{\omega} - \omega M_w\right), \qquad (4.4)$$

где R_w, C_w, M_w - активное сопротивление, жесткость и масса стенок, приходящиеся на единицу длины тракта. Жесткость стенок приблизительно оценивается как $C_w = E_w / h_w$, E_w- модуль упругости тканей, h_w - толщина стенок, а погонная масса $M_w = \rho_w h_w$, ρ_w- погонная плотность тканей ($\rho_w \approx 1.05 \div 1.1\, g/cm^2$). Ткани твердого неба обладают преимущественно упругим импедансом, который уменьшает локальную скорость звука, а мягкие ткани имеют импеданс инерционного типа, который скорость звука увеличивает. При этом влияние стенок сказывается, в основном, на низких частотах.

Резонансные частоты речевого тракта

Существует ряд методов решения волнового уравнения относительно резонансных частот.

Один из простейших методов применим к уравнению без потерь в предположении независимости площади поперечного сечения *A* от времени или, по крайней мере, ее достаточно медленного изменения за время пробега акустической волны от голосовой щели до губ и обратно. Этот метод состоит в разделении переменных (метод Фурье), где функция двух переменных - координаты *x* и времени *t* представляется в виде произведения двух функций, каждая из которых зависит только одного аргумента ***P(x,t)=X(x)T(t)***.

Это приводит к системе двух обыкновенных дифференциальных уравнений

$$(\boldsymbol{AX}')' + \lambda_{\boldsymbol{k}}^2 \boldsymbol{A}\,\boldsymbol{X} = 0,$$
$$\boldsymbol{T}'' + \lambda_{\boldsymbol{k}}^2 \boldsymbol{c}_0^2 \boldsymbol{T} = 0, \quad \boldsymbol{k} = 1,2,\ldots$$

$\lambda_{\boldsymbol{k}}$ - собственные числа волнового уравнения, $\boldsymbol{F_k} = 2\pi\lambda_{\boldsymbol{k}}\boldsymbol{c}_0$ - резонансные частоты.

Другой метод, метод длинных линий позволяет учесть податливость стенок и потери, нелинейно зависящие от частоты. Речевой тракт разбивается на N последовательных секций длины Δl. При расчете резонансных частот речевого тракта поперечное сечение каждой секции принимается круговым, что справедливо для всех длин волн, больших, чем $8\Delta l$. Таким образом, речевой тракт аппроксимируется последовательностью цилиндрических секций. Передаточная функция речевого тракта определяется как отношение объемной скорости акустических колебаний у губ к объемной скорости на входе в речевой тракт со стороны голосовой щели. Резонансные частоты тракта находятся как полюса передаточной функции.

Потери в речевом тракте

Потери в речевом тракте связаны с вязким сопротивлением, потерями на испарение, на теплопроводность, излучение и колебания стенок. Разные виды потерь по разному зависят от частоты, и ширина полосы *k*-го резонанса:

$$\Delta F_k = \frac{\alpha_1}{1+\alpha_2 F_k^2} + \alpha_3\sqrt{F_k} + \alpha_4 F_k^2 \qquad (4.5)$$

Первый член в этой формуле определяется потерями на колебание стенок, второй - потерями на вязкое сопротивление, третий - потерями на излучение. В дополнение к потерям акустических колебаний, необходимо учитывать и динамические потери, пропорциональные скорости воздушного потока *v* в речевом тракта.

Более точное вычисление коэффициента затухания $\delta_k = \Delta F_k / \pi$ может быть выполнено по формуле, полученной путем приравнивания изменения кинетической

энергии на периоде основного тона работе сил сопротивления в предположении достаточно малых потерь

$$\delta_k = \frac{\omega_k I_k}{4\omega_k + I_k}, \tag{4.6}$$

$$I_k = \frac{2\int_0^l QA\psi_k^2 dx}{\rho_0 \int_0^l A\psi_k^2 dx}, \tag{4.7}$$

$Q(x,t)$ - функция распределенных потерь, ψ_k - k-я собственная функция колебательной скорости, найденная из решения волнового уравнения без потерь.

Голосовой источник

Акустические колебания в речевом тракте вызываются сигналом голосового возбуждения, пропорциональным $G(t)=dW/dt$, W - объемная скорость воздушного потока через голосовую щель. Объемная скорость $W = vA_{vs}$ определяется решением уравнения

$$\rho_0 h_{vs} \frac{d(vA_{vs})}{dt} + k_{fr} vA_{vs} + \frac{\rho_0 c_{vs}}{2} v^2 A_{vs} = \Delta P A_{vs} \tag{4.8}$$

где h_{vs} - глубина голосовой щели, A_{vs} - площадь голосовой щели, $A_{vs} = A_{vs}(v)$, k_{fr} - коэффициент вязкого трения, v - скорость воздушного потока, $c_{vs} \approx 1{,}375$ - коэффициент динамического сопротивления, ΔP - перепад давления на голосовой щели, ρ_0 - плотность воздуха. Это уравнение имеет аналитическое решение при постоянной площади голосовой щели $A_{vs} = const$. Решение для переменной площади находится численным методом. Движение голосовых складок и, соответственно, площадь голосовой щели определяется упругостью складок и скоростью воздушного потока в щели. Перепад давления на голосовой щели ΔP заставляет голосовые складки расходиться. Движущийся поток создает силу, аналогичную подъемной силе крыла самолета, и эта сила заставляет складки сближаться. Таким образом возникают автоколебания складок, частота которых зависит от упругости складок и перепада давления.

На Рис. 4.4 показаны форма импульса объемной скорости и производная по времени от этого импульса, пропорционально которой и происходит возбуждение акустических колебаний в речевом тракте.

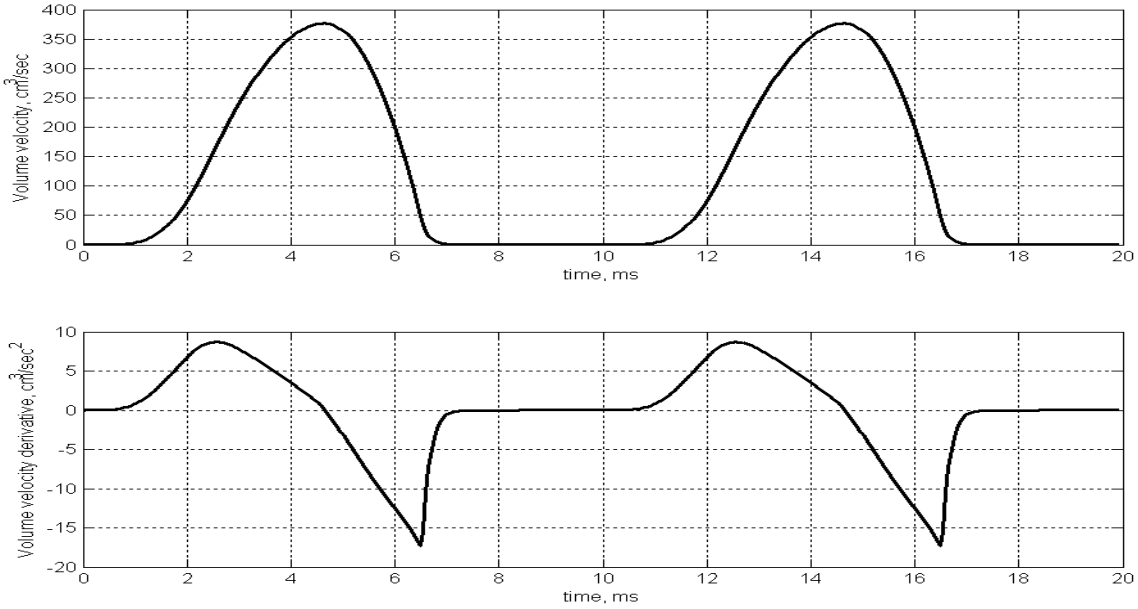


Рис. 4.4. Объемная скорость потока и ее производная (голосовой источник).

Турбулентный источник

В местах резкого расширения речевого тракта поток воздуха завихряется, порождая турбулентный источник давления. Такой источник возникает на выходе из голосовой щели, а глухие и звонкие фрикативные звуки /*s, sh, f, q, p, h,* z, *zh*/ образуются с участием турбулентного источника. На Рис. 5 показаны три фазы сближение голосовых складок от состояния дыхания до положения на глухом взрывном /*p*/, полученные с помощью рентгенографии. Видно, что имеется два резких расширения – над голосовыми складками и над ложными голосовыми складками. В этих местах происходит турбулизация воздушного потока и формируется шумовой источник возбуждения.

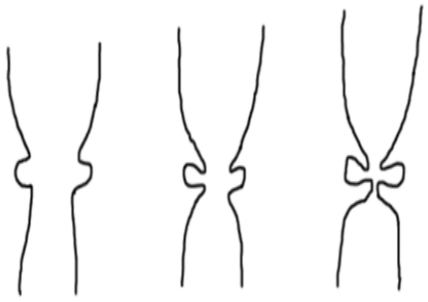


Рис. 4.5. Разрез гортани по голосовым складкам во фронтальной плоскости

Турбулизация потока возникает, если в каком-то месте число Рейнольдса Re,

$$\mathrm{Re} = \frac{\rho_0 \, vh}{\mu} \tag{4.9}$$

превышает критическое значение $\boldsymbol{Re}_{\mathrm{crit}}$, $\mathrm{Re} > \boldsymbol{Re}_{\mathrm{crit}}$. Здесь ρ_0 - плотность воздуха, v - линейная скорость воздушного потока, h - характерный геометрический размер поперечного сечения, μ - вязкость воздуха ($\mu = 1.86 \cdot 10^{-4}$ $g / cm\sec$). Критическое число

Рейнольдса в каждом месте речевого тракта и характеристики турбулентного шума зависят от его формы (в среднем $\boldsymbol{Re}_{\text{crit}} \approx 1800$).

Амплитуда, с которой возбуждается каждый резонанс под воздействием распределенного источника *f(x,t)* вычисляется как

$$a_k(t) = \frac{2\int_0^l f(x,t)A(x,t)\psi_k(x)dx}{\rho_0 l\int_0^l A(x,t)\psi_k^2(x)dx} \tag{4.10}$$

где ψ_k - *k*-я собственная функция акустического давления в речевом тракте. Если источник возбуждения сосредоточен в одной точке с координатой x_0, то

$$a_k(t) = \frac{2f(x_0,t)A(x_0,t)\psi_k(x_0)}{\rho_0 l\int_0^l A(x,t)\psi_k^2(x)dx} \tag{4.11}$$

Отсюда видно, что если источник возбуждения находится в области нуля какой-то собственной функции, то соответствующий резонанс не возбуждается. Это описывает свойства фрикативных звуков типа /*s, sh*, z, *zh*, *q, p*/, у которых подавлены низкочастотные компоненты спектра. Это же относится и к свойствам импульсного источника возбуждения, возникающего при взрыве смычки.

Импульсный источник

Импульсный источник возникает при раскрытии смычки в речевом тракте. Накопленное за время смычки давление в речевом тракте быстро падает, и амплитуда импульсного источника пропорциональна скорости этого падения .

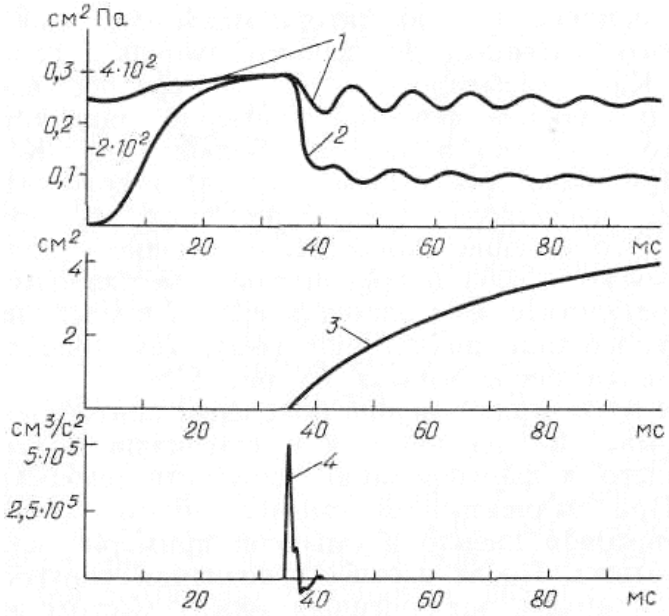


Рис. 4.6. Импульсный источник: 1— площадь голосовой щели, 2 - подсвязочное давление, 3 - площадь сужения в речевом тракте, 4 - производная от объемной скорости воздушного потока.

Лекция 5

Восприятие речи

Строение слухового аппарата

Наружное, среднее и внутреннее ухо.

Наружное ухо включает в себя ушные раковины, звуковой канал и барабанную перепонку.

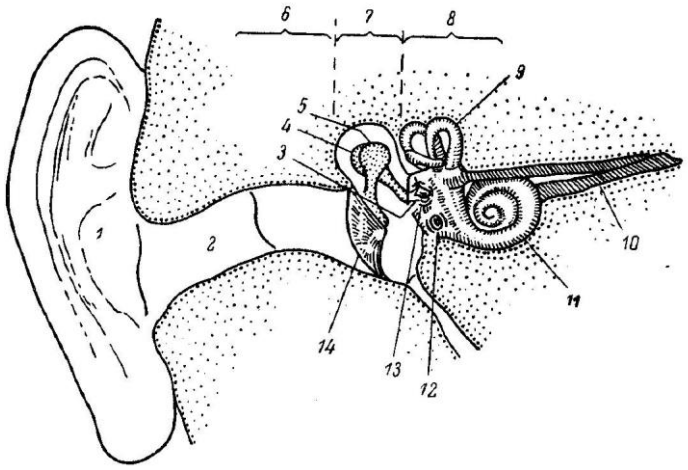
Среднее ухо содержит систему слуховых косточек.

Внутреннее ухо содержит улитку.

Наружное ухо оканчивается барабанной перепонкой диаметром 9 – 11 мм и толщиной 0.1 мм. К центру перепонки присоединен мускул, управляющей степенью ее натяжения. Наименьшее сопротивление перепонки на частотах 800 – 900 Гц.

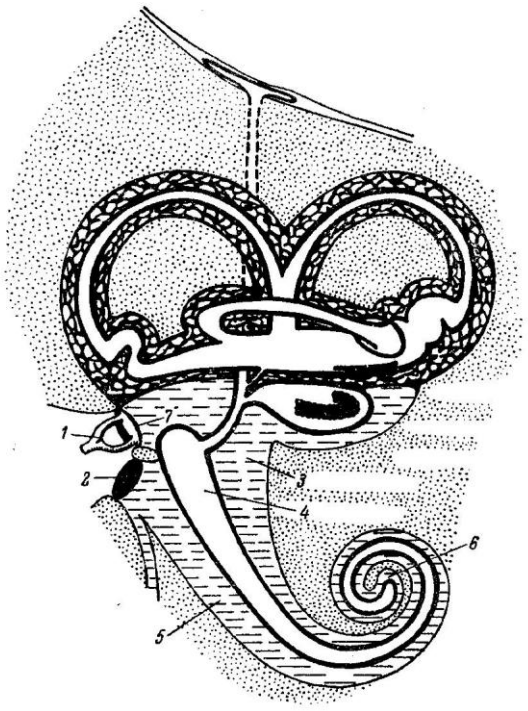
К барабанной перепонке прикреплена косточка – "молоточек". Она связана с другой косточкой – "наковальней", а та, в свою очередь – с косточкой, называемой "стремячком". Эта система слуховых косточек усиливает звуковое давление в 50 – 60 раз. "Стремячко" передает звуковые колебания овальному окну улитки.

Улитка представляет собой конус, свернутый у человека на 2.75 оборота.



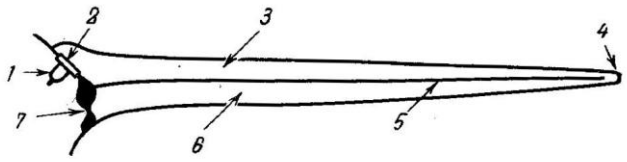
Схематическое изображение уха человека.

1 — ушная раковина; *2* — наружный слуховой проход; *3*, *4*, *5* — слуховые косточки, соответственно стремечко, наковаленка и молоточек; *6* — наружное ухо; *7* — среднее ухо; *8* — внутреннее ухо; *9* — вестибулярный аппарат; *10* — слуховой нерв; *11* — улитка; *12* — круглое окно; *13* — овальное окно; *14* — барабанная перепонка.



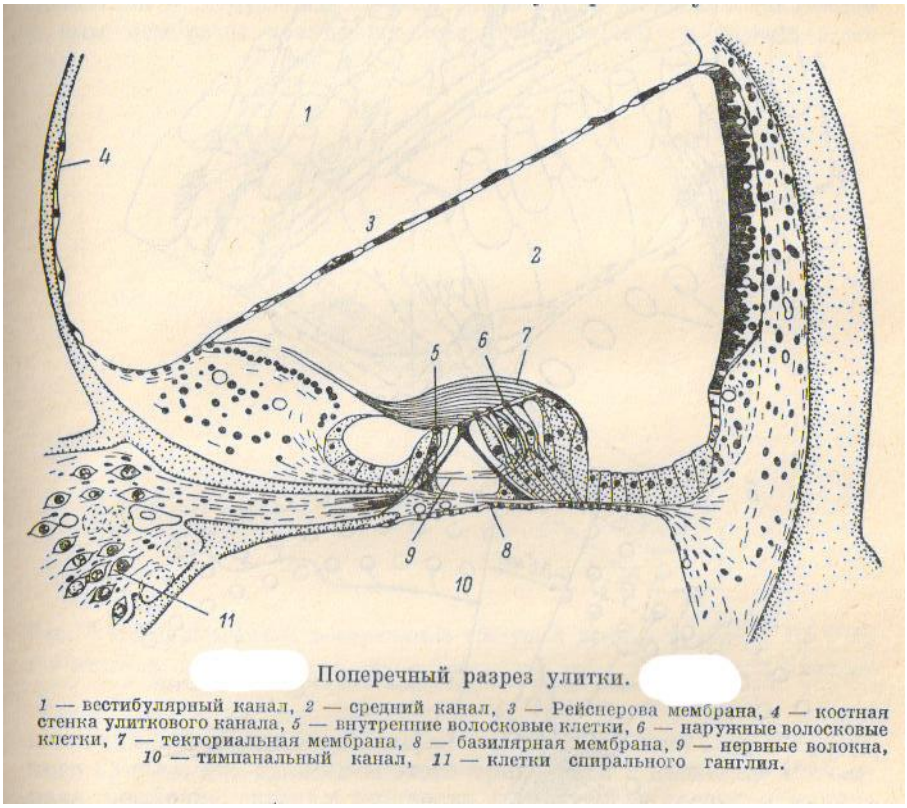
Схематическое изображение строения внутреннего уха млекопитающих.

1 — стремечко, *2* — круглое окно, *3* — вестибулярный канал, *4* — средний, или кохлеарный, канал, *5* — тимпанальный канал, *6* — геликотрема, *7* — овальное окно.



Упрощенная схема развернутой улитки:

1 — стремечко, *2* — овальное окно, *3* — преддверная лестница, *4* — геликотрема, *5* — перегородка улитки, *6* — барабанная лестница, *7* — круглое окно



Поперечный разрез улитки.

1 — вестибулярный канал, *2* — средний канал, *3* — Рейсснерова мембрана, *4* — костная стенка улиткового канала, *5* — внутренние волосковые клетки, *6* — наружные волосковые клетки, *7* — текториальная мембрана, *8* — базилярная мембрана, *9* — нервные волокна, *10* — тимпанальный канал, *11* — клетки спирального ганглия.

Длина базилярной мембраны – около 30 мм, причем у основания она уже (0.1 мм), чем у вершины улитки (0.5 мм).

Базилярная мембрана состоит примерно из 24000 слабо связанных поперечных волокон.

Кортиев орган содержит 5 рядов волосковых клеток. Внутренний ряд содержит около 3500 клеток, и внешние – по 5000 клеток. Считается, что только внутренние волосковые клетки передают информацию о колебаниях базилярной мембраны, а наружные клетки включены в цепь положительной обратной связи, увеличивающей амплитуду колебаний базилярной мембраны в области пучностей.

Колебания овального окна передаются жидкости, в которой над и под базилярной мембраной возникают вихри, положение которых связано с частотой колебаний (принцип "места"). Кроме того, информация о частоте звука содержится и в частоте колебаний в области пучности.

Уравнение колебаний базилярной мембраны

Простейший вид волнового уравнения колебаний базилярной мембраны:

$$\frac{\partial^2 P}{\partial x^2} = -\left(\frac{1}{S_1} + \frac{1}{S_2}\right)\frac{\rho}{Z}\frac{\partial P}{\partial t} - \left(\frac{1}{R_1} + \frac{1}{R_2}\right)\frac{P}{Z}$$

где P – звуковое давление в канале, S_1 – площадь поперечного сечения вестибулярного канала, S_2 - площадь поперечного сечения тимпанального канала, R_1, R_2 – трение в жидкостях этих каналов, ρ – плотность жидкости, Z – полный импеданс, $Z=P/V$, V – скорость изменения объема вестибулярного канала.

$$Z = r + j\left(\omega m - \frac{1}{\omega c}\right),$$

m, r,c – погонные масса, сопротивление и упругость базилярной мембраны.

Спектрально-временные характеристики базилярной мембраны могут быть представлены гребенкой так называемых гамма-тон фильтров:

$$g_i(t) = a_i t^{n-1} e^{-2\pi b_i t} \cos(2\pi f_i + \psi_i),$$

где *n* – порядок фильтра g_i - импульсный отклик *i*-го фильтра, f_i - центральная частота *i*-го фильтра, b_i - ширина полосы *i*-го фильтра. Огибающая отклика пропорциональна гамма-распределению.

Психофизические характеристики периферического отдела

Полоса частот, воспринимаемых молодым человеком – около 20 кГц, и уменьшается с возрастом. Чувствительность слуха зависит от частоты, и имеет максимум в области полосы частот речевых сигналов – 300 – 5000 Гц.

Закон Вебера-Фехнера: ощущение *A* пропорционально логарифму раздражения $\overline{A}$: $A = c\ln(\overline{A}/A_0)$, где A_0 – порог восприятия, *c* – некоторая константа. Поэтому амплитуды спектральных компонент речевого сигнала вычисляются в децибелах: $20\log_{10}$A. Порог восприятия по амплитуде составляет около 1 дБ.
Отношение сигнал/шум на выходе микрофона обычно составляет 60 дБ. Это означает, амплитуды спектральных компонент речевого сигнала нужно квантовать не более чем 64 уровня.
Порог чувствительности к положению пика форманты – от 1.5% на низких частотах, до 15% - на высоких.

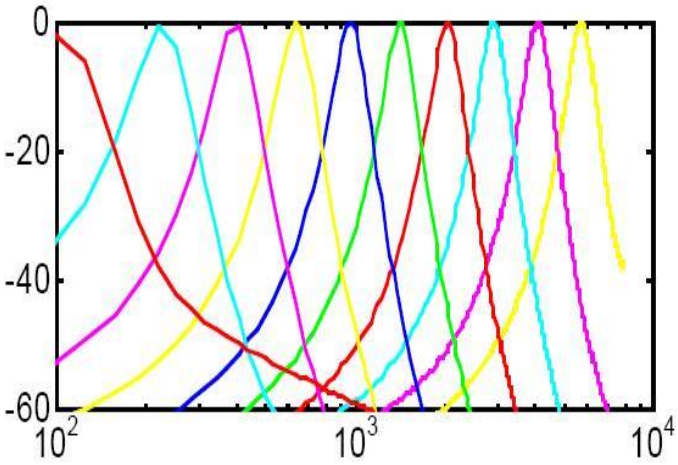
Критическая полоса – это такая полоса частотного диапазона речи, которая воспринимается как одно целое, и может быть заменена эквивалентным тоном.
Субъективное восприятие высоты тона описывается в системе *барк* или *мел*. Разность между двумя частотами, равная критической полосе, равна 1 *барк*.

$$M(f) = 1125 \cdot \ln(1 + f/700)$$

где f – частота в герцах, M – частота в мелах.

$$B = 13arctg(0.00076f) + 3.5arctg\left(\frac{f}{7500}\right)$$

B – частота в барках.
Совместное действие механики колебаний базилярной мембраны и частотных характеристик рецепторов и нейронов создает весовую функцию при анализе спектра речевого сигнала. Эта функция имеет крутой склон в сторону высоких частот, и более пологий склон в сторону низких частот.



Латеральное торможение приводит к обострению пиков в спектре сигнала.

Детекторы спектрально-временных неоднородностей

Речь - код, в котором передаваемая информация защищена, в том числе и с помощью разных видов **модуляции** - частотной, амплитудной, временной.

В слуховых системах млекопитающих и земноводных найдены нейроны, избирательно реагирующие на включение или выключение стимула. Обнаружены структуры, специализирующихся на обнаружении амплитудных модуляций, причем некоторые нейроны обладают порогом по скорости изменения огибающей звукового сигнала. Нейроны слуховой системы также реагируют и на частотные модуляции в сигнале, причем некоторые нейроны избирательно реагируют на возрастание или уменьшение частоты.

Описаны эффекты временной маскировки, т.е. зависимость восприятия искусственного стимула или сегмента речи от предыдущих и последующих по времени сигналов.

Во временной области найдено несколько постоянных времени, с которыми происходит сглаживание сигнала: 2 мс, 10 мс, 50 – 100 мс. Адаптация к изменению уровня громкости происходит с постоянными времени 200 – 300 мс.

В частотной области существует так называемый эффект латерального торможения, в результате которого обостряются пики огибающей спектра.

Аудио-визуальное восприятие речи.

Наблюдение за артикуляционной мимикой помогает лучшему восприятию речи в шумах, на фоне других разговоров и пониманию иностранной речи.

Эффект MacGurk

Взаимодействие аудио и визуального восприятия обладает и негативными свойствами. Просмотр видео записи с произнесением звукосочетания /гага/ на фоне звукового сопровождения с звукосочетанием /баба/ воспринимается как /дада/.

Активность слуховой и моторной коры при помехах.

При восприятии речи на фоне шумов активизируется не только слуховая, но и моторная зона коры головного мозга.

Факторы, влияющие на восприятие речи
Сложность задачи текущей деятельности
Социальный опыт и психолингвистический тип слушателя
Объем кратковременной памяти
Тренированность к особенностям речи диктора
Внешние условия – громкость, тип и уровень помех, реверберация помещения
Чем выше уровень помех, тем больше восприятие определяются частотой встречаемости слов.

Структура речевого сообщения – активные или пассивные грамматические формы, истинное или ложное высказывание, длительность сообщения, предсказуемость
В изолированном произнесении лучше всего распознаются существительные, а хуже всего – наречия.
В слитном произнесении лучше всего распознаются глаголы и наречия.

Оценка разборчивости и натуральности речи.

Разборчивость – это доля правильно распознанных элементов речи
Фонемная, слоговая, словесная и фразовая разборчивость

Уровень образования и жизненный опыт влияют на восприятие речевых элементов.
Слоговая разборчивость важна в диалоговой речи, а фонемная – в контекстной (описания, инструкции).

Тесты с открытым и закрытым выбором

Минимальные пары слов
Рифмованные тесты СГС
Тесты ГСГ
Артикуляционные таблицы
Тест "истина"/"ложь"

Тренировка аудиторов

Оценка натуральности: парные сравнения

Задержка восприятия. Тесты "истина"/"ложь"

Матрицы переходов.

Структура согласных звуков русской речи устанавливается в экспериментах по восприятию слогов /гласный-согласный-гласный/ на фоне белого шума .

Матрица переходов субъективного распознавания фонем, усредненная для отношений сигнал/шум -8 дБ, -4 дБ, 0 дБ, +4 дБ, +8 дБ демонстрирует группировку звуков по способу образования и источнику возбуждения:

/*b*, *d*, *g*/ - звонкие взрывные

/*z*, *zh*, *v*/ - звонкие фрикативные

/*m*, *n*/ - назальные

/*p*, *t*, *k*/ - глухие взрывные

/*s*, *sh*, *f*, *h*/ - глухие фрикативные

	b	d	g	ž	z	v	l	m	n	p	t	k	š	s	f	x	j	-
b	**84**	4	10			1												
d	4	**92**	3															
g	8	4	**80**		5													
ž			6	**49**	41	1											1	1
z		5	10	8	**70**	2	2										1	1
v	3		20	1	4	**59**	9										1	
l						1	**98**	1										
m							2	**90**	8									
n							1	6	**92**								1	
p										**57**	5	12		2	1	6		1
t										2	**77**	6		9	3	1		1
k										4	8	**73**		4	6	4		
š											11	4	**38**	37	5	3		2
s										1	18	4	1	**66**	4	2		1
f										15	3	15		6	**4**	1		1

Лекция 6

Теория речевого сигнала

Методы анализа речевого сигнала

Дискретизация и квантование

Если в спектре речевого сигнала необходимо сохранить частоты вплоть до F_{max}, то частота дискретизации, согласно теореме Котельникова-Найквиста, должна быть не меньше $F_s = 2F_{max}$. В системах передачи речи обычно используется частота дискретизации 8 кГц, а в системах распознавания – 16 кГц.

Равномерное квантование на 16 бит обеспечивает представление амплитуды сигнала с $2^{16} = 32000$ уровнями, достаточными для анализа сигнала с необходимой точностью. При таком квантовании амплитудный диапазон речи составляет 96 дБ (увеличение амплитуды вдвое, т.е. на 1 бит, соответствует 6 дБ).

Текущий спектр

В процессе артикуляции меняется форма речевого сигнала, а с ней и частотные характеристики речевого сигнала. Поэтому адекватным методом анализа является частотно-временной анализ сигнала.

Наиболее распространено преобразование Фурье на конечном интервале, так называемый текущий спектр

$$S_n(e^{j\omega}) = \sum_{m=-\infty}^{\infty} w(n-m)s(m)e^{-j\omega m}$$

где ***s*** – речевой сигнал, ω – круговая частота, ***w*** - весовая функция (окно). При частоте дискретизации 16 кГц обычно используется окно на 256 отсчетов, т.е. ***T*** = 16 мс. Интервалы анализа могут перекрываться.

Весовые функции

Используются разные виды окон, каждое из которых обладает разным уровнем вычислительных шумов, и по-разному представляет спектр сигнала.

Если необходимо сохранить периодичность колебаний, то используется прямоугольное окно, ***W***=1. Наиболее распространено окно Хемминга:

$$W_{ham}(n) = 0.54 - 0.46\cos(2\pi n / N)$$

N – ширина окна в отсчетах,
или его модификация, называемая окном Хеннинга:

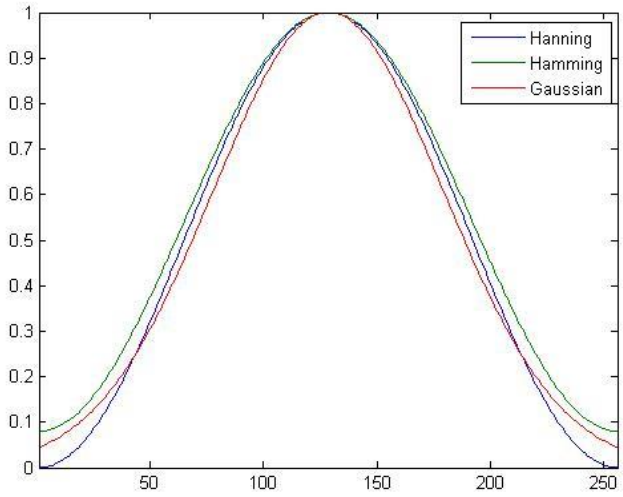
$$W_{han}(n) = 0.5(1 - \cos(2\pi n / N))$$

Окно Лапласа или Гаусса, так называемая колокольная функция,

$$W_{gaus}(n) = \exp(-0.5(2\alpha n / N))^2$$

обеспечивает в некотором смысле наилучший компромисс по разрешающей способности как во временной, так и в частотной области.

Применяются и другие виды окон, например, треугольное (окно Блэкмана), окно Кайзера-Бесселя.



В слуховой системе человека, как и в других сенсорных системах, отклик нелинейно связан с входным стимулом.

Кепстр (Cepstrum)

Текущий спектр речевого сигнала подвержен влиянию источника голосового возбуждения, в том числе частоты основного тона. Для компенсации этого влияния в большинстве современных систем автоматического распознавания речи для описания формы спектра используется его разложение в ряд по коэффициентам так называемого кепстрального преобразования:

$$\log\left|S(j\omega,t)\right|^2 = \sum_n c_n(t)e^{jn\Omega\omega}$$

где коэффициенты c_n определяются через обратное преобразование Фурье от логарифма энергетического спектра

$$c_n = \frac{1}{\Theta}\int_0^{\Theta} \log\left|S(j\omega,t)\right|^2 e^{-jn\Omega\omega}d\omega$$

$\Omega = 2\pi/\Theta$, Θ - верхняя частота в спектре речевого сигнала. Число кепстральных коэффициентов n зависит от требуемого сглаживания спектра, и находится в пределах от 20 до 40.

Экспериментально было установлено, что наилучшие результаты при автоматическом распознавании достигаются, если спектр речевого сигнала представлен в шкале *мел*. Такое кепстральное преобразование получило название MFCC – Mel Frequency Cepstral Coefficients.

Используются также первая и вторая производная по времени от кепстральных коэффициентов.

Линейное предсказание (LPC – Linear Prediction Coefficients)

Если некоторый сигнал характеризуется конечным интервалом корреляции, то каждый его отсчет может быть представлен в виде взвешенной суммы предыдущих отсчетов

$$s_i = \sum_{k=1}^{n} a_k s_{i-k} + G u_i$$

где s_i -й отсчет сигнала, u_i - сигнал возбуждения, $\boldsymbol{G}$ - коэффициент усиления. Число коэффициентов n определяется по эмпирическому правилу $n = F_s + 2$, где F_s –частота дискретизации сигнала в кГц.

Коэффициенты линейного предсказания a_k вычисляются путем минимизации среднеквадратичной ошибки

$$\varepsilon_i = \sum_{m} (s_{i+m} - \sum_{k=1}^{n} a_k s_{i+m-k})^2$$

По коэффициентам линейного предсказания может быть вычислена передаточная функция и сглаженный спектр речевого сигнала:

$$S(j\omega) = \frac{G}{1 - \sum_{k=1}^{n} a_k e^{-j\omega k}}$$

.

Формантный анализ

Пики в спектре речевого сигнала называют ***формантами*.** Частоты этих формант связаны с резонансными частотами речевого тракта.

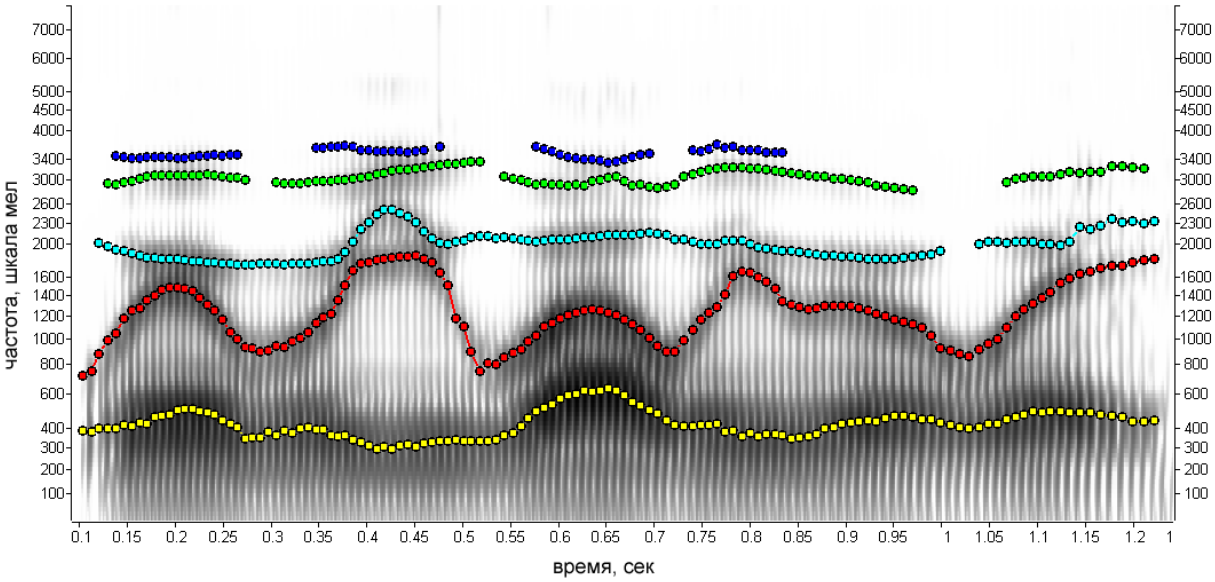
Существует множество алгоритмов определения формантных частот, использующих несколько основных методов.

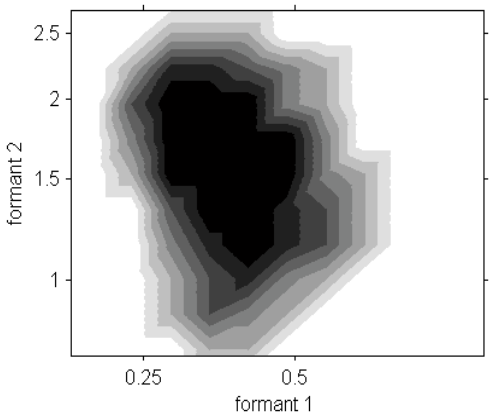
1.Частоты максимумов исходного спектра или спектра, полученного обратным преобразованием от кепстра.

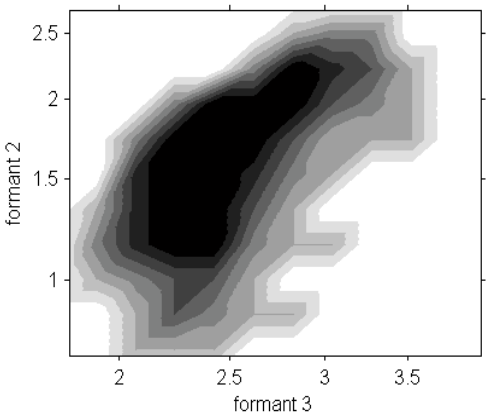
2. Полюса передаточной функции речевого тракта, вычисленные с помощью линейного предсаказания.

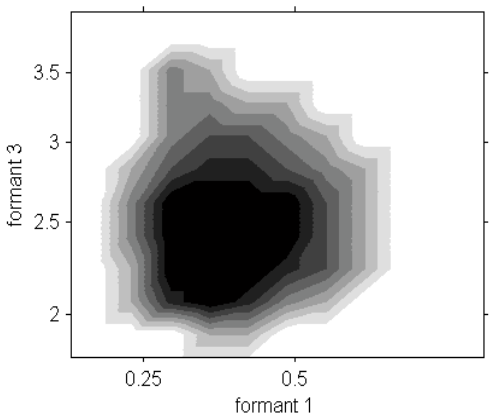
3. Оценка мгновенных частот методом анализа интервалов между пересечениями нуля сигналов в разных частотных областях.

4.Оценка частотных компонент в моделях слухового анализа.


частота, шкала мел
7000
6000
5000
4500
4000
3400
3000
2600
2300
2000
1600
1400
1200
1000
800
600
400
300
200
100
0.1 0.15 0.2 0.25 0.3 0.35 0.4 0.45 0.5 0.55 0.6 0.65 0.7 0.75 0.8 0.85 0.9 0.95 1 1.05 1.1 1.15 1.2 1
время, сек


formant 2
2.5
2
1.5
1
0.25
0.5
formant 1


formant 2
2.5
2
1.5
1
2
2.5
3
3.5
formant 3


formant 3
3.5
3
2.5
2
0.25
0.5
formant 1

Анализ основного тона

Основной тон – это частота голосового возбуждения, т.е. частота колебаний голосовых складок.

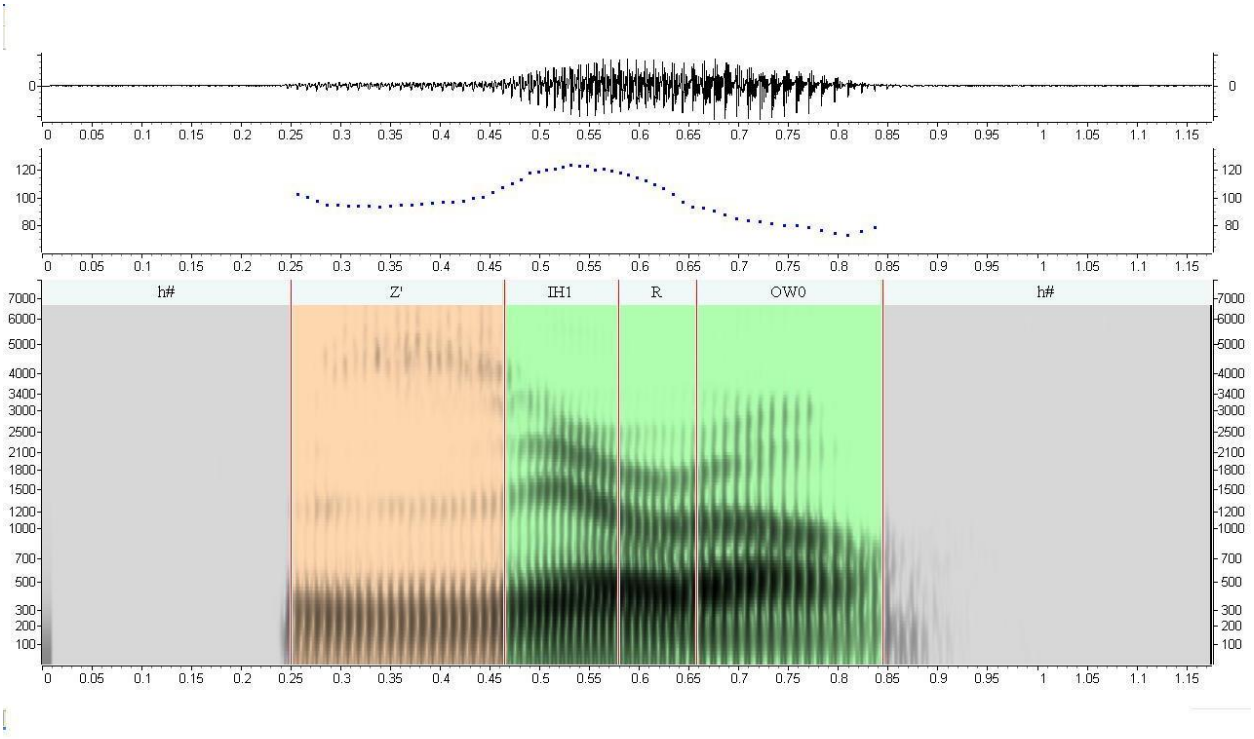
Анализ основного тона выполняется во временной или частотно-временной области.

Пики огибающей речевого сигнала.

Положение пиков автокорреляционной функции на оси времени.

Частота следования пиков сигнала-остатка в методе линейного предсказания.

Диапазон частот основного тона для мужских голосов – 80 – 250 Гц со средней частотой $F_0 = 129$ Гц. Диапазон частот основного тона для женских голосов – 120 – 350 Гц со средней частотой $F_0 = 250$ Гц.



Частота основного тона и сонограмма слова “ZERO”, произнесенного мужским голосом.

Детекторы спектрально-временных неоднородностей

Речь - код, в котором передаваемая информация защищена, в том числе и с помощью разных видов **модуляции** - частотной, амплитудной, временной.

В слуховых системах млекопитающих и земноводных найдены нейроны, избирательно реагирующие на включение или выключение стимула. Обнаружены структуры, специализирующихся на обнаружении амплитудных модуляций, причем некоторые нейроны обладают порогом по скорости изменения огибающей звукового сигнала. Нейроны слуховой системы также реагируют и на частотные модуляции в сигнале, причем некоторые нейроны избирательно реагируют на возрастание или уменьшение частоты.

Описаны эффекты временной маскировки, т.е. зависимость восприятия искусственного стимула или сегмента речи от предыдущих и последующих по времени сигналов.

Во временной области найдено несколько постоянных времени, с которыми происходит сглаживание сигнала: 2 мс, 10 мс, 50 – 100 мс. Адаптация к изменению уровня громкости происходит с постоянными времени 200 – 300 мс.

Модель неспецифических детекторов

Оператор

$$A(\omega,t)=\lg\frac{S(\omega+\Delta\Omega,\theta_1,t\pm\Delta T_1,\tau_1)+C}{S(\omega-\Delta\Omega,\theta_2,t\mp\Delta T_2,\tau_2)+C}, \qquad (6.1)$$

моделирует многие свойства детекторов, найденных в слуховой системе.

При τ_1=0, τ_2=0, $\Delta\Omega$=0, θ_1=0, θ_2=0, и $\Delta T_1\to 0$, $\Delta T_2\to 0$, оператор (1) вычисляет логарифмическую производную по времени:

$$A(\omega,t)=\lg S(\omega,t+\delta t)-\lg S(\omega,t-\delta t)\underset{\delta t\to 0}{=}2\delta t\left[\lg S(\omega,t)\right]'$$

При θ_1=0, θ_2=0, τ_1=0, τ_2=0, ΔT_1=0, ΔT_2=0, $\Delta\Omega\to 0$, в пределе $A(\omega,t)$ оценивает мгновенную производную спектра, но не по времени, а по частоте:

$$A(\omega,t)=\lg S(\omega+\Delta\Omega,t)-\lg S(\omega-\Delta\Omega,t)\underset{\Delta\Omega\to 0}{=}2\Delta\Omega\left[\lg S(\omega,t)\right]'$$

или

$$A(\omega,t)=2\Delta\Omega\frac{S'(\omega,t)}{S(\omega,t)}$$

Параметры θ_1 и θ_2 нормируют спектр сигнала к скользящему среднему:

$$\overline{S}(\omega,t)=\frac{\theta_2\int\limits_{\omega-\theta_1}^{\omega+\theta_1}S(\omega,t)}{\theta_1\int\limits_{\omega-\theta_2}^{\omega+\theta_2}S(\omega,t)d\omega},$$

Если среднее на интервале $2\theta_1$ равно среднему на интервале $2\theta_2$, то

$$\overline{A}(\omega,t)=0$$

Если потребовать, чтобы $\overline{S}(\omega,t)$=1 всякий раз, когда

$$\theta_1\int\limits_{\omega-\theta_2}^{\omega+\theta_2}S(\omega,t)d\omega>\theta_2\int\limits_{\omega-\theta_1}^{\omega+\theta_1}S(\omega,t),$$

то такие участки спектра будут "невидимы", поскольку $\overline{A}(\omega,t)$=0. Тем самым на спектральном разрезе выделяются локальные максимумы, определенные на различных частотных интервалах.

В частном случае, при $\theta = 0, \quad \theta_2 = \Theta$, где Θ - верхняя граница частотного спектра сигнала, выполняется нормировка к общей энергии сигнала во всем частотном диапазоне в каждый момент времени:

$$\overline{S}(\omega,t) = \frac{\Theta S(\omega,t)}{\int_0^{\Theta} S(\omega,t)d\omega},$$

Тогда нормированный детектор есть

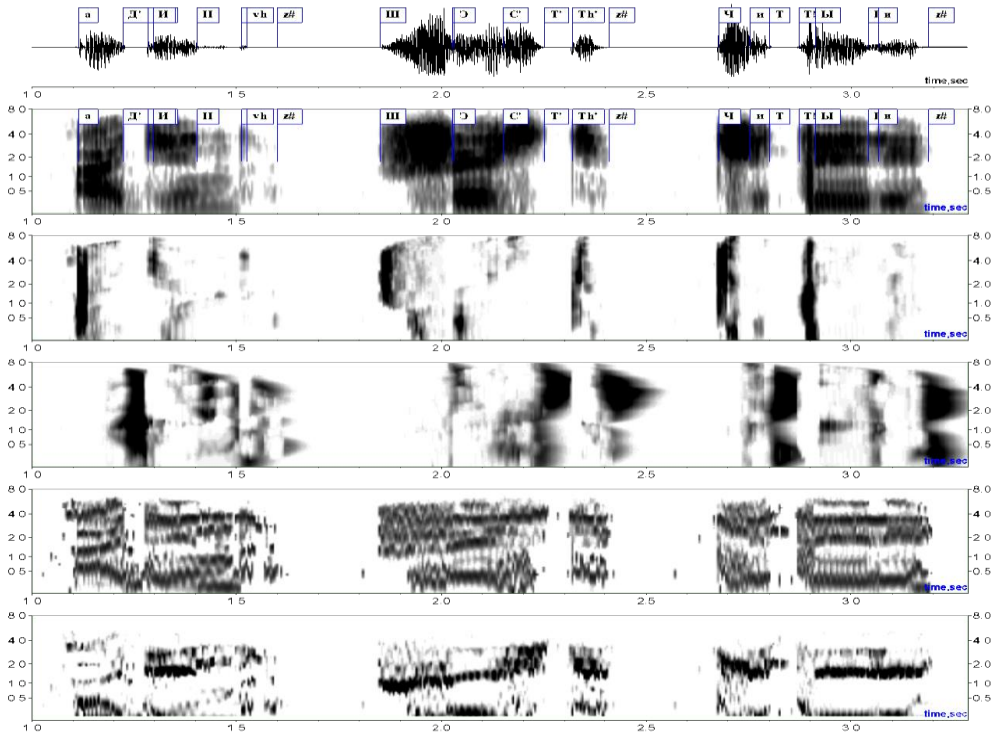
$$\overline{A}(\omega,t) = \lg \frac{\overline{S}(\omega+\Delta\Omega, t\pm\Delta T_1, \tau_1)}{\overline{S}(\omega-\Delta\Omega, \Theta, t\mp\Delta T_2, \tau_2)},$$

Пусть спектральный состав сигнала не меняется во времени, а меняется только амплитуда, т.е. *S(ω,t) = G(t)S(ω)*. Тогда нормированный спектр не меняется во времени:

$$\overline{S}(\omega,t) = \frac{\Theta G(t)S(\omega)}{\int_0^{\Theta} G(t)S(\omega)d\omega} = \overline{S}(\omega),$$

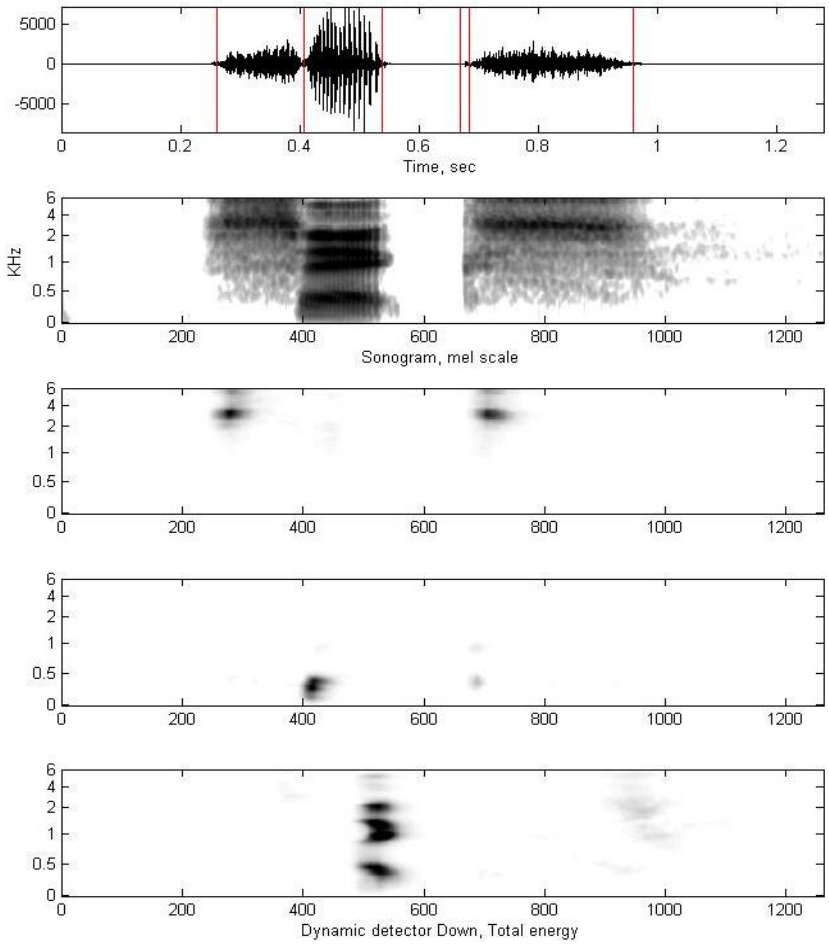
и при $\tau_1 = 0, \tau_2 = 0, \Delta\Omega = 0$

$$\overline{A}(\omega,t) = \lg \frac{\overline{S}(\omega, t\pm\Delta T_1, \tau_1)}{\overline{S}(\omega, t\mp\Delta T_2, \tau_2)} = 0$$



Детекторы артикуляторных событий

Первичные динамические детекторы реагируют на любое изменение спектрально-временных характеристик. На основе этих дектторов можно сформировать *детекторы артикуляторных событий*, которые реагируют только на переход из одного конкретного артикуляторного состояния в другое.



Слово /*six*/, мужской голос.

Отклики детекторов артикуляторных событий «Пауза - Фрикативный» и «Смычка - Фрикативный», отклики детектора перехода «Фрикативный - Гласный», и отклики детекторов переходов «Гласный - Смычка» и «Фрикативный - Пауза»,.

Элементы кодовой структуры речевого сигнала

Речевой код является нелинейным случайным кодом с каскадной структурой. На нижнем уровне элементами кода являются артикуляторные состояния, а на акустическом уровне элементами кода служит переход из одного артикуляторного состояние в другое.

Состоянием, т.е. длительным стационарным положениям артикуляторов, характеризуются все гласные и фрикативные звуки (в русском языке, это С, З, Ш, Ж, Ф, Ч и их мягкие варианты).

Смычные согласные – Б, П, Д, Т, Г, К, М, Н, характеризуются полным смыканием в каком-либо месте речевого тракта, называемом местом артикуляции, и, соответственно, нулевым значением площади поперечного сечения тракта. Это также состояния, которые могут поддерживаться в течение некоторого времени, но при этом на глухих смычных не происходит никакого излучения речевого сигнала, а на звонких смычных излучаемый сигнал несет очень мало информации о месте артикуляции.

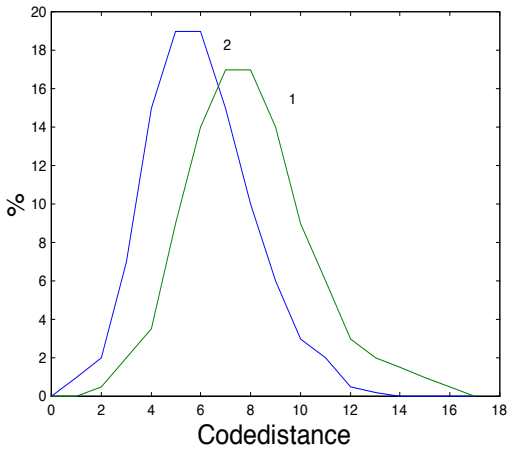
Восприятие смычных согласных, а также звуков Л, Р, В, осуществляется на основе переходных процессов между этими звуками и соседними гласными звуками, позволяя определить место артикуляции. В русском языке имеется три основных места артикуляции смычных звуков: на губах (Б, П, М), в передней части языка (Д, Т, Н) и в области небной занавески (Г, К).

Наличие или отсутствие голосового возбуждения кодирует звонкие и глухие согласные, турбулентный источник возбуждения кодирует фрикативные, а опущенная небная занавеска кодирует назальные звуки М и Н.

Следующий уровень речевого кода состоит из множества слогов, представляющих собой сочетания двух или трех звуков. Этот уровень обладает малой исправляющей способностью, поскольку его избыточность определяется ограничениями на сложность управления и энергетические затраты (разные в разных языках).

Спектры кодовых расстояний

Определим кодовое расстояние d между словами как число несовпадающих признаков. Тогда условие обнаружения ошибок кратности Q_{detect} есть $Q_{\text{detect}} \le d - 1$, а условие исправления ошибок кратности Q_{correct} есть $Q_{\text{correct}} \le (d-1)/2$.



Спектры кодовых расстояний для наиболее частых слов. Кодирование по фонемам (1), кодирование по признакам "голосовой источник, шумовой источник, назальный, гласный" (2).

Уровень слов обладает значительной исправляющей способностью, поскольку далеко не все возможные сочетания звуков в данном языке образуют осмысленные слова. Исправляющая способность на уровне фонем – более 99% одиночных ошибок, около 89% двойных ошибок, 57% тройных ошибок.

ПерНеходный → Переходный.

На фразовом уровне происходит коррекция ошибок восприятия, совершенных на нижних уровнях речевого кода, путем использования синтаксиса языка.

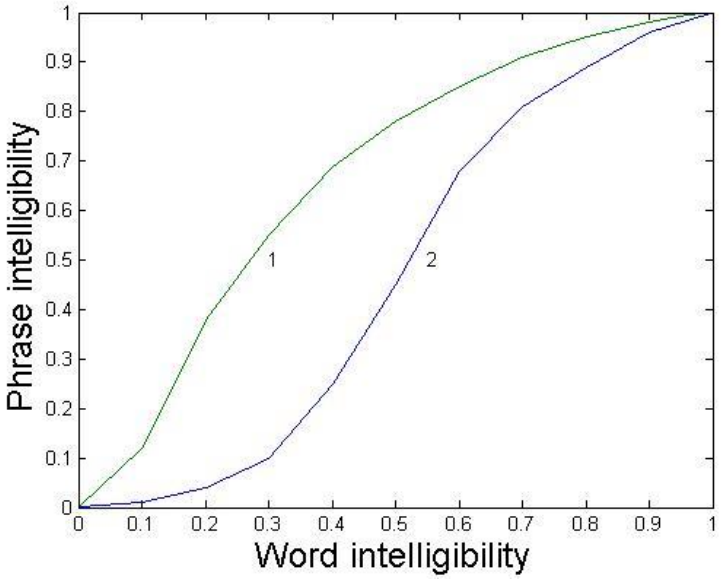
Девочка пришел домой → Девочка пришла домой.

Не все грамматически правильные фразы имеют какой-либо смысл, и это позволяет корректировать ошибки путем семантического анализа речевого высказывания.

Глокая куздра штеко будланула бокра и курдячит бокренка.
Веселая дверь скачет в небо.

На прагматическом уровне выполняется окончательная коррекция, поскольку не все семантически правильные высказывания могут иметь смысл в конкретной области или теме разговора.

В английском языке действуют более сильные ограничения, например, на порядок слов. Поэтому и зависимость фразовой разборчивости от словесной у английского языка лучше, чем у русского.



1 – разборчивость для английского языка, 2 – разборчивость для русского языка.

В английском языке существует множество слов, имеющих разный смысл при одинаковом произношении. Такие слова имеются и в русском языке. Иногда смысл слова определяется контекстом фразы.

Лист:

Возьмите чистый ЛИСТ бумаги.
Как ЛИСТ увядший падает на душу ...
Этот концерт написал ЛИСТ.

В других случаях необходимо знать "сценарий", в рамках которого появилась данная фраза.

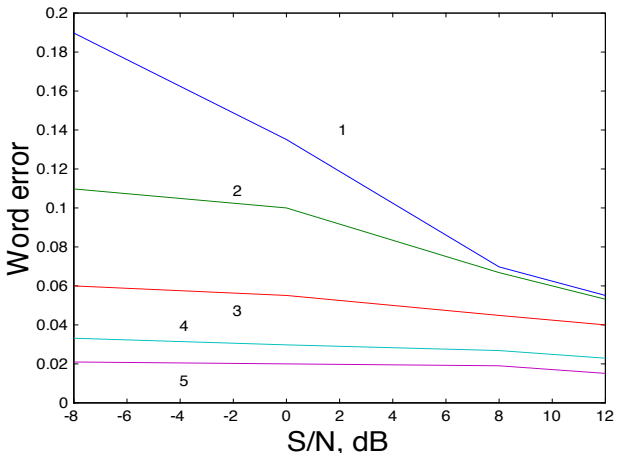
Ключ:

Я нашел КЛЮЧ (источник воды).
Я нашел КЛЮЧ (в своем кармане).
Я нашел КЛЮЧ (к расшифровке кода).

Потенциальная вероятность

Используя результаты психо-физических экспериментов по восприятию фонем при различных уровнях шума, можно оценить верхнюю и нижнюю границу вероятности правильного распознавания слов и сравнить ее с экспериментально полученными оценками в аудиторских испытаниях.

Оказалось, что при высоких отношениях сигнал/шум ***S/N*** вычисленные оценки и субъективная вероятность ошибки восприятия слова близки, тогда как при низких ***S/N*** потенциальная вероятность ошибки значительно ниже реально наблюдаемой. При отношении сигнал/шум = 0 дБ она в 5 раз лучше реально наблюдаемой надежности субъективного восприятия слов.



Субъективная словесная неразборчивость (1); верхняя граница ошибок при кодировании слов фонемами (2 - R=0.97, 3 - R=0.58); верхняя граница ошибок при кодировании слов независимыми признаками (4 - R=0.97, 5 - R=0.58);

Это свидетельствует о том, что на фонемной уровне алгоритм декодирования по фонемам у человека не использует полностью фонемную избыточность, возможно, в силу его сложности, предпочитая исправлять ошибки на более высоких уровнях, или просто переспрашивать.

Если кодировать все фонемы русского языка пятью признаками – Гласный, Смычка, Звонкий, Шумный, Назальный, то и в этом случае на уровне слов сохраняется довольно высокая избыточность, позволяющая корректировать ошибки. Эти признаки физически интерпретируемы и достаточно хорошо распознаются на акустическом уровне.

Полное описание фонем требует знания места артикуляции, которое для гласных и фрикативных определяется как координата наибольшего сужения в речевом тракте. Этот признак довольно плохо распознается на акустическом уровне, но он необходим в для успешного распознавания слов при высоких уровнях шума.

Префиксный код

Вероятность появление фонем и слов в речи не связана с их помехоустойчивостью, как это можно было бы ожидать, исходя из необходимости защиты передаваемой информации. Их распределение подчиняется закону Мандельброта

$$p(r) = \frac{A}{(r + B)^{\gamma}},$$

p - вероятность появления сообщения, r - ранг, A,B и γ - параметры языка. Ранг r = 1 приписывается наиболее часто встречающемуся слову, r = 2 – следующему по встречаемости, и т.д.

Этот закон был получен для кодов, сформированных из условия максимизации информации, передаваемой сообщением, при ограничении на стоимость сообщения.

Во всех исследованных языках частота встречаемости слов определяется их длиной в фонемах: чем короче слово, тем чаще оно встречается. Частота встречаемости фонем, по крайней мере в русском языке, определяется числом используемых артикуляторов и их массой.

Это означает, что частота появления слов или фонем определяется затратами на их генерирование.

Среди кодов, подчиняющихся закону Мандельброта, существуют так называемые неприводимые, или префиксные коды, у которых ни одно кодовое слов не является началом более длинного кодового слова. Для таких кодов доказано существование алгоритма сегментации слитной последовательности кодов без использования пауз между словами или специальных символов-разделителей.

Установлено, что в письменной речи не более 10% слов составляют начало других слов, причем в большинстве это союзы и предлоги. Таким образом, речевой код на уровне слов относится к классу префиксных кодов, что обеспечивает распознавание слов в слитном потоке речи.

Проблему распознавания слов или понимания речи можно представить с двух точек зрения.

Интерпретируя речевое сообщение как каскадный код, можно корректировать ошибки анализа речевого сигнала на фонемном, слоговом, лексическом, фразовом или семантическом уровне, используя ограничения на более высоких уровнях.

Процесс корректирования ошибок можно рассматривать и как общий случай решения некорректных обратных задач. На самых низких уровнях анализа речевого сигнала, например, при определении места артикуляции, обратная задача формулируется более традиционным образом. Она решается с помощью ограничений на механику и акустику речеобразования.

Лекция 7

Теория внутренней модели

Система управления артикуляцией

В системе управления артикуляцией существуют локальные и глобальные обратные связи.

Локальная обратная связь обеспечивает исполнение команд на изменение мышецной силы. В цепь обратной связи включены рецепторы - веретена мышц.

Внутренняя глобальная обратная связь обеспечивает координацию мышечных сокращений для достижения заданной формы речевого тракта. Форма и положение артикуляторов контролируются мышечными веретенами и тактильными рецепторами.

Внешняя обратная связь обеспечивает формирование заданных акустических характеристик речевого сигнала, которые вычисляются слуховой системой.

В глобальных обратных связях сигналы от тактильных рецепторов или, тем более, от слуховой системы, не могут быть заведены непосредственно на команды управления мышцами. Необходим какой-то модуль для согласования размерностей.

И внутренняя, и внешняя обратная связи требуют решения так называемой обратной задачи, т.е. вычисления входных сигналов по наблюдениям за выходными сигалами.

Возмущения артикуляции и восприятия

Функционирование системы управления артикуляцией исследуется в экспериментах с искусственным нарушением процессов артикуляции и восприятия, а также путем измерения электрической активности мозга.

Применяется ограничение движений нижней челюсти (bite-block) и губ (lips-block), изменение формы твердого неба с помощью специальной накладки. Динамическое вмешательство в движения артикуляторов осуществляется с помощью электро-механических устройств, изменения формы протеза твердого неба и электрической стимуляции мышц.

С помощью кодека изменяются текущие характеристики частоты основного тона или формантных частот речевого сигнала, который воспринимается диктором через наушники.

При задержке восприятия собственной речи возникает заикание (эффект Ломбарда).

В этих экспериментах обнаружено явление компенсации возмущений, причем задержка сопоставима с временем прохождения сигнала от рецепторов до центральной нервной системы и обратно. Это означает, что вычисление реакции происходит практически мгновенно, без пробных движений.

Важная информации о свойствах системы управления артикуляцией получается при наблюдении за патологией речи.
ларингэктомия,
удаленный язык,
паралич (парез) мышц нижней челюсти,
зубные протезы.

Модель тела.
Обучение родному и иностранному языку.

Ампутация, врожденное отсутствие конечностей.

Моторная теория восприятия речи.
Активность слуховой и моторной зоны коры при восприятии речи и наблюдением за речевой мимикой. Эффект МакГурка.

Речевые обратные задачи

Принципиальная некорректность речевых обратных задач.

Свойства некорректных обратных задач. Регуляризация.

Вариационный метод.

Вариационный метод требует использования математических моделей процессов речеобразования, и это совпадает с гипотезой существования таких моделей в системе управления артикуляцией. Эта модель задается в виде

$$A(x)=u,$$

где x - артикуляторные параметры, u - акустические параметры.

В методе Тихонова ищется приближенное решение обратной задачи путем минимизации функционала

$$M(z) = \alpha\Omega(z) + \rho^2(A_h z, u_\delta), \quad z \in Z$$

где A_h – оператор приближенной (с точностью h) математической модели, связывающей входные параметры инвертируемого процесса z и выходные параметры u_δ, измеренные с погрешностью δ. $\Omega(z)$ есть критерий оптимальности, $\alpha=\alpha(h,\delta)$ – параметр регуляризации. Величина $\rho(A_h z, u_\delta) = \|A_h z - u_\delta\|$ есть невязка между измеренными и вычисленными параметрами, а Z – данное множество ограничений. В нашем случае h и δ - погрешность в описании модели речеобразования и ошибки измерения акустических параметров.

Критерии оптимизации

Процесс минимизации состоит в поиске условного экстремума при наличии ограничений на значения артикуляторных и акустических параметров.

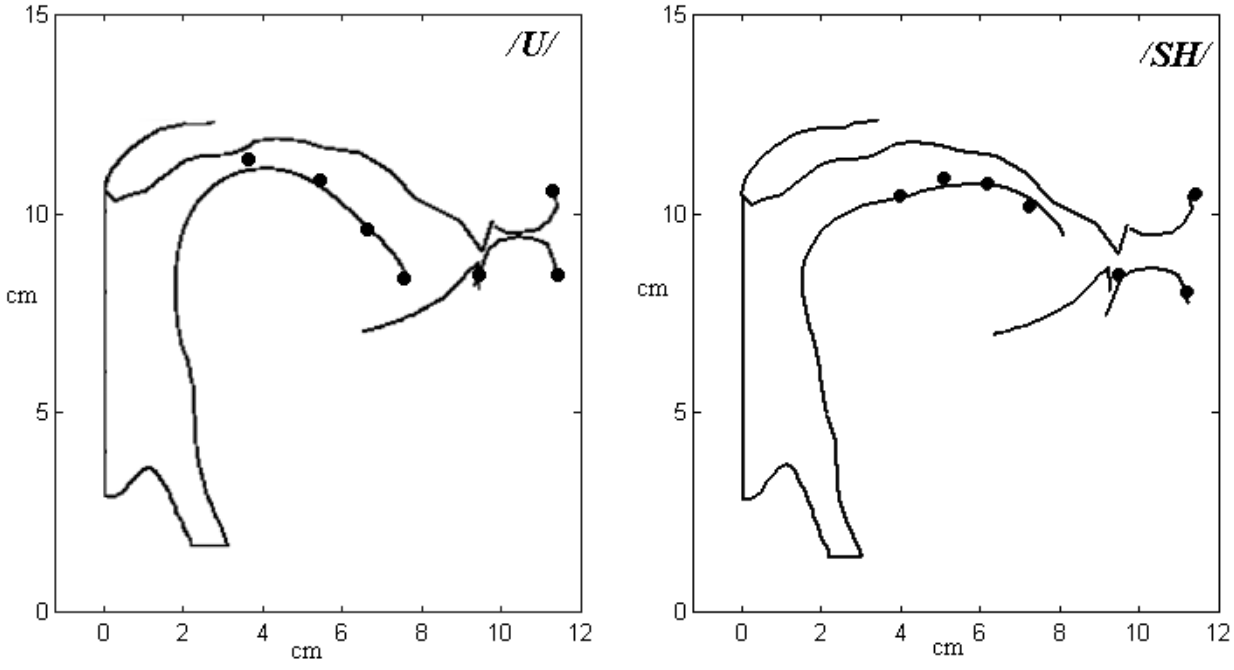
Критерий минимума работы артикуляторов оказался эффективным при решении обратных задач для стационарных сегментов гласных или фрикативных звуков.

$$\Omega_W = \frac{1}{2T}\sum_k \int_t^{t+T} c_k (x_k - x_k^{(o)})^2 d\tau$$

C_k – коэффициент упругого сопротивления движению артикулятора,

Погрешность определения координат точек на поверхности артикуляторных

органов сравнима с погрешностью их измерения.



При решении динамических задач необходимо использовать составной критерий $\Omega = \Omega_T + \Omega_w$, где

$$\Omega_W = \frac{1}{2T} \sum_k \int_t^{t+T} c_k (x_k - x_k^{(o)})^2 \, d\tau ,$$

$$\Omega_T = \frac{1}{2T} \sum_k \int_t^{t+T} (m_k x_k'')^2 \, d\tau ,$$

Здесь c_k – коэффициент упругого сопротивления движению артикулятора, m_k -масса артикулятора, $x_k^{(0)}$ - значение артикуляторного параметра в нейтральном состоянии. Эти критерии интерпретируются соответственно как средняя за время T суммарная работа упругих сил (Ω_W) и средний квадрат полной силы, приложенной к артикуляторам (Ω_T) .

Ограничения

Решение речевых обратных задач невозможно без использования ограничений. Внешние ограничения связаны со свойствами языка, а внутренние – с анатомией и ограничениями на мышечные силы.

Шесть видов ограничений:

1. Ограничение на силу f_n развиваемую n-й мышцей, $0 \le f_n \le f_{n\max}$, определяет максимальную скорость и ускорение артикулятора.
2. Значения артикуляторных параметров, включая массу, упругое и вязкое сопротивление, находятся в определенных пределах, $z_{j\min} \le z_j \le z_{j\max}$.

3. Задан класс преобразований вектора артикуляторных параметров $\mathbf{z} = z(z_1, z_2, ..., z_N)$, в площадь поперечного сечения речевого тракта $S(x,\mathbf{z})$, x – координата вдоль средней линии тракта.
4. Ограничение на форму речевого тракта.
   Для гласных: $S(x,z) \geq S_{\min}$, где $S_{\min}$ определяется условиями возникновения турбулентного шума.
   Для фрикативных $S_{\min} \leq S(x,z) \geq S_{\max}$, где S_{min} и $S_{\max}$ определяются условиями поддержания турбулентных шумов, т.е. $\mathrm{Re} \geq \mathrm{Re}_{crit}$.
   Для смычных: $S_{\min} = 0$ в месте артикуляции.
   Для назальных: опущенная небная занавеска, т.е. $S_{vp} > 0$, тогда как для всех остальных звуков $S_{vp} = 0$.
5. Акустические ограничения в виде $0 \leq \rho(u, u_\delta) \leq \Delta\rho$, где $\rho(u, u_\delta) = \|u - u_\delta\|$- невязка между измеренными акустическими параметрами u_δ и параметрами u, вычисленными моделью.
   Для звуков с формантной структурой оптимальным оказалось $\rho = \max|1 - F_k^* / F_k|$, k=1, 2, 3, где F_k – k-я формантная частота, измеренная в речевом сигнале, F_k^* - частота, вычисленная акустической моделью.
   Для фрикативных звуков

$$\rho = 1 - \frac{\int_0^l S(\omega) S^*(\omega) d\omega}{\left[\int_0^l S^2(\omega) d\omega \int_0^l S^{*2}(\omega) d\omega\right]^{1/2}}$$

   где $S(\omega)$ и $S^*(\omega)$ – измеренный и вычисленный спектр.
6. Сложность планирования и программирования нейромоторных команд также является ограничивающим фактором.

Кодовая книга

Задача минимизации функционала оказывается многоэкстремальной, и нет гарантий нахождения глобального минимума.

Для нахождения локального минимума, обеспечивающего приемлемую точность решения, необходимо повторять процесс оптимизации определенное количество раз, начиная с разных начальных условий. Начальные значения выбираются не произвольно, а из так называемой кодовой книги, в которой каждому вектору акустических параметров соответствует некоторое множество векторов артикуляторных параметров.

Формирование кодовой книги для управления артикуляцией и восприятия чужой речи.

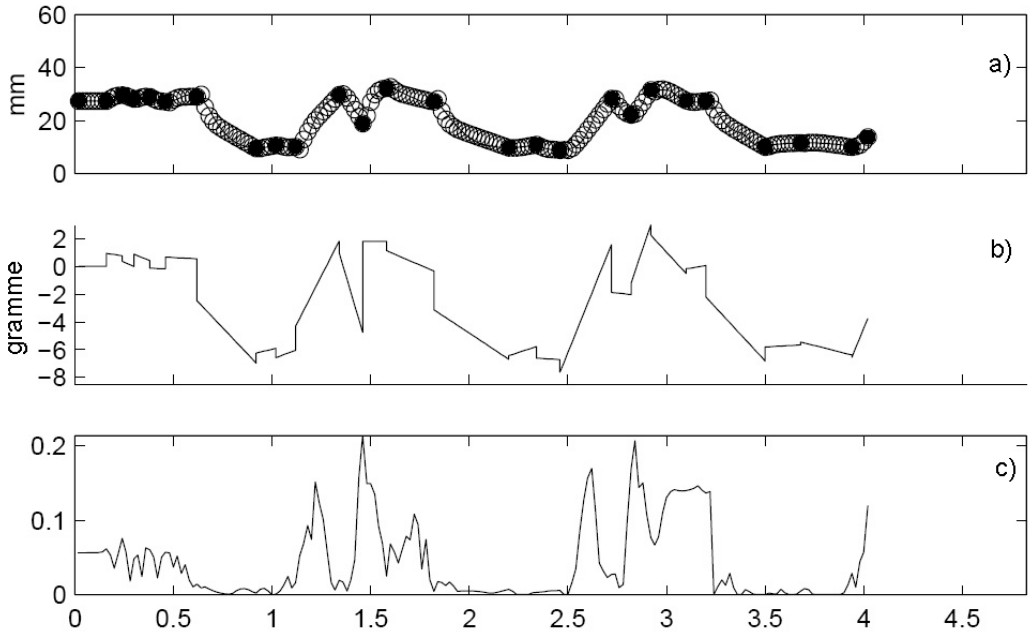
Обратная задача для нейромоторных команд

При решении динамической обратной задачи, т.е. задачи относительно управлений, необходимо создать некоторую модель управлений. Эта *динамическая модель* связывает переменный вектор артикуляторных параметров $x(t) = (x_1(t), x_2(t), \ldots\ldots, x_n(t))$ с вектором $u(t) = (u_1(t), u_2(t), ..., u_n(t))$ управляющих воздействий посредством системы обыкновенных дифференциальных уравнений

$$x_i''+2g_i x_i'+\omega_i^2 x_i = u_i(t), \quad i=1,\ldots,n.$$

Параметры g_i и ω_i этой системы характеризуют динамические свойства *i*-го артикулятора. Координата u_i вектора управления интерпретируется как ускорение $u_i = G_i / m_i$, создаваемое силой G_i, которая развивается мышцами, связанными с *i*-м артикулятором массы m_i.

Модель управлений в виде кусочно-непрерывных линейных функций оказалось достаточно точной и соответствующей свойствам управления сокращением мышц.



Кусочно-линейное управление движениями небной занавески.

Моделирование эффектов возмущений и ускорения.

Энергетические критерии оптимальности при управлении артикуляцией обеспечивают перестройку партитуры команд управления – амплитуд и относительных фаз команд на каждый артикулятор.

Такая перестройка наблюдается как в экспериментах с возмущениями движений артикуляторов, так и в экспериментах по отведению электро-миографических потенциалов (ЭМГ) мышц при изменении темпа артикуляции.

Компьютерное моделирование подтвердило существование эффекта перестройки партитуры команд в условиях, аналогичных реальным наблюдениям.

Внутренняя модель при восприятии речи

Эксперименты по измерению электрической активности различных зон коры головного мозга человека показали, что при высоком отношении сигнал/шум активна только слуховая зона.

При низких отношениях сигнал/шум, наряду со слуховой зоной коры, активизируется и моторная зона.

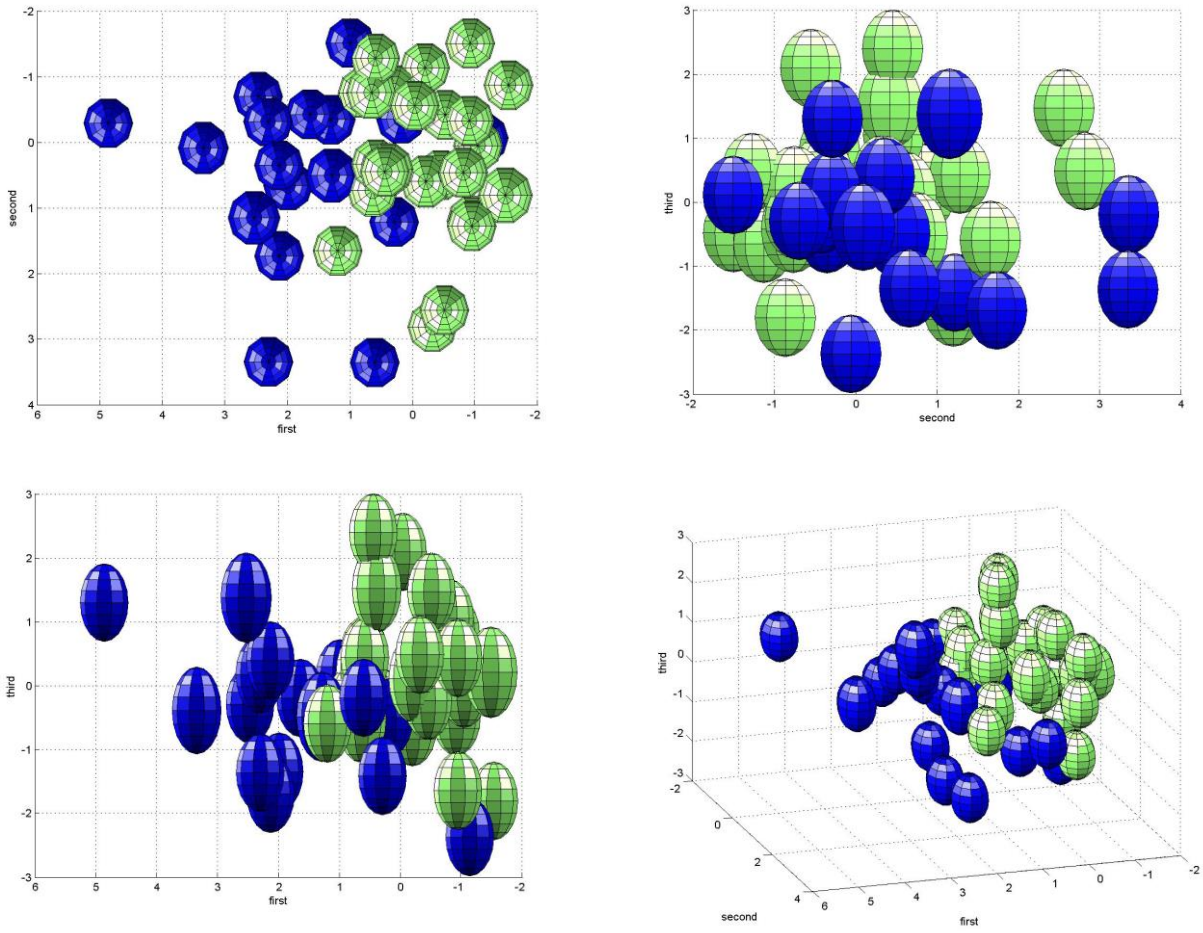
Эти наблюдения подтверждают предположения о том, что, при определенных условиях, в распознавании речи человеком участвует и моторная компонента.

Эти же наблюдения указывают на связь процессов восприятия речи с кодовой структурой речевого сигнала.

Как обсуждалось ранее, анализ потенциальной вероятности правильного распознавания слов указывает на то, что высоком отношении сигнал/шум избыточность речевого кода на уровне слов обеспечивает достаточно хорошие показатели при использовании лишь признаков, легко вычисляемых при акустическом анализе.

При увеличении уровня шумов необходимо использовать признаки места артикуляции которые определяются лишь при переходе уровень формы речевого тракта путем решения обратной задачи, т.е. при использовании моторных компонент сигнала.

Анатомические параметры черепа и внутренних поверхностей ротовой полости содержат информацию, необходимую для распознавания человека



Синий цвет – женщины, зеленый – мужчины.

Решение обратных задач относительно формы речевого тракта, артикуляторных параметров или команд управления артикуляторами перспективно для распознавания речи или диктора.

Параметры голосового источника могут использоваться для распознавания диктора.

Обратная задача для голосового источника

Форма импульса объемной скорости воздушного потока, протекающего через голосовую щель, и площадь голосовой щели определяются путем решения обратной задачи.

Наиболее популярный способ решения обратной задачи состоит в вычислении сигнал-остатка методом линейного предсказания и последующего интегрирования. Этот метод называется обратной фильтрацией. Полученный сигнал принимается за источник голосового возбуждения. От этого источника можно перейти к функции площади голосовой щели. В пространстве параметров этой функции вероятность распознавания мужского голоса составила 94.7%, а вероятность распознавания женского голоса - 95.9%.

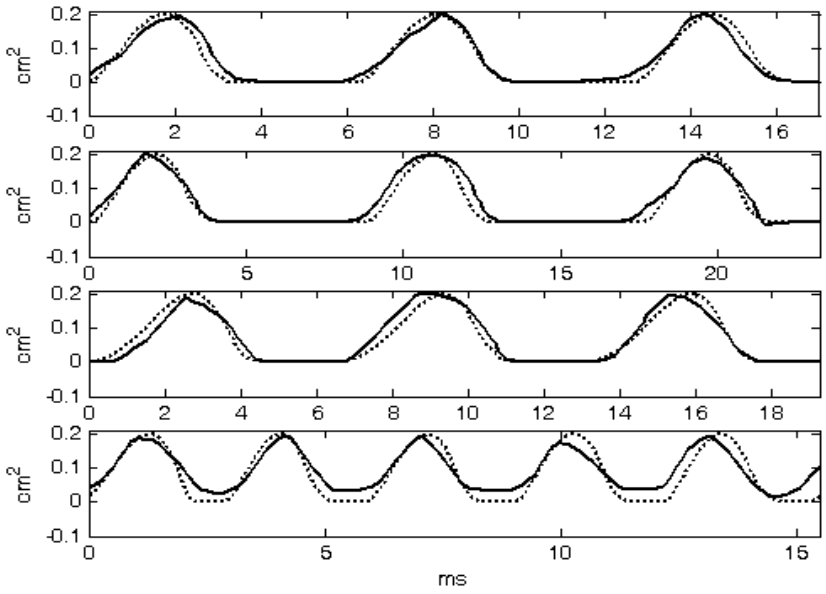
Уравнение потока через голосовую щель может быть решено относительно площади голосовой щели:

$$S(t+\Delta t)=\frac{-b\pm\sqrt{b^2-4ac}}{2a},\quad S(t)\geq 0,$$

где w - объемная скорость потока, полученная интегрирование импульса голосового источника

$$a=-2\alpha\beta\Delta p(t+\Delta t),$$
$$b=2[w(t+\Delta t)-(1-\alpha)w(t)],$$
$$c=\alpha\beta c_x\rho_0 w^2(t+\Delta t).$$

c_x - параметр в уравнении потока, ρ_0 - плотность воздуха,



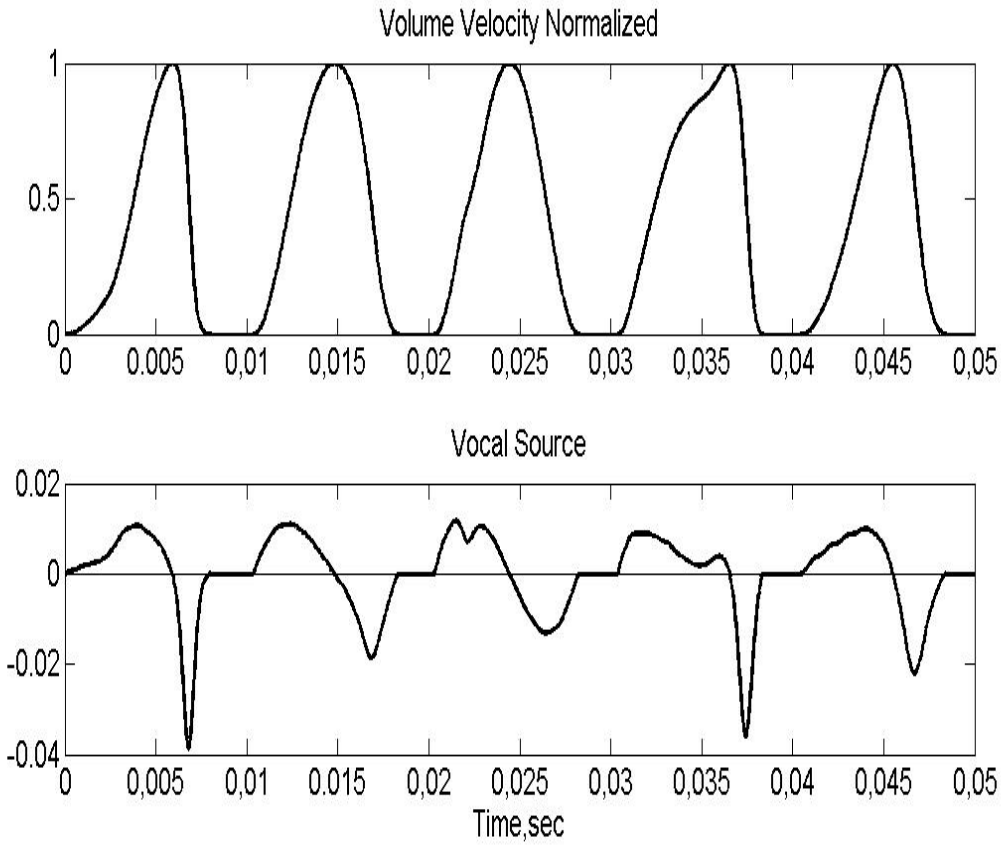
Измеренная (-) и вычисленная (---) площадь голосовой щели

В пространстве параметров функции площади голосовой щели вероятность распознавания мужского голоса составляет 94.7%, а вероятность распознавания женского голоса - 95.9%.

Второй способ заключается в непосредственном вычислении формы голосового источника через отношение мгновенных спектров на интервалах открытой и закрытой голосовой щели

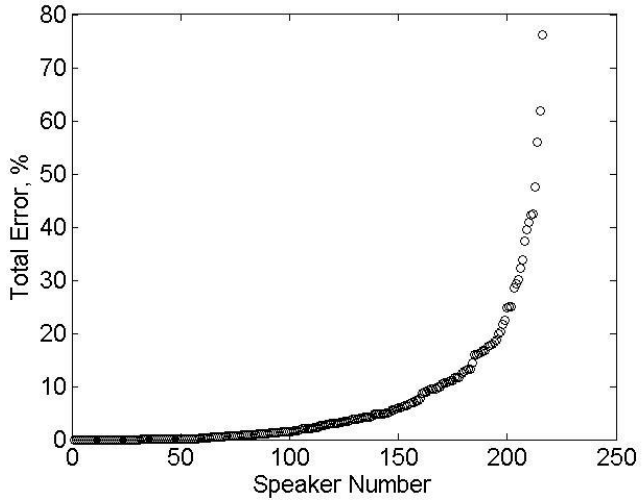
$$S_{vs}(j\omega) = \frac{S_{cl}^{*}(j\omega)S_{op}(j\omega)}{S_{cl}^{*}(j\omega)S_{cl}(j\omega) + \alpha(1+\omega^{2})}.$$

Здесь ω - круговая частота, $\alpha(\delta)$ - параметр регуляризации, $S_{op}(j\omega)$ и $S_{cl}(j\omega)$ - кратковременные спектры на интервалах открытой и закрытой голосовой щели, а $S_{cl}^{*}(j\omega)$ - комплексно-сопряженная функция к $S_{cl}(j\omega)$, j – мнимая единица. Голосовой источник $G(t)$ восстанавливается как действительная часть обратного преобразования кратковременного спектра голосового источника S_{vs}.

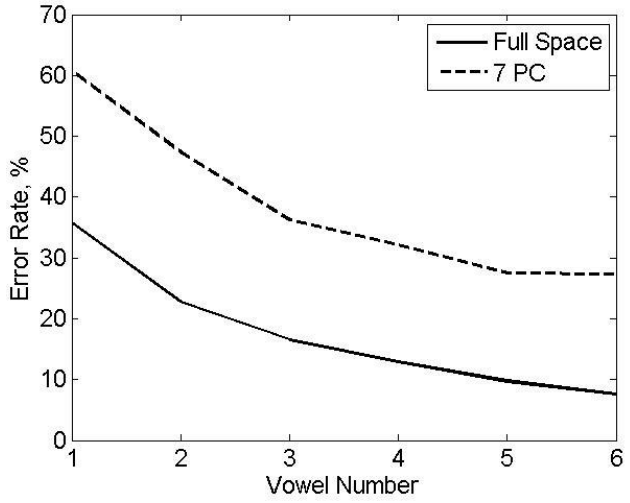


Средняя ошибка распознавания мужских голосов в пространстве периода основного тона и 7 коэффициентов при собственных функциях формы импульса

объемной скорости, 6 гласных. . Средние ошибки FAR=3.7%, FRR=3.9%, но около 25% дикторов имеют суммарную ошибку более 10%. Поэтому критерием качества системы распознавания должна быть не средняя ошибка, а доля дикторов, для которых ошибка не превышает некоторый порог.



Зависимость ошибки распознавания от числа гласных.

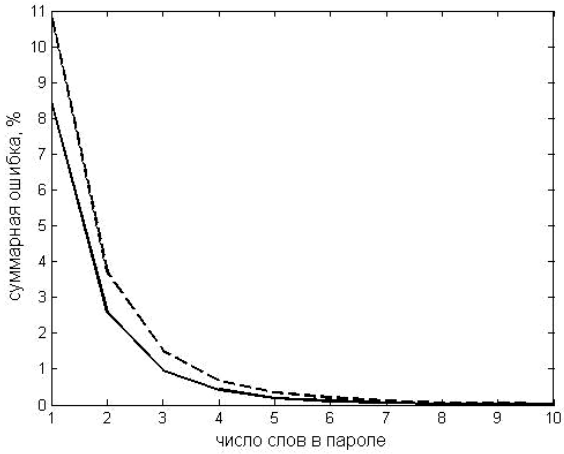


(—) мужчины, (---) женщины

Верификация диктора в пространстве спектрально-временных параметров.

FAR и FRR для мужских голосов

число цифр в пароле	1	2	3	4	5	6	7	8	9	10
ложные отказы, %	3.54	1.067	0.347	0.168	0.063	0.038	0.015	0.006	0.006	0.003
ложные пропуски, %	4.496	1.513	0.628	0.267	0.121	0.062	0.037	0.015	0.007	0.003



Пример синтеза речи по артикуляторным командам, найденным путем решения обратной задачи

- *You wish to know all about my grand father*.
  Well he's nearly ninety three years old.
  And he still thinks as swiftly as ever."

original Microsoft TTS IITP TTS

Решение обратной задачи относительно артикуляторных параметров обеспечивает скорость передачи, близкую к теоретически минимальной.

На передающем конце канала связи вычисляются артикуляторные команды, передаются по каналу, и на приемном конце речевой сигнал синтезируется артикуляторным синтезатором.

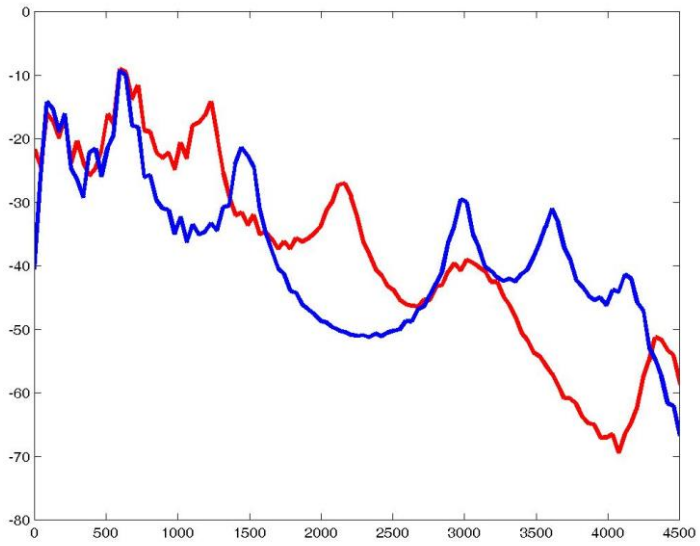
Пример фразы "*the other one is too big*"

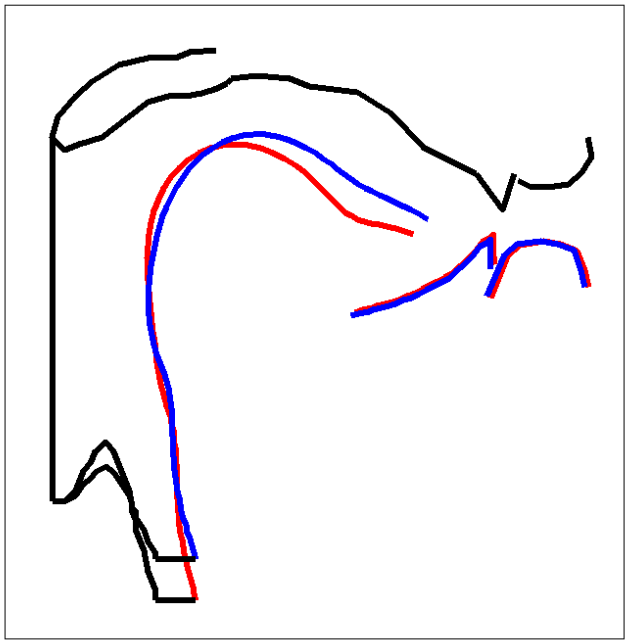
исходный сигнал

9.6 kbps ACELP (FS1016) кодер

1.8 kbps IITP coder

Решение обратной задачи относительно формы речевого тракта обеспечивает визуальную обратную связь в дополнение к акустической обратной связи





Математические модели процессов речеобразования и восприятия речи создают теоретическую основу для создания эффективных методов

автоматического распознавания и понимания речи,

распознавания диктора,

синтеза речи по произвольному тексту,

обучения языку,

сжатия речевого сигнала в каналах связи.

Литература

В.Н.Сорокин, Речевые процессы, 2012, М., Народное образование
В.Н.Сорокин, Синтез речи, 1992, М., Наука.
В.Н.Сорокин, Теория речеобразования, 1985, М., Радио и Связь.
Г.Фант, Акустическая теория речеобразования, 1964, М., Наука.
Д.Фланаган, Анализ, синтез и восприятие речи, 1968, М., Связь.
Д.Маркел, А.Грей, Линейное предсказание речи, 1980, М., Связь.
Л.Рабинер, Р.Шафер, Цифровая обработка речевых сигналов, 1981, М., Радио и Связь.
Физиология речи. Восприятие речи человеком, 1976, Л., Наука.
I.R.Titze, The myoelastic aerodynamic theory of phonation, 2006, NCVS.
K.N.Stevens, Acoustic phonetics, 1996, MIT.