В. Н. Сорокин, Структура проблемы автоматического распознавания речи, *ИТиВС*, 2004, выпуск 2, 25–40

# Структура проблемы автоматического распознавания речи

В.Н. Сорокин

**Аннотация.** Системы автоматического распознавания речи могут достичь необходимого уровня надежности только в том случае, если будет воспроизведена структура процесса распознавания речи человеком. Сюда входят математическая модель восприятия речи, решение обратной задачи (от речевого сигнала к форме речевого тракта) и модель кодовой структуры речевого сообщения вместе с механизмом декодирования.

## Введение

Современные коммерческие системы автоматического распознавания речи достигли уровня надежности распознавания слов близкой, к 90%, особенно при условии адаптации к диктору. Объем распознаваемого словаря расширился до сотен тысяч словоформ. Основные усилия теперь концентрируются на создании интерфейса "человек - машина", наиболее удобного для пользователя. Складывается впечатление, особенно для внешнего наблюдателя, что принципиальные проблемы распознавания речи уже решены и остались лишь технические и эргономические задачи. Это привело к свертыванию фундаментальных исследований за рубежом и закрытию исследовательских отделов в крупнейших частных компаниях. Однако появление коммерческих систем распознавания обусловлено не столько прорывом в решении принципиальных проблем, сколько значительным расширением возможностей персональных компьютеров.

Практика применения систем распознавания речи показала, что они неустойчивы к помехам и искажениям канала речевой связи. Типичным является катастрофическое снижение словесной надежности распознавания до величин порядка 40-60% при появлении относительно слабых шумов, смене типа микрофона или канала связи [1,2]. По некоторым экспертным оценкам исходная надежность распознавания слов в диктофоне компании IBM до адаптации к диктору составляет около 70%, что исключает использование этой системы в режиме независимости от диктора. Сравнительно высокая словесная надежность достигается только при раздельном произнесении слов и только в условиях, близких к тем, при которых происходило обучение.

Подавляющее большинство известных систем распознавания речи основано на применении метода скрытых марковских моделей (СММ). Скрытой марковской моделью является марковский процесс, который не наблюдается непосредственно, а искажен некоторым случайным процессом. Параметрами скрытой марковской модели являются:

- возможные состояния процесса,
- вероятность перехода из одного состояния в другое,
- вероятность искажения наблюдаемого состояния.

Развитие метода скрытых марковских моделей в применении к распознаванию речи нача-

лось после того, как была доказана сходимость простого итеративного алгоритма к параметрам модели [3]. 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.

Поэтому высказываются серьезные сомнения в возможности такого развития метода СММ, которое привело бы к окончательному решению задачи распознавания для любых приложений [4, 5]. Эти сомнения вполне обоснованы, поскольку метод СММ, по существу, не является специфическим для речи и не учитывает фундаментальные свойства речи. Этот метод вполне применим к распознаванию любых акустических сигналов, а не только к распознаванию речи. К числу специфических свойств речи, в первую очередь, относится тот факт, что речевой сигнал предназначен для передачи сообщений и содержит в себе код, специально сконструированный для коррекции ошибок, возникших в процессе речеобразования и передачи речевого сигнала по какому-то каналу связи.

Задача распознавания или понимания речи является обратной задачей в том смысле, что по принятому речевому сигналу нужно восстановить фонетический состав или смысл переданного сообщения. Как известно, обратные задачи часто некорректны, т.е. их решение неоднозначно и неустойчиво относительно помех и искажений. Устойчивое решение обратной задачи может быть получено только при условии использования математической модели распознаваемого процесса и определенных ограничений на возможные решения. Это приводит к необходимости разработки моделей процессов речеобразования и восприятия речи, включая модель кодовой структуры речевого сообщения, поскольку для защиты от помех и искажений речь должна обладать свойствами кодов, исправляющих ошибки.

Постановка задачи автоматического распознавания речи зависит от практического приложения. Собственно распознавание речи подразумевает распознавание или, в более широком смысле, понимание того, что было сказано. Но возможны и другие постановки задачи, например, при распознавании (верификации или идентификации) диктора (кто сказал), распознавании состояния диктора (как сказал) или распознавании среды, окружающей диктора (в каких условиях сказал). Все эти задачи приходится решать с большей или меньшей степенью подробности при любом практическом приложении.

## Изменчивость

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

Отсюда, в частности, вытекают требования к формированию базы данных для обучения системы распознавания. Чтобы избежать настройки на фиксированные условия записи речи, база данных должна быть, по возможности, неоднородной. В ней должны быть представлены разнообразные виды помех и искажений.

Ни один из известных формальных "математических" методов не в состоянии компенсировать все виды изменчивости. Это относится к когда-то популярному методу неоднородной деформации временной оси, скрытым марковским моделям и нейронным сетям. "Физический" подход уделяет большее внимание структуре речевого сигнала и поиску адекватных единиц распознавания. Этот подход в настоящее время, в основном, представлен системами, построенными на основе экспертных знаний, почерпнутых из опыта чтения сонограмм ("видимой речи") [6]. Эти знания весьма субъективны, и задача борьбы с изменчивостью в явном виде в них не формулируется.

В системах понимания речи и в задачах в ограниченной предметной областью любой метод должен дополняться лингвистическим анализом лексических, грамматических, семантических и прагматических связей в речевом потоке.

Концептуальный характер данной статьи и ограничения по ее объему не позволяют вдаваться в обсуждение важных деталей. Поэтому ниже основные элементы проблемы распознавания речи описываются лишь в общих чертах.

## Модели восприятия

Неспособность метода скрытых марковских моделей к подавлению помех и искажений вызвали новый интерес к разработке математических моделей восприятия с целью использования их в системах распознавания речи. В частности, сообщается о значительном повышении устойчивости к шуму при использовании нелинейных моделей периферического отдела слухового анализатора [2, 7-10].

Имеется ряд хорошо установленных свойств восприятия речи и неречевых стимулов. Сюда относятся эффекты адаптации к акустическим свойствам канала связи, включения и выключения стимула, прямой и обратной временной маскировки, спектрально-временного латерального торможения, логарифмическая шкала частот и амплитуд, а также сведения о существовании детекторов амплитудных и частотных модуляций. Это позволяет создать феноменологическую модель первичного анализа речи, обладающую малой чувствительностью к квази-стационарным амплитудно-частотным характеристикам канала связи. В этой модели используются только операции задержки во времени, усреднения по времени или по частоте, а также логарифмирование [11].

Оператор

$$A(\omega,t)=\lg\frac{S(\omega+\Delta\Omega,\theta_1,t\pm\Delta T_1,\tau_1)+C}{S(\omega-\Delta\Omega,\theta_2,t\mp\Delta T_2,\tau_2)+C} \quad (1)$$

описывает акустические (неспецифические) детекторы спектрально-временных неоднородностей сигнала и моделирует многие известные свойства слухового восприятия. Здесь $S$ - спектр мощности принятого сигнала, очищенного от аддитивных шумов, $\Delta\Omega$ - сдвиг отсчета спектра по частоте, $\Delta T_1$ и $\Delta T_2$ - сдвиг отсчета спектра по времени, $\theta_1$ и $\theta_2$ - скользящие интервалы сглаживания спектра по частоте, $\tau_1$ и $\tau_2$ - постоянные времени сглаживания спектральных компонент фильтром первого порядка, $C \geq 1$.

Описывая динамический спектр искаженного речевого сигнала на входе системы распознавания как

$$S(\omega,t)=K(\omega,t)[V(\omega,t)X(\omega,t)+\zeta(\omega,t)], \quad (2)$$

где $X(\omega,t)$ - частотно-временная характеристика речевого тракта, $V(\omega,t)$ - характеристика источника возбуждения, $\zeta(\omega,t)$ - динамический спектр аддитивной помехи, $K(\omega,t)$ - передаточная функция канала, можно показать, что при определенных значениях параметров $\Delta\Omega\ \theta_1\ \theta_2\ C$ функция $A(\omega,t)$ инвариантна к постоянному во времени коэффициенту усиления $K(\omega)$ или мало зависит от него. Таким образом, обеспечивается устойчивость описания речевого сигнала при различных стационарных амплитудно-частотных характеристиках канала. Различные значения параметров в операторе анализа создают многослойное описание речевого сигнала с различными частотно-временными свойствами.

Анализ электрической активности мозга показал, что динамические и статические сегменты звукового стимула вызывают доминирующую реакцию в разных местах слуховой коры [12]. В соответствии с этим можно предположить существование разных алгоритмов обработки статических и динамических сегментов речевого сигнала. Анализ речевого сигнала на стационарном участке позволяет использовать накопление (усреднение) во времени, повышая устойчивость оценки спектральных параметров. Распознавание типа стационарного сегмента - ядра ударного гласного, фрикативного, назального, глухой или звонкой смычки - необходимо не только при декодировании речевого сигнала в акустической области, но и при выборе метода решения обратной задачи относительно формы речевого тракта.

При определенных значениях параметров $\tau_1\ \tau_2\ \Delta T_1\ \Delta T_2$, соответствующих динамическим характеристикам артикуляции и акустических процессов, оператор () подчеркивает переходные процессы разной длительности в спектрально-временной области. Это создает возмож-

ность детектирования так называемых артикуляторных событий, т.е. переходов из одного артикуляторного состояния в другое. Сочетание параметров динамических и статических детекторов позволяет сегментировать речевой сигнал на дискретные участки, классифицируемые в терминах выбранной системы кодовых элементов речевого потока.

На Рис. 1, 2 показаны спектрально-временные изображения речевого сигнала на некоторых уровнях описываемой модели восприятия. Словосочетание "шестьдесят один" , произнесенное слитно, одновременно записывалось через телефонную трубку (Рис. 1) и направленный микрофон, расположенный на расстоянии примерно 25 см от рта диктора (Рис. 2). На интервале времени около 0.5 сек после начала записи производилась оценка спектральных характеристик аддитивного шума ( в каждом канале - своя), а затем была выполнена очистка от шума. Под осциллограммой звукового давления расположена сонограмма, полученная путем локально-частотной нормировки с параметрами $\theta_1$ =20*мел*, $\theta_2$ =800 *мел*. Остальные параметры $\tau_1$ $\tau_2$ $\Delta T_1$ $\Delta T_2$ и $\Omega$ были равны нулю, т.е. временная обработка не производилась. Ниже показана сонограмма, полученная в результате применения спектрально-временного торможения с интегрированием каждой частотной компоненты на площадке, определенной параметрами частоты $\theta_1$ =0, $\theta_2$ =300 *мел*, и на интервале 2 мс. Такой способ использования эффекта латерального торможения позволяет относительно просто решить проблему оценки спектрального профиля синхронно с импульсами голосового источника, не прибегая к измерению мгновенного периода основного тона. На тех сегментах речевого сигнала, на которых отсутствует голосовое возбуждение, отсчеты спектра берутся через 1 мс.

Далее следуют изображения сигналов от двух слоев динамических детекторов. Один слой реагирует на возрастание энергии частотных компонент. Показаны отклики детектора с параметрами $\tau_1$ = 5мс, $\tau_2$ =15мс, $\Delta T_1$=0, $\Delta T_2$=25мс. Второй слой соответствует откликам детектора, реагирующего на спад энергии. Его параметры: $\tau_1$ = 5мс, $\tau_2$ =15мс, $\Delta T_1$=0, $\Delta T_2$= –25мс

Показан также результат действия одного из слоев сегментации на квази-стационарные интервалы, на каждом из которых спектральный профиль изменяется меньше, чем на некоторую величину, оцениваемую по нормированному скалярному произведению между накопленным средним профилем и вновь измеренным профилем. Из этих рисунков видно, что для одного и того же произнесения, но при записи через разные микрофоны и АЦП с разным уровнем шумов, на уровне неспецифических детекторов число и границы квази-стационарных сегментов несколько отличаются. Использование динамических детекторов улучшает устойчивость сегментации, но полной инвариантности все же достичь не удается. Устойчивая сегментация речевого сигнала на элементы, соответствующие артикуляторным состояниям, может быть получена только на следующем уровне анализа, использующем специфические детекторы артикуляторных состояний и переходных процессов.

Модель восприятия речевого сигнала является существенным элементом системы распознавания. Она создает первичное описание, устойчивое к помехам и искажениям в канале связи. Это описание служит входом для последующих блоков - восстановления формы речевого тракта и декодера.

## Обратная задача для речевого тракта

Наблюдения за процессами усвоения речи, компенсации естественных и искусственных нарушений процессов речеобразования и восприятия, привели к гипотезе о существовании в системе управления артикуляцией так называемой внутренней модели [13]. Предполагается, что внутренняя модель располагает информацией о свойствах механики, аэродинамики и акустики речеобразования, а также о фонетических свойствах языка, и выполняет текущий контроль за процессом речеобразования, вычисляя команды обратной связи, необходимые для поддержания параметров речи в заданном диапазоне. Формирование сигналов обратной связи внутренней

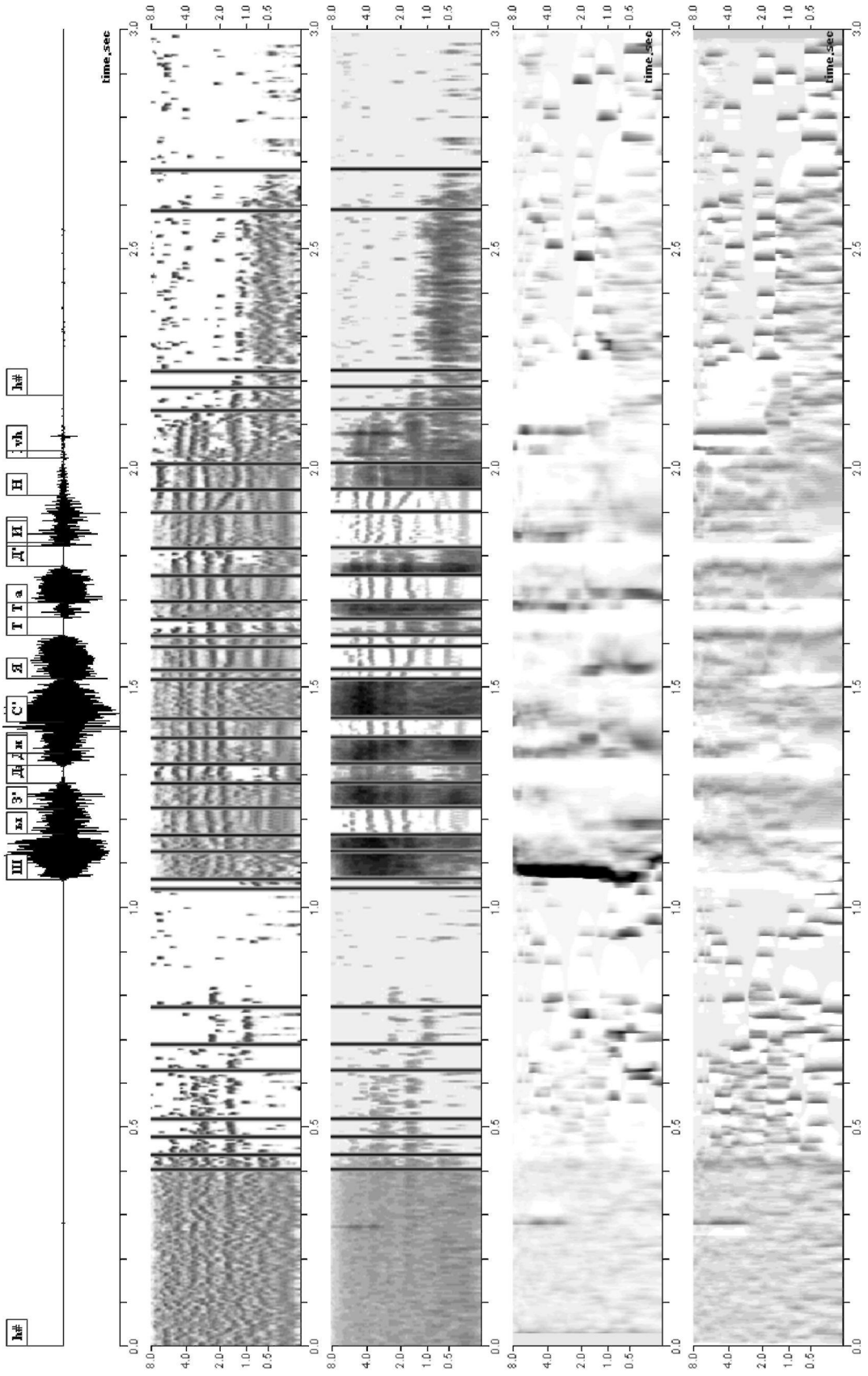


Рис. 1. Последовательность обработки речевого сигнала для словосочетания "шестьдесят один". Телефонная трубка. Описание в тексте

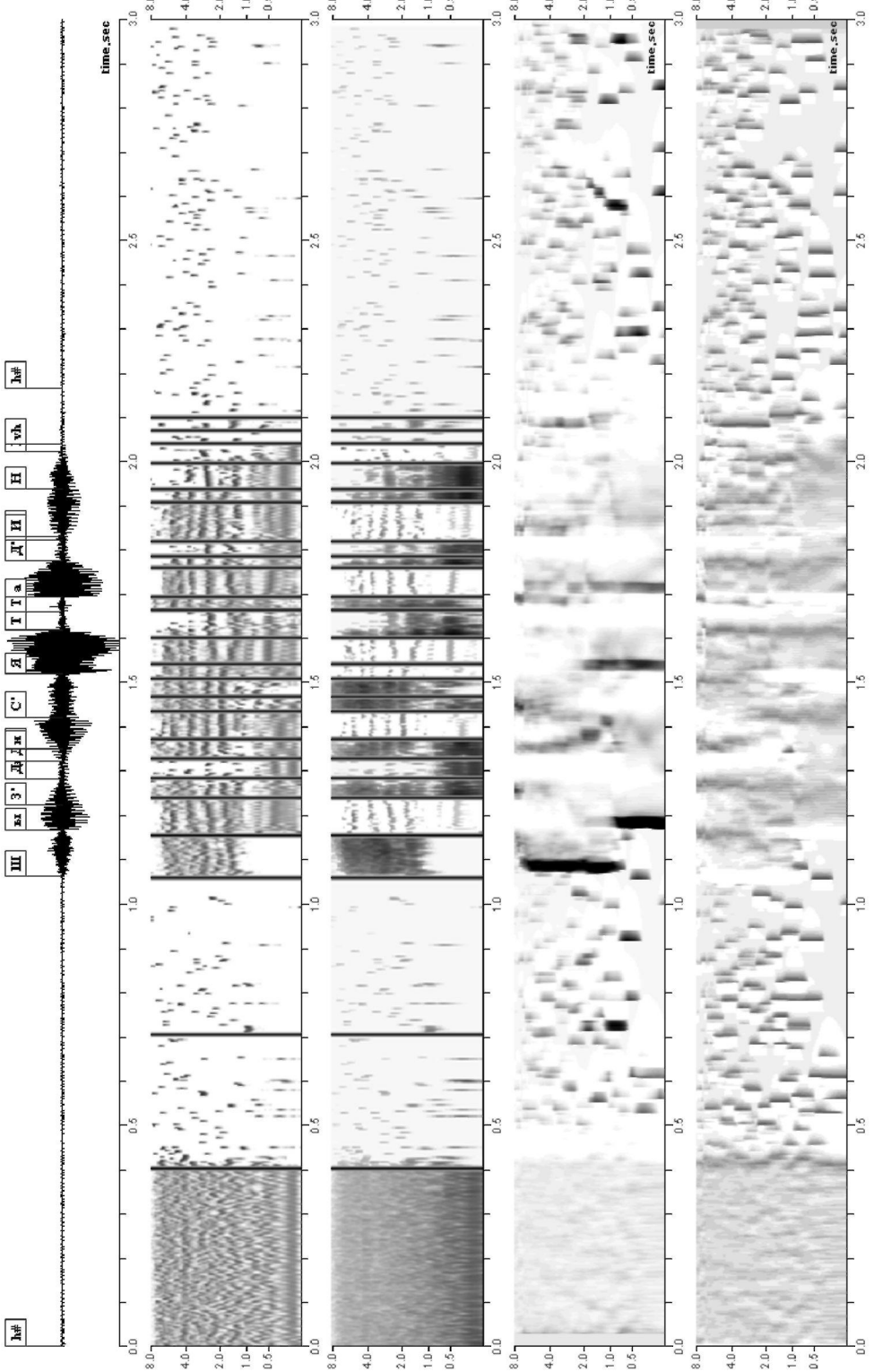


Рис. 2. Последовательность обработки речевого сигнала для словосочетания "шестьдесят один". Направленный микрофон. Описание в тексте

моделью возможно только в том случае, если она решает обратные задачи, вычисляя нейромоторные команды управления артикуляцией по сигналам проприоцепторов внутри речевого тракта или оценкам акустических параметров собственной речи слуховым анализатором этого диктора.

Способность к коррекции параметров собственной речи позволяет предположить, что аналогичный механизм решения обратной задачи "от акустики к артикуляции" может использоваться и при распознавании речи других людей. Это предположение было сформулировано в так называемой моторной теории восприятия, декларирующей, что процесс распознавания речи происходит не только в пространстве акустических параметров, но также и в пространстве управлений артикуляцией.

Долгое время моторная теория восприятия не находила экспериментального подтверждения и критиковалась вследствие трудностей решения обратной задачи и кажущейся достаточности чисто акустического описания речевого сигнала для его распознавания. Однако новые экспериментальные методики позволили найти требуемые доказательства. Была обнаружена активность слуховой зоны коры головного мозга у людей, наблюдающих артикуляторные движения дикторов, тогда как неречевая мимика не вызывала подобной активности [14, 15]. Более того, с помощью метода функционального магнитно-резонансного анализа было установлено, что при восприятии речи в условиях шумов возникает активность в моторной зоне коры головного мозга, тогда как при хороших условиях активизируется только область слуховой коры [16, 17]. Анализ корректирующей способности речевого кода на уровне слов показал, что при хороших условиях речевой связи достаточно использования фонетических признаков, непосредственно измеряемых в речевом сигнале (типа звонкости/глухости, назальности, турбулентности, гласный/согласный). При ухудшении отношения сигнал/шум необходимо использовать информацию о так называемом месте артикуляции - положении наибольшего сужения или смычки вдоль продольной оси речевого тракта. Для этого нужно решить обратную задачу от акустики к форме речевого тракта, т.е. перейти из пространства акустических параметров в пространство артикуляций. Активизация моторной зоны коры в условиях помех указывает на реальность участия моторной компоненты в распознавании речи.

В силу кинематической неоднозначности все обратные задачи для речи являются некорректными по Адамару, т.е. формально для них не гарантируется однозначное и устойчивое решение волнового уравнения относительно площади поперечного сечения речевого тракта и, тем более, относительно артикуляторных параметров. Однако вариационный метод и регуляризация по Тихонову [18] в совокупности с сильными ограничениями на значения и динамику артикуляторных параметров позволяют получить устойчивые и достаточно точные решения речевых обратных задач. 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

$$A(x)=u, \qquad (3)$$

где x - артикуляторные параметры, u - акустические параметры. Решение обратной задачи состоит в минимизации функционала

$$M(x_{h\delta}) = \alpha\Omega(x) + \rho^2(A_h x, u_\delta), \qquad (4)$$

где $\alpha=\alpha(h,\delta)$ - параметр регуляризации, $\rho(A_h x, u_\delta) = \|A_h x - u_\delta\|$ - невязка между вычисленными и входными данными, функционал $\Omega(x)$ - критерий оптимальности, $h$ и $\delta$ - погрешности в описании модели речеобразования и ошибки измерения акустических параметров.

Входными акустическими параметрами для гласных служат три резонансные частоты, а для фрикативных и взрывных сегментов согласных - огибающая спектра. Для назальных необходимо использовать акустическую модель речевого тракта с разветвлением в носовую полость. Акустическими признаками назального сегмента служат дополнительные резонансы и динамика энергии в низкочастотной области. На звонкой

и, особенно, глухой смычках энергия либо сосредоточена в области радиального резонанса речевого тракта, либо ее нет совсем. Таким образом, для выбора метода решения обратной задачи необходимо сначала определить тип сегмента - гласный, назальный, фрикативный, смычка. Это означает предварительную сегментацию речевого сигнала на такие элементы. Процесс минимизации состоит в поиске условного экстремума при наличии ограничений на значения артикуляторных и акустических параметров.

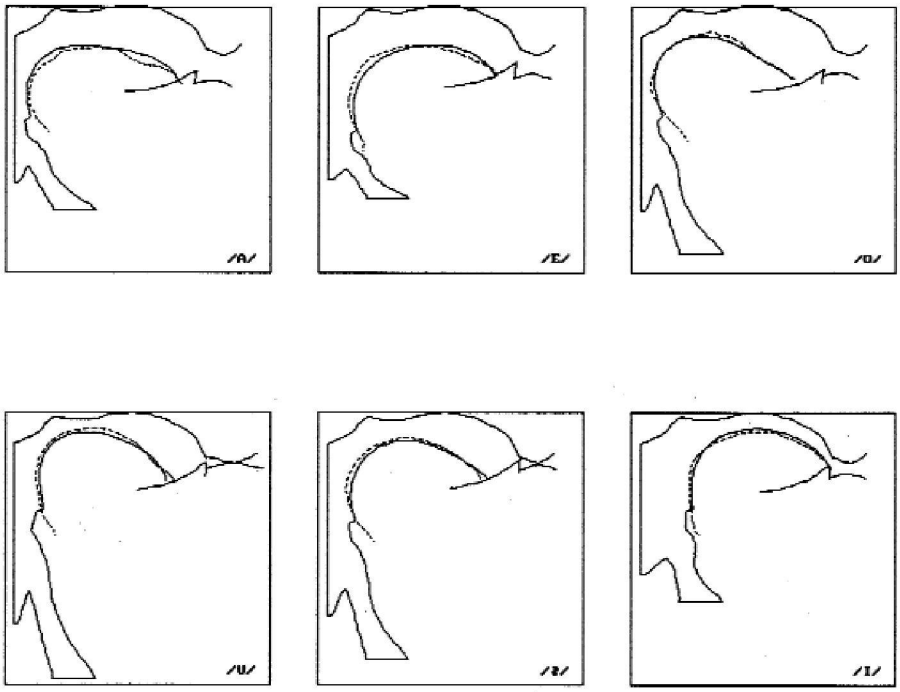

Рис. 3. Измеренная форма языка (- - - ) и решение обратной задачи для гласных (АА)

Критерий минимума работы артикуляторов оказался эффективным при решении обратных задач для стационарных сегментов гласных звуков или фрикативных [19, 20]. На Рис. 3 показаны профили речевого тракта в средне-сагиттальной плоскости, измеренные с помощью рентгенографии, и вычисленные формы тракта для гласных русского языка. Видно, что точность восстановления формы и положения языка вполне удовлетворительна. Погрешность восстановления формантных частот при этом была сравнима с погрешностью их измерений в речевом сигнале. Одновременно восстановлена и форма остальных участков речевого тракта с правдоподобными деталями. Например, на гласном /y/ произошло так называемое огубление (вытягивание губ вперед) и опускание гортани, а на гласном /u/ - подъем гортани.

При решении динамических задач необходимо использовать составной критерий $\Omega = a\Omega_W + b\Omega_T$, $a+b=1$, где

$$\Omega_W = \frac{1}{2T}\sum_k \int_t^{t+T} c_k (x_k - x_k^{(o)})^2 d\tau, \qquad (5)$$

$$\Omega_T = \frac{1}{2T}\sum_k \int_t^{t+T} (m_k x_k'')^2 d\tau, \qquad (6)$$

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 ($\Omega_W$) и средний квадрат полной силы, приложенной к артикуляторам ($\Omega_T$) [21].

Решение динамических обратных задач в ряде случаев также оказывается вполне удовлетворительным. Ошибка аппроксимации движений некоторых точек внутри речевого тракта, измеренных с помощью микролучевого рентгеноскопа, и акустических параметров находятся в пределах погрешности измерений. На Рис. 4 показаны измеренные и вычисленные траектории формант в слоге /au/.

Гласные, сгенерированные артикуляторным синтезатором по результатам решения обратной задачи, субъективно оказываются весьма похожими на оригинальные звуки [22]. Решение ди-

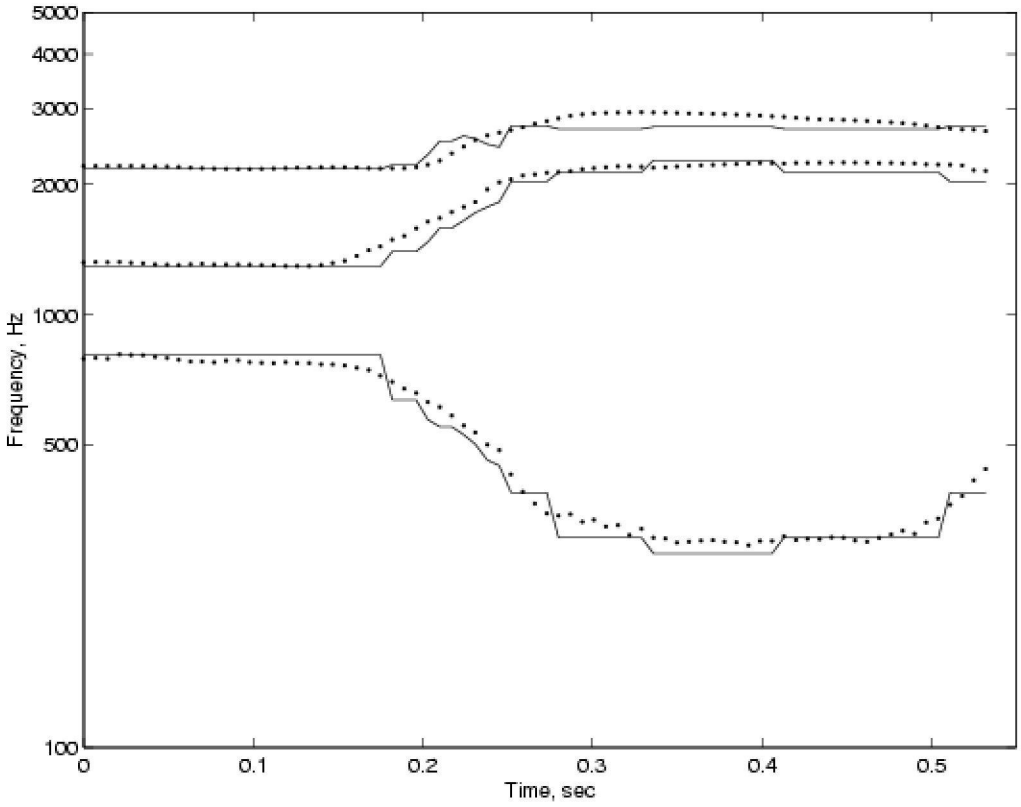


Рис. 4. Измеренные (···) и вычисленные (ААА) траектории формант для дифтонга /аи/

намической обратной задачи для слогов, содержащих фрикативные, также демонстрирует высокую точность восстановления артикуляторных и акустических параметров и похожее звучание. На Рис. 5 показаны сонограммы оригинального слога */аша/* и сонограммы слога, ресинтезированного по решению обратной задачи.

Задача минимизации функционала (4) 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. Первоначально при исследовании речевых обратных задач кодовая книга формировалась с использованием артикуляторного синтезатора [23, 24]. Однако при этом необходимо было либо произвольно задавать геометрические размеры речевого тракта, либо подставлять значения, измеренные на конкретном дикторе. Попытки применения фиксированных анатомических параметров к разным дикторам показали необходимость создания кодовой книги, в которой, помимо артикуляторных параметров, присутствуют и анатомические параметры для разных дикторов [25]. Таким образом, поиск методов решения обратной задачи для речевого тракта приводит к традиционному подходу при распознавании речи, т.е. к необходимости обучения кодовой книги на представительном множестве дикторов.

Создание таких кодовых книг стало возможным с появлением относительно дешевых систем типа Articulograph для измерения движений некоторых точек внутри речевого тракта. Эти системы используют принцип возбуждения тока в крошечных катушках-сенсорах, перемещающихся в неоднородном электромагнитном поле, генерируемым излучателями с разной частотой.

Теперь кодовая книга становится многослойной, и каждый слой содержит геометрические параметры речевого тракта диктора в дополнение к артикуляторным параметрам и соответствующим акустическим параметрам. Процесс решения обратной задачи для произвольного диктора включает определение типа анатомии, доставляющей наименьшую ошибку в реконструкции измеренных акустических параметров. Тем самым производится как бы адаптация к индивидуальным характеристикам диктора, но это совершенно не похоже на системы с настройкой на диктора.

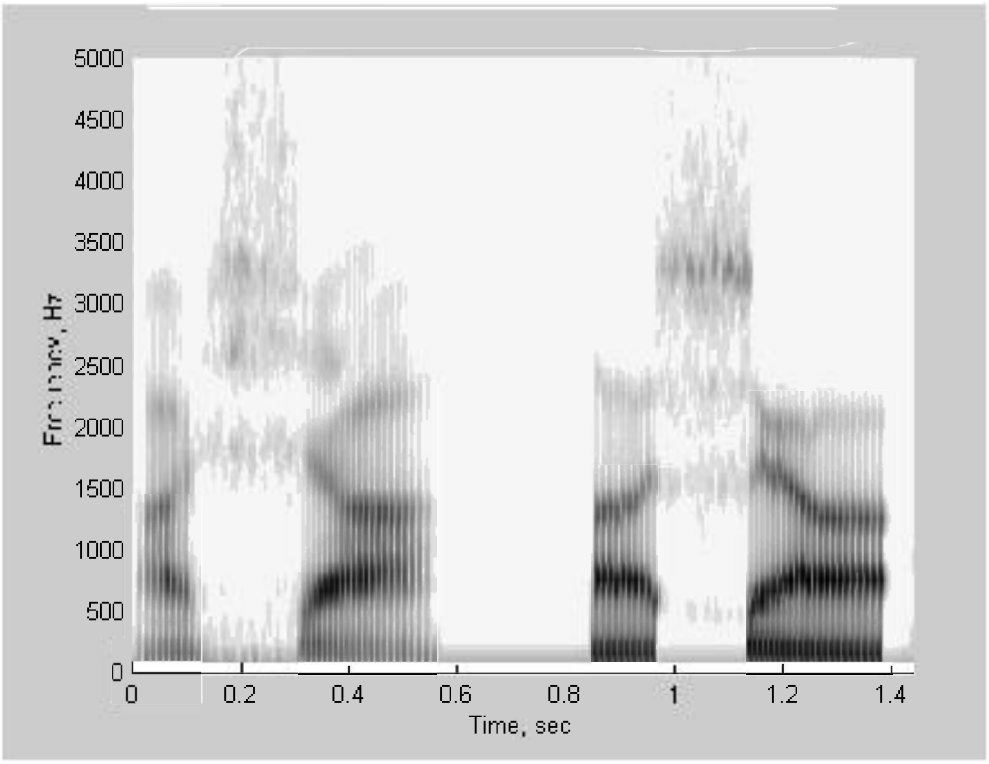


Рис. 5. Сонограммы слога /аша/. Оригинальное (слева) и ресинтезированное (справа) звукосочетание

Решение обратной задачи относительно формы речевого тракта может оказаться особенно полезным при определении так называемого места артикуляции - наибольшего сужения или смычки. Этот признак необходим для различения согласных, но он плохо определяется на акустическом уровне. Таким образом, дополнение пространства акустических признаков пространством артикуляторных параметров дает надежду на повышение устойчивости и надежности систем автоматического распознавания речи.

## Кодовые свойства речи

С точки зрения современной теории кодов, корректирующих ошибки, речь принадлежит к классу нелинейных кодов, поскольку всегда найдется хотя бы одна пара слов, которая при любом методе их сложения не образует новое осмысленное слово. Декодирование таких кодов возможно только полным перебором всех возможных слов. Такой перебор может быть реализован с помощью динамического программирования, в частности, методом Витерби, или методом последовательного декодирования. Первые эксперименты по применению метода последовательного декодирования к распознаванию речи были выполнены в [26, 27].

В организации речевого кода просматривается структура, аналогичная каскадным кодам, поскольку коррекция ошибок возможна за счет использования избыточности на уровне артикуляции (не все последовательности артикуляторных состояний физически реализуемы), признаков фонетических элементов, слогов, слов и фраз. Существуют также уровни семантических и прагматических ограничений. Декодирование речевого сигнала с использованием предсказания, полученного от разных уровней, позволяет быстро уменьшить число конкурирующих вариантов вместо их экспоненциального роста, когда

используется только информация о прошлых состояниях.

В [28] было показано, что слова, по крайней мере русской речи, записанные в фонетическом коде, обладают свойствами так называемых префиксных кодов, у которых ни одно кодовое слово не служит началом другого. Для 2500 наиболее часто встречающихся слов найдено менее 7% слов-префиксов, которые состоят из одно-двухбуквенных союзов, предлогов и местоимений. Основное свойство префиксных кодов состоит в возможности декодирования слитных сообщений, в которых кодовые слова не разделены паузами или специальными символами. Это имеет принципиальное значение для распознавания слитной речи. Одновременно выяснилось, что вероятность появления фонем в речи определяется не их помехоустойчивостью, а сложностью их образования.

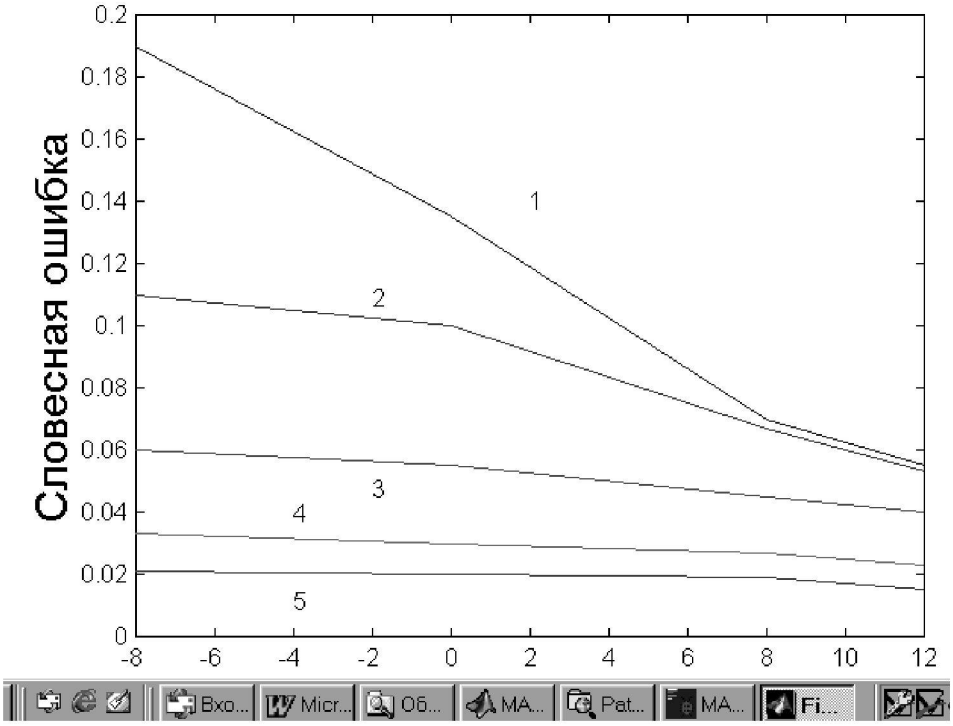


Рис. 6. Вероятность ошибки распознавания слов при разных оценках скорости передачи информации фонетическим кодом речи

1 - субъективная (по [29]),
2 - 3 - декодирование по независимым признакам,
4 - 5 - декодирование по фонемам

С помощью теоремы о кодировании и результатов психоакустических экспериментов по восприятию речи в присутствии белого шума с разным отношением сигнал/шум в [28] была оценена потенциальная надежность распознавания слов в случае, когда не используется синтаксическая, семантическая и прагматическая избыточность речи (Рис. 6).

Как видно, при достаточно хороших отношениях сигнал/шум реальная словесная разборчивость и теоретические оценки близки независимо от того, выполняется ли декодирование по независимым признакам или по сложным комплексам признаков, которыми являются фонемы. При более высоких уровнях шумов теоретически достижима меньшая ошибка распознавания, но человек почему-то не использует все возможности для коррекции ошибок на словесном уровне. Похоже, что при плохих условиях восприятия человек либо использует корректирующую способность более высоких уровней, либо прибегает к переспросу. Это может быть связано с какими-то ограничениями на сложность переработки информации в мозгу человека. Аналогичные явления наблюдаются и в технических системах. Например, ограничения на сложность декодирования могут привести к тому, что неоптимальный метод декодирования, не использующий полностью кодовую избыточность, обеспечивает меньшую ошибку, чем метод, потенциально способный использовать всю кодовую избыточность для исправления ошибок, но требующий чрезмерного количества вычислений.

С другой стороны, полученные оценки могут свидетельствовать о том, что системы автоматического распознавания речи способны достигнуть гораздо большей устойчивости к аддитивным шумам при достаточных вычислительных ресурсах.

Признаки фонем разделяются на две группы. В одну из них входят признаки, сравнительно легко вычисляемые на акустическом уровне. Это признаки голосового и шумового источников возбуждения, смычки и назальности. При хороших акустических условиях эти признаки обеспечивают достаточно высокую различимость слов и, в совокупности с избыточностью высших уровней, гарантируют приемлемую разборчивость речи. Эти же признаки эффективно работают и при быстрой сортировке эталонов больших словарей [30]. Как упоминалось выше, эти признаки также участвуют при решении обратной задачи относительно формы речевого тракта.

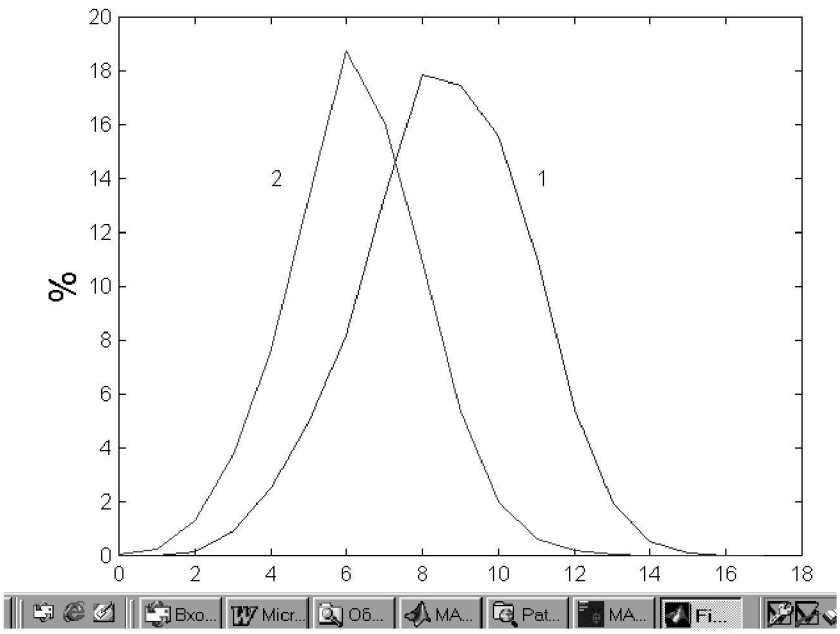


Рис. 7. Спектр кодовых расстояний для числительных

1 - кодирование фонем по полному множеству признаков,
2 - кодирование только признакам "гласный, смычка, назальный, фрикативный"

Однако по мере ухудшения условий речевого общения все большую роль играют признаки места артикуляции, которые нередко плохо вычисляются на акустическом уровне. Пример ухудшения различимости слов в очень ограниченном словаре числительных в 53 слова в задаче голосового набора телефонного номера показан на Рис. 7. Слова, произнесенные сорока семью дикторами в различных контекстах, записывались в фонетических терминах, а затем вычислялось минимальное расстояние по Хеммингу при всевозможных сдвигах слов относительно друг друга. На Рис. 7 изображены распределения (спектр) кодовых расстояний этих слов. Видно, что по сравнению с полным кодированием фонем с участием признака места артикуляции, усеченные коды с использованием только "акустических" признаков существенно сдвигают спектр кодовых расстояний в сторону меньших значений, а значит, уменьшают и исправляющую способность фонетического кода.

Как уже упоминалось выше, имеются экспериментальные свидетельства того, что при высоком уровне помех человек прибегает к вычислению каких-то артикуляторных компонент для улучшения надежности восприятия речи. Похоже, что для определения места артикуляции действительно необходимо решение обратной задачи относительно формы речевого тракта. Потребность в таком решении тем выше, чем менее доступна информация о синтаксисе, семантике и прагматических ограничениях в задаче понимания речи. Роль высоких уровней видна из сопоставления субъективной оценки надежности восприятия фраз и оценки распознавания, например, семизначного телефонного номера в предположении независимости появления числительных (Рис. 8).

В английском языке действуют более сильные синтаксические ограничения, чем в русском. Например, порядок слов во фразе задается гораздо более жестко, чем в русском, что обеспечивает большую предсказуемость следования

разных частей речи. Поэтому и зависимость фразовой разборчивости от словесной у английского языка лучше, чем у русского. В задаче телефонного набора практически нет никакой грамматики, семантики или прагматики. Рабочая характеристика для надежности распознавания семизначного номера определяется произведением вероятности правильного распознавания каждого слова. Как видно, в этой задаче предъявляются значительно более высокие требования к надежности распознавания слов, чем в задачах, в которых можно использовать исправляющую способность высших уровней. Ясно, что надлежащее использование корректирующей способности кодовых уровней выше лексического позволит существенно повысить надежность распознавания для "литературно правильной" речи. Проблема, однако, состоит в том, что произвольная речь не подчиняется правилам, выработанным для письменной речи (т.е. для текстов) и эти правила предстоит только установить.

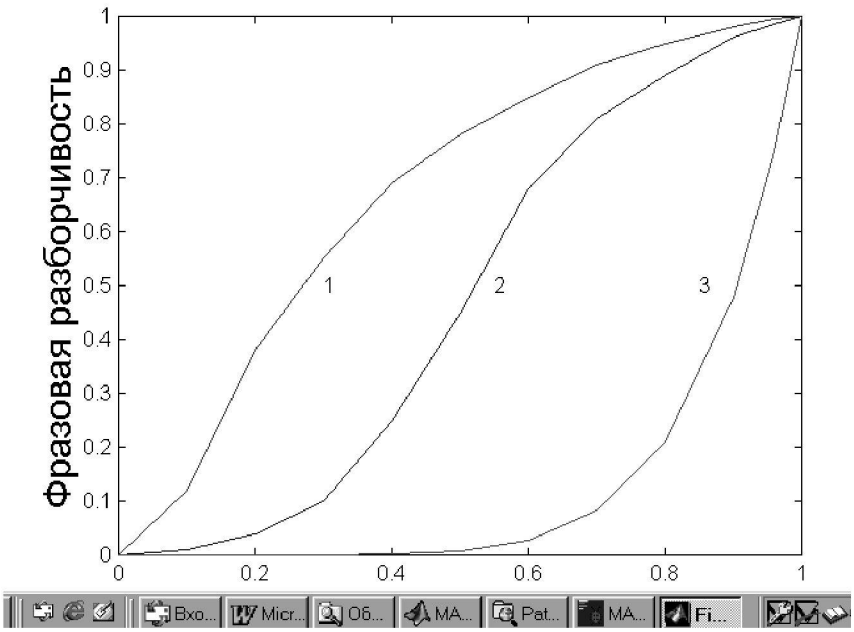


Рис. 8. Зависимость измеренной фразовой разборчивости от словесной

1 - для английского языка (по [31]), 2 - для русского языка (по [29]). 3 - вычисленная разборчивость семизначного телефонного номера

## Распознавание числительных

Распознавание числительных является хорошим тестом для системы распознавания, позволяя оценить эффективность применяемых алгоритмов только на лексическом уровне. Кроме того, набор телефонного номера с голоса, ввод цифровых данных в компьютер или даже простое расширение возможностей калькулятора являются вполне востребованными практическими приложениями.

Первая версия системы распознавания числительных была испытана в начале 90-х годов с использованием базы данных для английского языка TI46. В этой базе содержались речевые сигналы для 16 дикторов (8 мужчин и 8 женщин), записанные в тихой комнате через АЦП на 12 бит. Словарь состоял из 10 числительных и 10 команд. Каждое слово повторялось диктором 20 раз. Детекторы артикуляторных событий вычислялись в 8 частотных полосах, а в число статических признаков входили коэффициент автокорреляции, три формантные частоты и некоторые другие признаки. Каждое слово представлялось в виде матрицы апостериорных вероятностей признаков, а мера сходства вычислялась в метрике L4. При тестировании системы использовался скользящий алгоритм, позволяющий практически полностью использовать для обучения все произнесения. В режиме распознавания независимо от диктора была достигнута надежность порядка 93%. Для того времени это была очень высокая надежность, но надо учитывать, что она была получена на однородной базе данных с хорошими акустическими условиями. Даже с использованием сигнального процессора на компьютере 486 серии, отношение времени принятия решения к длительности произнесения составляло около 200. Описание системы представлено в патенте [32].

Вторая версия системы распознавания числительных была разработана для русского языка. Были предприняты меры для исключения настройки на базу данных. Для каждого из 47 дикторов использовалось по два микрофона с синхронным вводом речи в компьютер, причем для каждой трети дикторов применялись разные наборы микрофонов и расстояния до них. Всего использовалось 2 типа телефонных трубок и 3 типа микрофонов. В дополнение примерно 2/3 сигналов из базы данных были пропущены через симулятор телефонного канала. В развитие первой версии были использованы уточненные параметры детекторов артикуляторных событий и статические признаки и исследовались различные алгоритмы обучения системы. Декодер представлял собой вариант метода списочного декодирования. Решение принималось по принципу максимума апостериорной вероятности. При распознавании слов из базы в режиме независимости от диктора была получена надежность распознавания несколько выше 88%. Испытания системы в реальном времени (задержка решения - около 70 мс) на дикторах, не участвовавших в формировании базы данных, показали, в среднем, около 6% ошибок и 6% переспросов [33]. Достигнутая надежность распознавания слов, согласно Рис. 8, обеспечила бы надежность понимания фраз около 93 - 95%, конечно, при условии надлежащего использования фразовой избыточности. Однако разброс показателей надежности даже для одного и того же диктора оказался слишком велик для того, чтобы рекомендовать эту версию к коммерческому применению.

Ухудшение показателей испытанных систем по сравнению с ожидаемыми частично объясняется ограниченностью вычислительной мощности, что приводило иногда к весьма грубому упрощению алгоритмов анализа и декодирования. Частично это было связано с недостаточной представительностью обучающей выборки. Необходимо отметить, что надежность распознавания слов около 90%, декларируемая различными коммерческими системами, достигается с использованием "модели языка", т.е. с поддержкой корректирующей способности высших уровней, и для тех же условий распознавания, что и для обучения, тогда как сравнимые показатели надежности для числительных получены исключительно с коррекцией на лексическом уровне и на неоднородной базе данных. Опыт разработки и испытания этих систем позволил определить направление дальнейших исследований. В частности, выяснилась необходимость более полного использования каскадной структуры речевого кода и, соответственно, построения адекватного метода декодирования.

## Заключение

Для того, чтобы создать систему автоматического распознавания или понимания речи, которая обладала бы такой же надежностью, как и система слухового восприятия, необходимо воспроизвести основные механизмы анализа речи человеком. Модель первичного анализа должна обеспечить подавление помех и искажений канала связи, сегментацию на акустически однородные участки и детектирование спектрально-временных неоднородностей с разным частотным и временным разрешением. Детекторы артикуляторных событий и состояний должны быть обучены на артикуляторно-фонетический состав конкретного языка. Эти детекторы формируют поток дискретных (кодовых) элементов речи с оценками апостериорных вероятностей. При высоком уровне помех и неопределенности грамматики должна решаться обратная задача относительно формы речевого тракта. Необходимо найти адекватный метод декодирования потока кодовых элементов, используя ограничения лексического, синтаксического, семантического и прагматического уровня.

## Литература

1. Martin, A., Fiscus, J., Przybocki, M., Fisher, B., (1998). The evaluation: word error rates and confidence analysis. Proc. 9th Hub-5 Conversational Speech Recognition Workshop, Linthicum Heights, MD.
2. Tchorz, J., Kollmeier, B., (1999). A model of auditory perception as front end for automatic speech recognition. J. Acoust. Soc. Amer., v. 106, N4. Pt.1. pp. 2040-2050.

3. Baum L.E. (1972), An inequality and associated maximization technique in statistical estimation of probabilistic functions of Markov process, Inequalities, v. 3, pp. 1-8.
4. Lippmann, R.P., (1997). Speech recognition by machines and humans. Speech Communication, v. 22, pp. 1-16.
5. Hermansky, H., (1998). Should recognizers have ears? Speech Communication, v. 25, pp. 3-28.
6. Zue, V., Glass, J., Phillips, D., Seneff, S., (1990). The SUMMIT speech recognition system: phonological modelling and lexical access. Proc. ICASSP-90, Glasgow, Scotland, pp. 389-392.
7. Dau T., Puschel D., Kohlrausch A., (1996). A quantitative model of the "effective" signal processing in the auditory system: I. Model structure. J. Acoust. Soc. Amer., v. 99. N6. pp. 3615-3622.
8. Lippmann R.P., (1997).Speech recognition by machines and humans. Speech Communication, v. 22, pp. 1-16.
9. Kingsbury B.E.D., Morgan N., Greenberg S., (1998). Robust speech recognition using the modulation spectrogram. Speech Communication, v. 25, pp. 117-132.
10. Hermansky H., (1990). Perceptual linear predictive (PLP) analysis of speech. J. Acoust. Soc. Am., v. 87, pp. 1738-1752.
11. В.Н.Сорокин, (2003). Модель многослойного первичного анализа речевых сигналов. Труды 13-й сессии Российского акустического общества, с. 11-16.
12. Seifritz E., Esposito F., Hennel F., Mustofic H., Neuhoff J.G., Bilecen D., Tedeschi G., Scheffler K., Di Salle F., (2002). Spatiotemporal pattern of neural processing in the human auditory cortex. Science, v. 297, N 5587, pp. 1706-1708.
13. Sorokin V., Olshansky V. & Kozhanov L., (1998). Internal model in articulatory control: Evidence from speaking without larynx. Speech Communication, v. 25, pp. 249-268.
14. Sams M., Aulanko R., Hamalainen H., Lounasmaa O., Lu S., & Simola J., (1991). Visual information from lip movements modifies activity in the human auditory cortex. Neuroscience Letters, v. 127, pp. 141-145.
15. Calvert G., Brammer M., Bullmore E., Campbell R., Williams S., McGuire P., Voodruff P., Iversen S., & David A., (1997). Activation of auditory cortex during silent lipreading. Science, v. 276, pp. 593-596.
16. Callan D.E., Callan A.M., Kroos Ch., Vatikiotis-Bateson E., (2000). Neural processes underlying perception of audio-visual speech production. Proc. 5th Seminar on Speech Production, Kloster Seeon, pp. 273-276.
17. Sekiyama K, Sugita Y., (2002), Auditory-visual speech perception examined by brain imaging and reaction time, Proc. 7th Int. Conf. On Spoken Language Processing, Denver, pp. 1693-1696.
18. Тихонов А.Н., Леонов А.С., Ягола А.Г., (1995). Нелинейные некорректные задачи. М.: Наука. 310 с.
19. Sorokin V.N., (1994). Inverse problem for fricatives. Speech Communication, v. 14, pp. 249 - 262.
20. Sorokin V.N., Leonov A.S., Trushkin A.V., (2000). Estimation of stability and accuracy of inverse problem solution for the vocal tract. Speech Communication, v. 30, pp. 55-74.
21. Леонов А.С., Сорокин В.Н., (2003). Энергетические критерии оптимальности в речевых обратных задачах, Доклады Академии наук, т. 392, № 5, с. 694-699.
22. Леонов А.С., Макаров И.С., Сорокин В.Н., Цыплихин А.И., (2003). Артикуляторный ресинтез гласных. Информационные процессы, том 3, № 2, с. 71-82.
23. Atal B.S., Chang J.J., Mathews M.V., Tukey J.W., (1978). Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer sorting technique, J. Acoust. Soc. Am., v. 63, pp. 1535-1555
24. Sorokin V.N., Trushkin A.V., (1996). Articulatory-to-acoustic mapping for inverse problem. Speech Communication, v. 19, pp. 105-118.
25. Sorokin V.N., (1992). Determination of vocal tract shape for vowels. Speech Communication, v. 11, p. 71-85.
26. Jelinek, F., Bahl, L.R., Mercer, R.L., (1975). Design of linguistic statistical decoder for recognition of continuous speech. IEEE Trans. Inform. Theory, v. 1, pp. 250-256.
27. Зигангиров К.Ш., Сорокин В.Н., (1977). Применение последовательного декодирования к распознаванию слитной речи. Проблемы передачи информации, N 4, с. 81- 28.
28. Сорокин В.Н., (1985). Теория речеобразования, Радио и связь, М., 313 с.
29. Покровский Н.Б., (1976). Расчет и измерение разборчивости речи. Связь, Москва.
30. Sorokin V.N, (2003), Some coding properties of speech. Speech Communication, v. 40, pp. 409-423.
31. Фланаган Дж., (1968). Анализ, синтез и восприятие речи, Связь, Москва.
32. Сорокин В.Н., (1995). Способ распознавания изолированных слов речи с адаптацией к диктору. Патент на изобретение, Бюллетень изобретений, N31.
33. Сорокин В.Н., Ижнин А.Н., Цыплихин А.И., Чепелев Д.Н., (2003), Артикуляторно-ориентированная система распознавания речи, Труды Международного семинара "Диалог", с. 657-662.

**Сорокин Виктор Николаевич**. Родился в 1938 г. Доктор физико-математических наук. Автор более 120 научных работ, в том числе двух монографий. Специалист в области речевых технологий. Ведущий научный сотрудник Института проблем передачи информации РАН. Член правления Российского акустического общества, член Американского акустического общества.