Информационные процессы, Том 6 , № 3, стр. 177-207

====== ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ В ТЕХНИЧЕСКИХ ======
====== И СОЦИАЛЬНО-ЭКОНОМИЧЕСКИХ СИСТЕМАХ ======

Сегментация речи на кардинальные элементы

А.И.Цыплихин, В.Н.Сорокин
Институт проблем передачи информации, Российская академия наук, Москва, Россия

Поступила в редколлегию 05.2006

Аннотация—Для сегментации речевого сигнала выполнялся поиск границ квазистационарных и переходных процессов, основанный на корреляции между кратковременными спектрами равноотстоящих по времени участков сигнала. Распознавание кардинальных типов сегментов (гласноподобные, назальные, фрикативные глухие и звонкие, смычные глухие и звонкие) выполнялось в пространствах акустических параметров, установленных в результате исследования. Моделирование плотностей вероятности выборок осуществлялось разработанной модификацией EM-алгоритма. Анализ результатов сегментации производился на материале представительной речевой базы для нескольких типов телефонных трубок и микрофонов с ручной разметкой на артикуляторно-акустические сегменты. Средняя погрешность положения границ составила 4,52 мс, среднее число вставок было равно 1,26 на один сегмент разметки, а среднее число пропусков – 0,95%. В 96,3% случаев правильный тип сегмента по вероятности входил в первую двойку, в 85% был на первом месте.

1. ВВЕДЕНИЕ

Кардинальные элементы речевого сигнала – это группы звуков, созданных с использованием существенно различающихся механизмов речеобразования. Рассматривается шесть типов кардинальных элементов: гласноподобные, назальные, фрикативные глухие и звонкие, смычные глухие и звонкие звуки речи. Такая классификация сегментов речевого потока основана на характерных значениях минимальной величины площади поперечного сечения речевого тракта и прохода в носовую полость и на наличии голосовых импульсов.

Гласноподобные звуки характеризуются значением минимальной площади поперечного сечения ротовой полости речевого тракта, большим некоторой величины, $S_{\min} > S_{vow}$. *Фрикативные* сегменты речи формируются с участием турбулентного источника возбуждения акустических колебаний в речевом тракте. Одним из условий возникновения турбулентности воздушного потока в речевом тракте является $0 > S_{\min} > S_{fric}$. *Смычные* звуки содержат сегмент с полным перекрытием речевого тракта в каком-либо месте, т.е. $S_{\min} = 0$. *Назальные* сегменты, наряду с перекрытием речевого тракта, характеризуются открытием прохода в носовую полость, $S_{nas} > 0$, тогда как остальные типы звуков образуются с поднятой небной занавеской, и $S_{nas} = 0$. *Звонкие* звуки характеризуются присутствием импульсов, порождаемых голосовыми складками, *глухие*, соответственно, отсутствием этих импульсов.

Такая классификация, конечно, несколько условна. Благодаря особенностям артикуляции, смычка взрывных звуков, таких как /*б, д, г, п, т, к*/, может быть неполной, сопровождаясь турбулентными шумами. Непроизвольное опускание небной занавески приводит к назализации гласноподобных звуков, а эффекты коартикуляции назализуют сегменты гласных в окрестности назальной смычки. Тем не менее, шесть кардинальных типов сегментов объективно существуют в речевом потоке.

В речевых исследованиях существует задача, в которой необходимо различать, по крайней мере, шесть типов сегментов, которые как раз и являются кардинальными [1]. Это так называемая обратная задача, где форма речевого тракта вычисляется по параметрам речевого сигнала. При решении речевой обратной задачи минимизируется невязка между измеренными

и вычисленными акустическими параметрами. Поскольку каждый тип сегментов характеризуется своими, отличными от других акустическими параметрами, это необходимо принимать во внимание при решении обратной задачи. Гласноподобные сегменты, куда относятся все гласные, полугласные /*в, л, й*/, а также переходные процессы в окрестности взрывных и фрикативных согласных и назальных, описываются траекториями резонансных частот в спектре этих звуков. При решении обратной задачи для назальных звуков небная занавеска должна быть опущена, что создает иное распределение резонансных частот, чем на гласноподобных звуках. Поэтому нужна информация о том, является ли сегмент просто гласноподобным, назализованным или назальной смычкой. В спектре фрикативных звуков (особенно в спектре звонких фрикативных), иногда видны резонансные частоты. Однако решение обратной задачи опирается на характеристики шумовой компоненты спектра – частоты центра тяжести и нижней или верхней частоты среза шумовой компоненты [2]. На сегменте смычки обычно отсутствует излучение сигнала в области частот выше частоты основного тона. Распознавание смычки служит указанием на то, что в каком-то месте речевого тракта площадь поперечного сечения должны стать равной нулю на определенном интервале времени.

Казалось бы, что распознавание небольшого числа кардинальных сегментов должно быть более легкой задачей, чем распознавание фонетических элементов языка. Однако это не так. Акустические характеристики для каждого из представителей кардинального элемента настолько различаются, что для них невозможно отыскать инвариантные, контекстно-независимые признаки. Классификация какого-либо типа кардинальных сегментов чаще всего выполняется путем распознавания характерных для него фонетических элементов или артикуляторных состояний. Поэтому задача сегментации на кардинальные элементы является частным случаем задачи сегментации речевого сигнала на фонетически или артикуляторно значимые элементы.

Если бы удалось с высокой надежностью сегментировать речевой сигнал на фонетические элементы, то задача автоматического распознавания речи была бы значительно облегчена. Она свелась бы к перебору небольшого числа вариантов фонетических последовательностей с использованием лексической, синтаксической, семантической и прагматической избыточности речевого сообщения. Поэтому проблеме сегментации речевого сигнала постоянно уделяется большое внимание, хотя успехи в ее решении не слишком впечатляют.

Наиболее очевидный подход к различению сегментов гласных и согласных звуков речи состоит в использовании различия их акустической энергии. Так, в [3], сначала вычислялась функция громкости речевого сигнала путем взвешивания кратковременного спектра, а затем границы слогов устанавливались в момент времени, где разница между изменением громкости во времени и ее выпуклой оболочкой максимальна. Хотя такой подход и физически обоснован, но он применим далеко не всегда. Энергия фрикативных /Ш/ или /Ж/ может быть больше энергии некоторых гласных, а энергия назальной смычки может быть сопоставима с энергией гласных.

Поскольку простое сопоставление полной энергии речевого сигнала на различных участках не обеспечивает необходимой надежности классификации сегментов, то следующий шаг состоит в анализе спектрального состава сегмента. В [4] сегмент фрикативных звуков распознавался путем определения частоты центра тяжести спектра, длительности глухого участка, и степени изрезанности сглаженного спектра, вычисляемой как отношение максимальной производной по частоте на средней трети фрикативного к максимальной энергии на всем звукосочетании. При этом была достигнута точность детектирования 87% при хорошем отношении сигнал/шум SNR +20 dB, и 81% - при +10 dB.

Анализ статических характеристик сегмента, например, путем усреднения его спектра, далеко не всегда приводит к точной сегментации. Поэтому значительное внимание уделяется динамическим свойствам речевого сигнала. Простейший вариант использования динамики состоит в оценке различия амплитудного спектра в соседних кадрах, как, например, в [5]. Предлагается также использовать и фазовые различия в разных частотных областях [6], хотя и с не очень хорошими результатами. В частности, на базе данных для числительных английского языка, на 70% слогов базы была получена ошибка длительности слогов меньше 20%. Однако, учитывая, что средняя длительность слога составляла около 350 мс, то в абсолютных величинах ошибка была в пределах 70 мс.

Детектирование моментов быстрого изменения амплитудно-частотных характеристик речевого сигнала на границах глухой, звонкой и назальной смычек, а также на границах между гласными и согласными, было предпринято в [7], на основе анализа энергии в 6 частотных полосах. Для сигналов с отношением сигнал/шум 30 дБ, 73% границ находились на расстоянии не более чем 10 мс, а точность детектирования составляла около 90%.

В [8] смычка детектировалась по наличию взрыва, характеризующегося быстрым изменением энергии. Это изменение оценивалось оптимальным фильтром, содержавшим три компоненты: логарифм полной энергии сигнала в момент времени t, логарифм энергии в полосе частот выше 3 кГц, и Винеровской энтропии

$$W = \int \log\left(S\left(f,t\right)\right)df - \log\left(\int S\left(f,t\right)df\right), \tag{1.1}$$

где $S\left(f,t\right)$ – амплитудный спектр сигнала. Для базы данных TIMIT (отношение сигнал/шум 40 дБ), суммарная ошибка ложного пропуска и ложного срабатывания EER = 16%, а для аддитивного белого шума с отношением сигнал/шум 0 дБ – EER = 22%.

К сегментации применяется и популярная технология скрытых марковских моделей [6]. При этом достигается достаточно высокая относительно ручной разметки точность оценки границ согласный – согласный, гласный – согласный, согласный – гласный или гласный – гласный в китайском и японском языках. В базах данных для одного диктора средняя ошибка для японского языка составила 5.79 мс, а китайского – 6.61 мс. Это довольно хорошие оценки, однако нужно учитывать, что структура японского языка содержит последовательность звуков «согласный – гласный», а в китайском языке доля таких слогов очень велика. Такая структура существенно облегчает сегментацию.

Результаты, полученные для разных подходов к сегментации речи, во многих случаях имеют ограниченное значение в приложениях типа автоматического распознавания речи, поскольку использованный речевой материал сдержит «чистые» сигналы с высоким отношением сигнал/шум и для одного типа микрофонов. Этот недостаток пытаются преодолеть путем подмешивания различных шумов к исходным «чистым » сигналам. Так, в [9] к сигналам из базы данных TIMIT, телефонной базы данных NTIMIT, и базы данных для мобильного телефона CTIMIT добавлялся белый шум с уровнем 10 дБ или записанный шум в кабине автомобиля, в кабине пилота самолета или вертолета, а также шум вентилятора. Для метода скрытых марковских моделей с очисткой шума (стандартное или нелинейное вычитание спектров, Винеровская фильтрация) была получена точность определения границ с ошибкой менее 10 мс по базе данных TIMIT в 64.5% случаев, NTIMIT – в 63.8% случаев, CTIMIT – в 52.5% случаев, тогда как для чистого сигнала такая точность была достигнута в 86% случаев.

В [10] к анализу спектрограмм применялся такой же подход, как и к анализу зрительных образов: выделялись контуры, и оценивалась их ориентация на плоскости время – частота. При классификации на 52 фонетических сегмента для чистого речевого сигнала точность сегментации составила 58.5%, а для 8 обобщенных классов – 83.8%.

Из всего вышеперечисленного следует необходимость проведения дальнейшего исследования проблем поиска границ сегментов на речевом сигнале и распознавания кардинальных типов сегментов в интересах решения практических задач.

Исследования проводились на материале собранной в ИППИ РАН базы речевых данных русского языка, состоящей из образцов речи 47 дикторов. Каждый диктор произнес примерно по 1000 слов, входящих в словарь системы. В базе имеются изолированные, раздельные и слитные произнесения. В качестве примеров сигналов, содержавшихся в базе данных, приведём следующие: “ноль”, “один” ... “девять”; “десять”, ... “девяносто”; “сто”, ... “девятьсот”; “один – ноль – два – ноль – ... – девять – ноль”; “сто пятьдесят восемь тринадцать пятьдесят два”; “стоп”, “отмени”, “повтори”. При записи базы использовались два типа телефонных трубок и три типа микрофонов. Все произнесения базы размечены на фонетико-артикуляторные сегменты опытным лингвистом вручную. Алфавит разметки состоял из 127 артикуляторных и фонетических элементов. Частота дискретизации сигналов составляла 16 КГц. В исследовании участвовали сигналы с соотношением сигнал-шум от 12 дБ.

2. СЕГМЕНТАЦИЯ ПО УСРЕДНЕННОМУ НОРМИРОВАННОМУ СПЕКТРУ

В работе [1] рассматривались алгоритмы сегментации, основанные только на квазистационарных свойствах речевого сигнала. В соответствии со свойствами периферического отдела слухового анализатора человека, динамический спектр $S(f,t)$ речевого сигнала (так называемая сонограмма) преобразуется в шкалу мел, выполняется локальное взвешенное интегрирование в частотной области на интервалах 10 мел и 300 мел, а затем в каждый момент времени берется отношение этих спектров, обеспечивая локальную нормировку [11]:

$$G(f,t)=\log\left(\frac{F_2\int\limits_{f-F_1}^{f+F_1}S(\phi,t)\cdot d\phi}{F_1\int\limits_{f-F_2}^{f+F_2}S(\phi,t)\cdot d\phi}+1\right). \qquad (2.1)$$

Эта процедура нормировки соответствует известному из теории речевосприятия эффекту латерального торможения. В нормированной сонограмме $G(f,t)$ ярко выделена частотная структура: пики спектра обострены, зависимости от уровня сигнала и от наклона спектра по оси частот слабы (рис. 1).

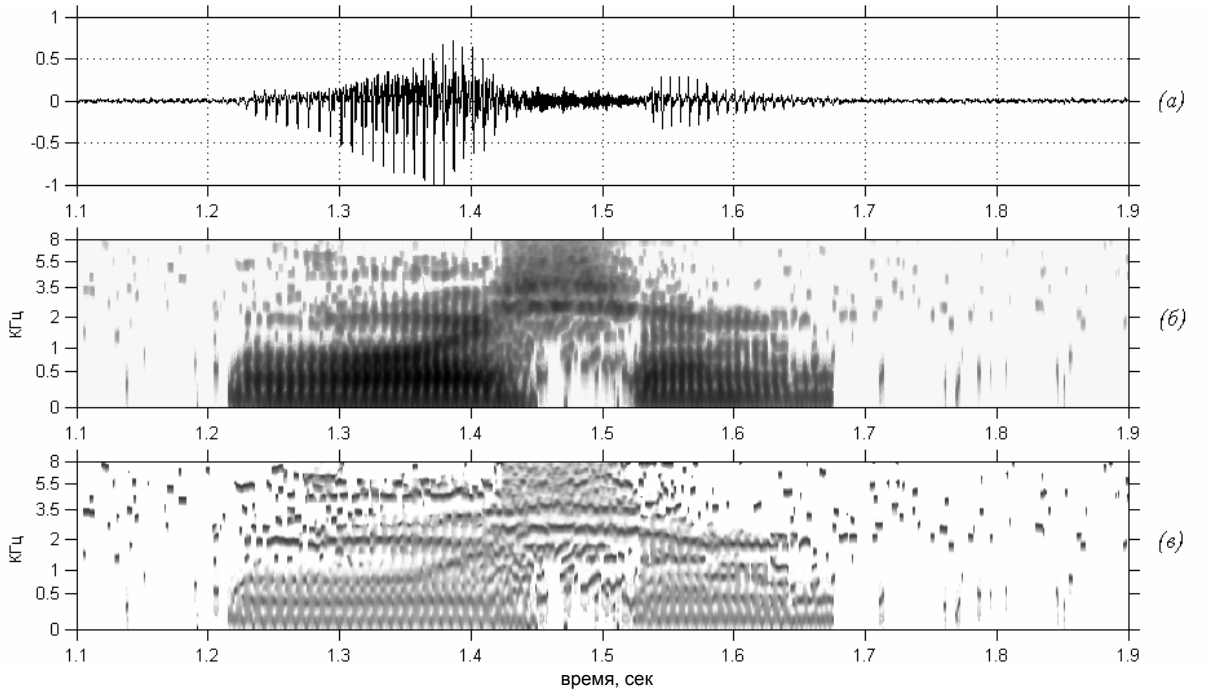

Рис. 1. Осциллограмма (а), сонограмма (б) и нормированная сонограмма (в) для слова «ВОСЕМЬ».

Свойства нормированной сонограммы позволяют использовать её для детектирования изменений частотных характеристик сигнала и поиска границ сегментов. Детектирование выполнялось с использованием меры L_2 расстояния между спектрами (средний квадрат разности). Мера L_2 связана с коэффициентом корреляции Коши-Буняковского $m_{КБ}$:

$$m_{L_2}(f_1,f_2)=2-2m_{КБ}=2-2\frac{\sum\limits_i f_1^i f_2^i}{\left(\sum\limits_i\left(f_1^i\right)^2\sum\limits_i\left(f_2^i\right)^2\right)^{1/2}}. \qquad (2.2)$$

Здесь f_1 и f_2 – сравниваемые неотрицательные функции. Мера L_2 не определена, если хотя бы одна из функций тождественно равна нулю. Для регуляризации меры L_2 вводится константа c_{reg}:

$$m_{\mathrm{L}_2}\left(f_1, f_2, c_{reg}\right) = 2 - 2\frac{\sum_i \left(f_1^i + c_{reg}\right)\left(f_2^i + c_{reg}\right)}{\left(\sum_i \left(f_1^i + c_{reg}\right)^2 \sum_i \left(f_2^i + c_{reg}\right)^2\right)^{1/2}}. \qquad (2.3)$$

Значение c_{reg} выбирается порядка значений сравниваемых функций. Для нормированного спектра c_{reg} = 10.

Алгоритм сегментации выполнял усреднение нормированного спектра $G\left(f,t\right)$ по времени начиная от последней найденной границы. На каждом шаге сонограммы t_i значение сегментирующей функции $\Psi\left(t_i\right)$ вычислялось как минимальное расстояние от усредненного нормированного спектра $\bar{G}\left(f,t_i\right)$ до спектров $G\left(f,t_j\right)$, $j = i+1,\ldots i+\tau_{delay}$:

$$\Psi\left(t_i\right) = \min_j m_{L_2}\left(\bar{G}\left(f,t_i\right), G\left(f,t_j\right)\right). \qquad (2.4)$$

Полученное значение $\Psi\left(t_i\right)$ использовалось для принятия решение об установке границы. Если $\Psi\left(t_i\right)$ превышало порог Ψ^*, то в граница устанавливалась в t_i, и накопление спектра начиналось снова. Поиск минимума расстояния на τ_{delay} производился для того, чтобы уменьшить влияние на $\Psi\left(t_i\right)$ кратковременной изменчивости спектра. Для этих же целей выполнялось предварительное сглаживание нормированного спектра по времени треугольной функцией длительностью 30 мс. В ходе исследования установлено оптимальное значение параметра τ_{delay} = 15 мс.

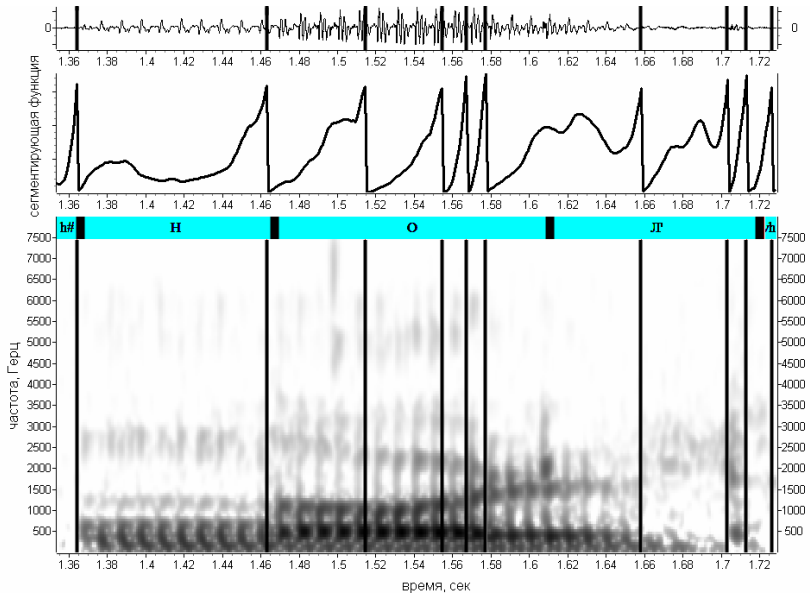

Рис. 2. Пример сегментации по усредненному нормированному спектру для слова «НОЛЬ».

Типичный вид сегментирующей функции $\Psi\left(t_i\right)$ при значении порога $\Psi^* = 1.3\cdot10^{-4}$ показан на рис. 2 для слова «НОЛЬ». Наверху показана осциллограмма сигнала, посередине – сегментирующая функция, внизу – сонограмма. Границы на осциллограмме и сонограмме расставлены автоматически. Сверху над сонограммой показана ручная разметка сигнала. Из

рисунка видно, что в данном случае достаточно точно найдены границы «пауза»-«Н» и «Н»-«О». Произошла потеря границы «О»-«Л», что связано со слабой выраженностью изменения частотной структуры на этой переходе (формантный состав не меняется, но происходит изменение амплитуды). Три лишние границы возникли на второй половине «О»: здесь происходит значительное изменение частотных характеристик сигнала. Расстояние между текущим и усредненным спектром быстро увеличивается, значение сегментирующей функции превышает порог и принимается решение об установке границы. Явление «пересегментации» на переходных участках характерно для алгоритма, использующего усредненный спектр.

Соотношение между числом лишних границ (вставок) и числом пропущенных границ можно менять, сдвигая значение порога Ψ^*. В таблице 1 приведены характеристики алгоритма сегментации по усредненному спектру в зависимости от величины порога Ψ^*: средняя погрешность определения положения границ, средний процент пропусков границ, среднее число вставок на каждый сегмент разметки. Тестирование выполнялось на материале базы речевых данных ИППИ.

Таблица 1. Характеристики алгоритма сегментации по усредненному спектру в зависимости от величины порога.

показатель	$\Psi^* = 0.8\cdot10^{-4}$	$\Psi^* = 1.3\cdot10^{-4}$	$\Psi^* = 1.8\cdot10^{-4}$
погрешность положения границ	3.87 мс	5.97 мс	8.30 мс
процент пропусков границ	0.95%	3.53%	13.09%
число вставок на сегмент разметки	2.60	1.04	0.30

Видно, что с увеличением порога уменьшается число лишних границ, но значительно увеличивается число пропущенных границ и падает точность. На рис. 3. показан результат сегментации того же слова «НОЛЬ» для более низкого (слева) и более высокого порогов (справа). Видно, что оба варианта имеют недостатки: при низком пороге число лишних границ неприемлемо высоко, а при высоком увеличивается погрешность и число пропусков.

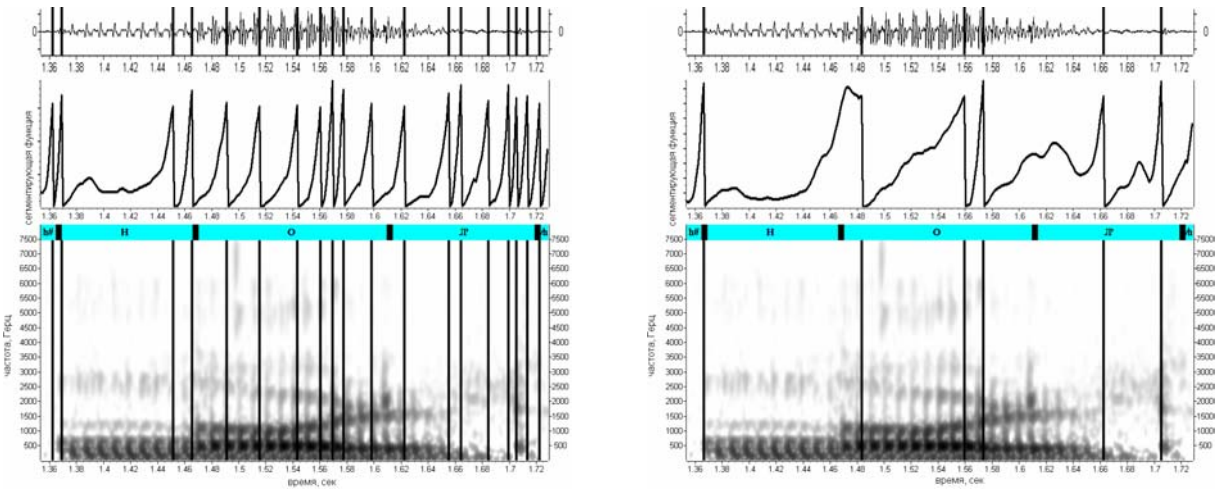

Рис. 3. Результат сегментации того же слова «НОЛЬ» для более низкого (слева) и более высокого порогов (справа).

Исследование показало, что основные недостатки алгоритма, использующего усредненный нормированный спектр, заключаются в неспособности найти границу между сегментами, если основное изменение характеристик сигнала происходит в амплитудной области. Логично попытаться использовать для сегментации спектр, в котором информация об уровне сигнала сохранена. Самый простой вариант – применить прежние правила сегментации к ненормированной сонограмме сигнала. Сегментирующая функция при этом будет равна

$$\Psi(t_i) = \min_j m_{L_2}\left(\overline{S}(f,t_i), S(f,t_j)\right). \qquad (2.5)$$

Были найдены значения параметров алгоритма, дающие хорошие результаты: τ_{delay} = 15 мс, c_{reg} = 1, Ψ^* = $3\cdot10^{-2}$. Работа алгоритма иллюстрируется рис. 4 для того же слова «НОЛЬ», что и на рис. 2. На рисунке хорошо видно, что «пересегментация» на переходном участке второй половицы «О» исчезла, однако появилось несколько лишних границ на «Л». Эти границы являются результатом высокой чувствительности к малым амплитудно-частотным изменениям при низком уровне сигнала. Эту чувствительность можно компенсировать увеличением c_{reg}, однако это приведет к увеличению потерь, что не желательно: даже в этом примере на рисунке граница «О»-«Л» найдена с очень большой ошибкой. Таким образом, вариант алгоритма сегментации по усредненному ненормированному спектру не является предпочтительным.

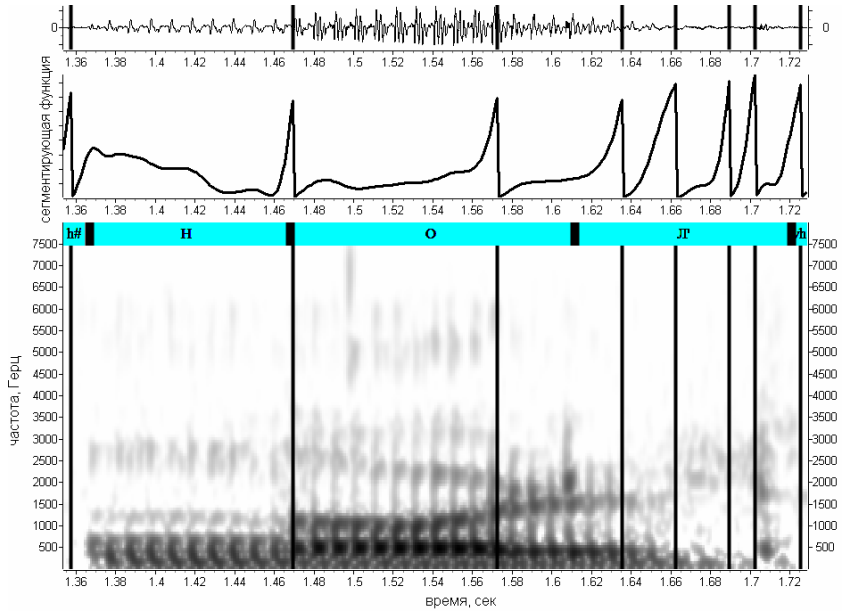

Рис. 4. Результат сегментации того же слова «НОЛЬ»
при использовании ненормированной сонограммы.

3. СЕГМЕНТАЦИЯ ПО ДИНАМИЧЕСКИМ ДЕТЕКТОРАМ

Был рассмотрен вариант использования сонограммы, обработанной так называемыми динамическими детекторами [11]:

$$D_+(f,t) = \max\left[0,\ +\log\frac{S(f,t+\Delta t/2)+c_{dyn}}{S(f,t-\Delta t/2)+c_{dyn}}\right]$$
$$D_-(f,t) = \max\left[0,\ -\log\frac{S(f,t+\Delta t/2)+c_{dyn}}{S(f,t-\Delta t/2)+c_{dyn}}\right] \qquad (3.1)$$

Здесь $D_+(f,t)$ – положительный динамический детектор, $D_-(f,t)$ – отрицательный динамический детектор, c_{dyn} – константа регуляризации, Δt – расстояние между спектрами относительно момента вычисления. Сонограмма $S(f,t)$ предварительно сглаживалась по времени Гауссовой функцией шириной 20 мс. Расстояние между спектрами Δt выбиралось равным ширине функции сглаживания, константа c_{dyn} = 1. Записанные в виде (3.1) положительный и отрицательный динамические детекторы выделяют на сонограмме места соответственно нарастания и спада энергии сигнала. Например, значение $D_+(f,t)$ тем больше, чем больше $S(f,t+\Delta t/2)$ относительно $S(f,t-\Delta t/2)$, и равно нулю, если $S(f,t+\Delta t/2) \le S(f,t-\Delta t/2)$. Вид спектров, полученных после обработки динамическими

детекторами (так называемых детектограмм) для того же слова «НОЛЬ», показан на рис. 5 снизу. Правая часть рисунка соответствует положительному детектору, левая – отрицательному. На детектограммах черные участки показывают повышение энергии (правый рисунок) или понижение (левый рисунок). Посередине показана суммарная энергия детекторов, вычисленная в полной полосе частот.

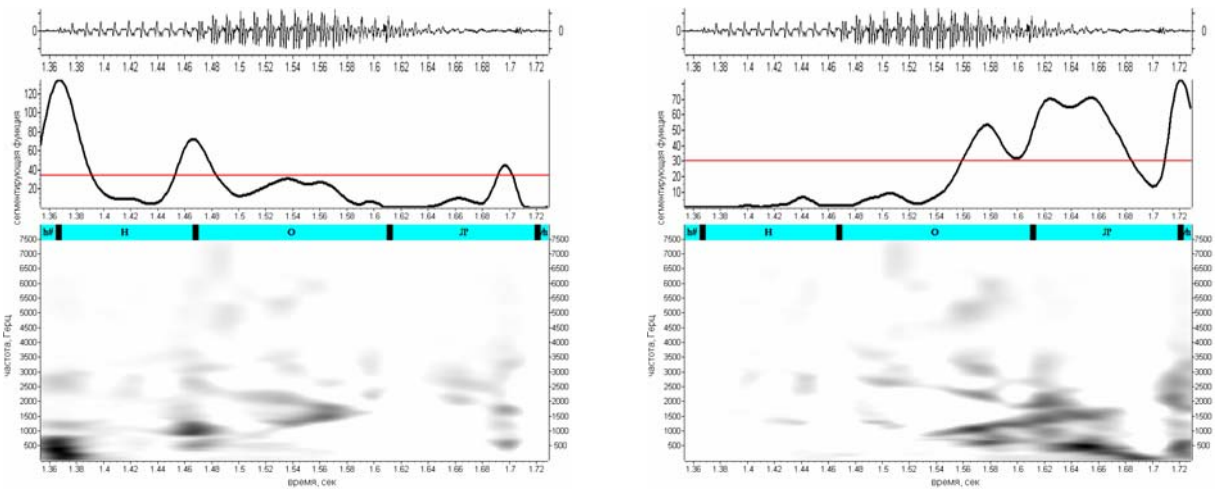

Рис. 5. Использование динамических детекторов на примере того же слова «НОЛЬ».
Справа – обработка положительным детектором, слева – отрицательным.

Анализ вида функций суммарной энергии показывает, что некоторые их пики в данном случае соответствуют эталонным границам сегментов. Для формализации алгоритма в данном случае требуется введение двух сегментирующих функций:

$$\Psi_+(t) = \int D_+(f,t)\,df, \quad \Psi_-(t) = \int D_-(f,t)\,df. \tag{3.2}$$

Для каждой функции должен быть найден порог, позволяющий ликвидировать низкие пики. Границы ставятся в моменты пиков обоих функций (3.2), превышающих порог. Из рис. 5 видно, что пороги должны быть примерно следующие: $\Psi_+^* = 35$, $\Psi_-^* = 30$. При этих порогах границы на рассматриваемом сигнале будут найдены правильно, число лишних границ будет невелико. Однако, оказывается, встречаются сигналы, для которых такой алгоритм сегментации будет работать некорректно.

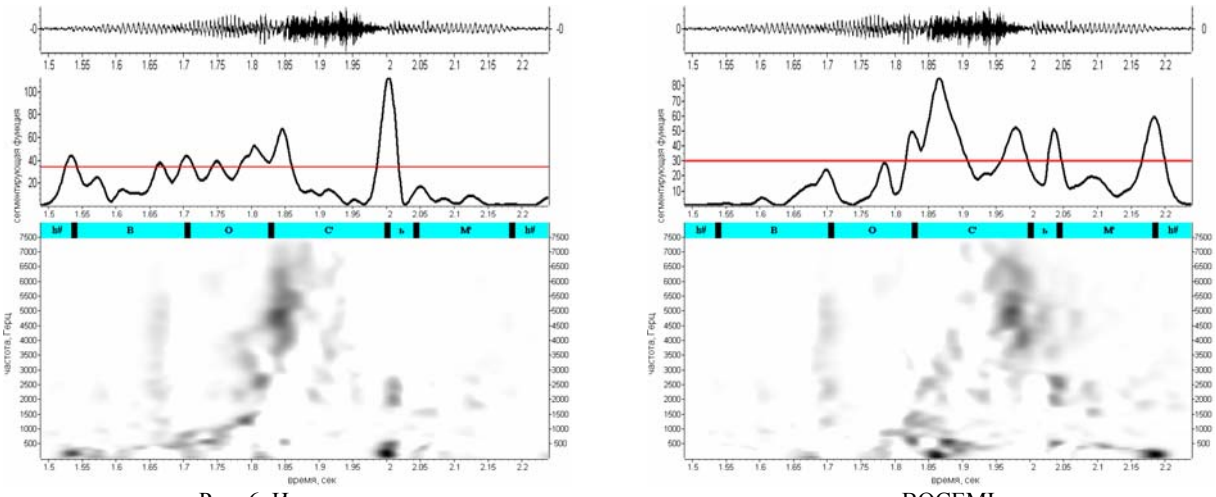

Рис. 6. Использование динамических детекторов на примере слова «ВОСЕМЬ».
Справа – обработка положительным детектором, слева – отрицательным.

На рис. 6 показан результат обработки динамическими детекторами сигнала, содержащего произнесение слова «ВОСЕМЬ». Из рисунка видно, что, например, функция $\Psi_+(t)$ при заданных значениях порогов, приведет к возникновению большого числа ложных границ. При этом порог Ψ_+^* не может быть повышен, иначе возникает риск потери начала слова и других

границ. Этот пример заставляет сделать вывод об ограниченности рассматриваемого алгоритма сегментации по динамическим детекторам.

4. СЕГМЕНТАЦИЯ ПО КОРРЕЛЯЦИИ МЕЖДУ РАВНООТСТОЯЩИМИ СПЕКТРАМИ

На основании проведенных экспериментов и наблюдений был синтезирован алгоритм, адекватно обрабатывающий вышеописанные затруднительные случаи.

Сегментирующая функция вычислялась следующим образом:

$$\Psi(t) = m_{L_2}\left(S(f, t-\Delta t/2), S(f, t+\Delta t/2), c_{reg}\right). \quad (4.1)$$

Здесь $S(f,t)$ – ненормированная сонограмма в шкале Герц, сглаженная по времени Гауссовой функцией шириной 15 мс, m_{L_2} – мера расстояния между спектрами с регуляризующей константой c_{reg}, Δt – расстояние между спектрами относительно момента вычисления. Эксперименты показали, что Δt должно быть равно ширине функции сглаживания, то есть Δt = 15 мс, значение константы c_{reg} = 5.

Определенная так сегментирующая функция инвариантна к квазистационарным шумам, и имеет большие значения в моменты, когда происходит значительное изменение амплитудно-частотных характеристик сигнала. Пики этой функции указывают на возможные положения границ. При этом длительные переходные участки сигнала не будут насекаться на большое число сегментов, поскольку, несмотря на большие значения $\Psi(t)$, на этих участках не будет содержаться лишних пиков.

Для установки границ пики сегментирующей функции должны подвергнуться дополнительному анализу. Самый простой способ анализа – сравнение значения пика функции $\Psi(t)$ с неким порогом: если пик $\Psi(t)$ превышает порог, граница ставится. Однако практика показывает, что этого оказывается недостаточно, и легко найти пример случая, когда высота пика, на котором должна быть установлена граница, заведомо ниже возможных значений порога. Следовательно, требуется более сложный способ анализа.

Наибольшую эффективность показал рекуррентный способ анализа пиков сегментирующей функции. Пусть известно положение n-й границы, соответствующее одному из пиков функции $\Psi(t)$, требуется определить положение следующей, $n+1$-й границы. Рассмотрим первый локальный минимум $\Psi(t_1)$, расположенный правее n-й границы. Найдем ближайший к нему минимум $\Psi(t_2)$, такой, что $t_1 < t_2$, а мера расстояния между спектрами, взятыми в положениях этих минимумов, превышает некий порог m^*:

$$m_{L_2}\left(S(f,t_1), S(f,t_2), c_{reg}\right) > m^*. \quad (4.2)$$

В этом случае следует считать, что между t_1 и t_2 произошло значительное изменение характеристик сигнала, и, следовательно, где-то между этими минимумами должна быть установлена граница. Согласно свойствам сегментирующей функции, граница должна быть установлена на одном из пиков $\Psi(t)$. Если между t_1 и t_2 находится только один пик, решение тривиально. Если между t_1 и t_2 лежит несколько пиков, необходимо выполнить дополнительный анализ.

Ясно, что граница должна быть установлена в момент пика функции $\Psi(t)$. Но между t_1 и t_2 может лежать несколько пиков (см. рис. 7). Для выбора правильного пика найдем самый правый минимум $\Psi(t_3)$, лежащий между t_1 и t_2, такой что

$$m_{L_2}\left(S(f,t_3), S(f,t_2), c_{reg}\right) > m^*,\ t_1 \le t_3 < t_2. \quad (4.3)$$

В общем случае t_3 может совпадать с t_1. Таким образом, найден более узкий интервал поиска границы. Если между t_3 и t_2 по прежнему лежит несколько пиков, граница ставится по наивысшему пику, так как он указывает на наибольшее изменение свойств сигнала.

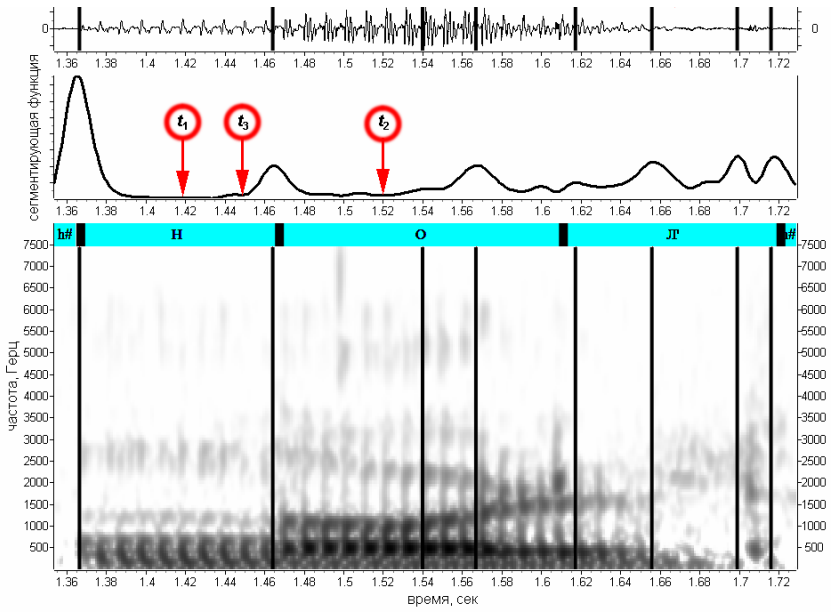

Рис. 7. Результат сегментации по корреляции

между равноотстоящими спектрами для слова «НОЛЬ».

В процессе расстановки границ происходит уточнение границ начала взрывов. Для этого построен простейший детектор переходов от смычных сегментов к взрывам. Сначала проверялось условие на смычку: энергия сегмента выше 1300 Гц должна быть ниже некого порога. Затем проверялось условие на взрыв: на конце сегмента должно наблюдаться резкое нарастание энергии в достаточно широкой полосе частот. При выполнении обоих условий граница окончания сегмента смычки устанавливалась на момент начала фронта нарастания энергии, то есть на начало взрыва.

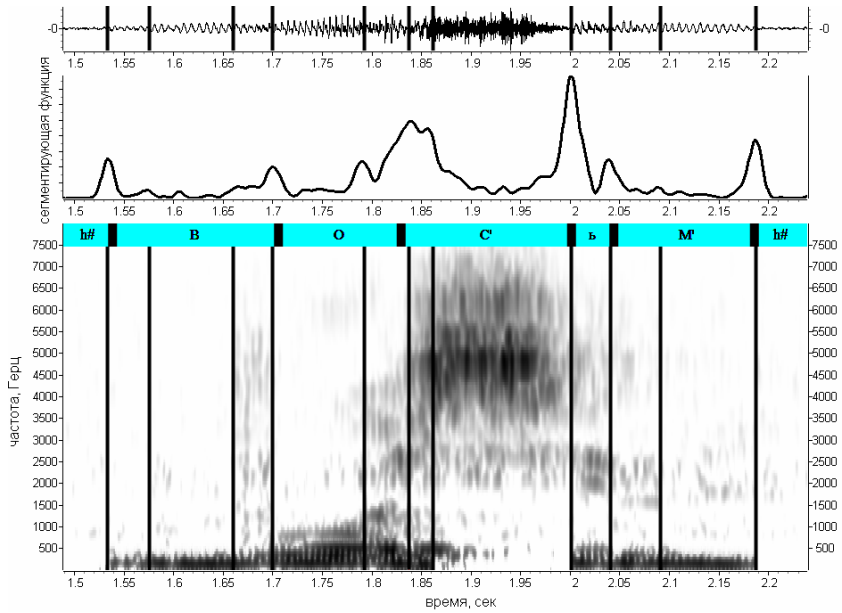

Рис. 8. Результат сегментации по корреляции

между равноотстоящими спектрами для слова «ВОСЕМЬ».

Построенный таким образом алгоритм оказывается предпочтительным для использования в практических речевых задачах. На рис. 7 и 8 показаны результаты сегментации упоминавшихся выше слов «НОЛЬ» и «ВОСЕМЬ». Видно, что границы артикуляторно-акустических сегментов

найдены с малой погрешностью. Лишние границы соответствуют переходным и квазистационарным участкам, и их число приемлемо.

Подробный анализ результатов сегментации будет приведен ниже в разделе «Результаты».

5. КЛАССИФИЦИРУЮЩИЕ АКУСТИЧЕСКИЕ ПАРАМЕТРЫ

Задача распознавания кардинальных элементов речи при указанных границах сегментов требует для своего решения использования специальных классифицирующих характеристик. Эксперименты показали, что никакой отдельно взятый акустический параметр не позволяет распознавать, например, тип сегментов «назальные» или любой другой. Дело в том, что «назальные» отличаются от «фрикативных» по одному набору параметров, а от «гласноподобных» по другому набору. Таким образом, каждый раз надо решать задачу дихотомии, поскольку для разделения каждой пары типов эффективны свои параметры.

Каждый сегмент речевого сигнала характеризуется определенными спектрально-временными параметрами. Их можно разделить на качественно разные группы: параметры, определяющиеся формой спектра, изменением энергии во времени, степенью периодичности и т. д. Рассмотрим эти группы подробнее. Будем считать при этом, что сигнал разбит на сегменты так, что на протяжении одного сегмента акустические параметры меняются незначительно.

В соответствии со свойствами восприятия, при вычислении акустических параметров используется логарифмический спектр в шкале мел.

Наклон спектра, усредненного на сегменте, вычисляется посредством аппроксимации отчетов спектра (x_i, y_i) прямой линией $a \cdot x_i + b$, $i = 1 \ldots n$. В соответствии с методом наименьших квадратов коэффициенты a и b находятся как

$$a = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^{\,2} - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}, \; b = \frac{\frac{1}{n}\left(\sum_{i=1}^{n} x_i^{\,2} \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i\right)}{\sum_{i=1}^{n} x_i^{\,2} - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}. \qquad (5.1)$$

Частотные отсчеты x_i могут браться в шкале Герц или мел, а отсчеты спектра y_i могут браться по энергетическому, амплитудному или логарифмированному спектру. Инвариантность к полной энергии при вычислении наклона достигается нормированием по площади к единице. Для вычисления наклона могут использоваться отчеты спектра, взятые из разных частотных диапазонов. В зависимости от этого наклон будет описывать различные явления. Например, наклон спектра в полном диапазоне будет характеризовать тип источника, голосовой или турбулентный: известно, что спектр голосового источника спадает с наклоном примерно на 12 dB на октаву, а спектр турбулентного источника, напротив, возрастает к высоким частотам. В спектре высоких гласных (например, «И», «Я», «Е») между первой и второй формантами существует провал, который может оказать негативное влияние на классифицирующие свойства наклона спектра. Поэтому представляется целесообразным исследовать наклоны, вычисленные на спектре с исключенным диапазоном средних частот.

Пересечение спектра с аппроксимирующей наклонной линией (рис. 9) позволяет оценить границы частотных интервалов, в которых сосредоточена энергия сигнала. Для вычисления частот, на которых происходит пересечение, спектр следует предварительно сгладить, чтобы ликвидировать резонансную структуру. Используется три пересечения в диапазоне от 0,3 до 7 кГц: верхняя частота первого (низкочастотного) пика, нижняя частота последнего (высокочастотного) пика, верхняя частота последнего пика. В тех случаях, когда в спектре присутствует только один пик, считается, что первый и последний пики совпадают. Заметим, что использование для вычисления пересечений вместо наклонной линии горизонтальной средней линии менее предпочтительно. Действительно, при наличии большого перепада энергии между первым и последним пиками спектра средняя линия может вообще не пересечь пик с меньшей амплитудой, и содержащаяся в нем информация будет утрачена.

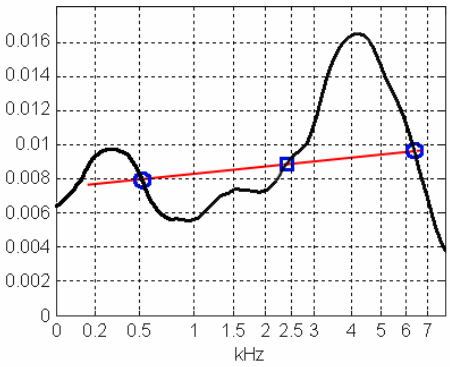

Рис. 9. Пересечение спектра с аппроксимирующей наклонной линией.
Логарифмический спектр, шкала частот в мелах.
Обозначения: ○ – верхняя граница пика, □ – нижняя граница пика.

Центры тяжести сглаженного спектра вычисляются отдельно для двух пиков спектра, найденных по вышеописанному алгоритму с использованием пересечений с наклонной линией. Положения этих центров позволяют различать гласные и фрикативные звуки. Заметим, что использование центра тяжести спектра эквивалентно использованию количества переходов через нуль (zero-crossing rate). Положения центров тяжести и пересечений спектра могут участвовать не только в детектировании кардинальных типов сегментов, но и в распознавании фрикативных звуков («С», «Ш», «Ф», «Х»).

Баланс энергий между различными частотными интервалами на спектре сегмента также может быть информативен для распознавания кардинальных типов. Например, для различения назальных и гласных можно использовать тот факт, что у назальных, как правило, в низкочастотной области (до 350 Гц) сосредоточено больше энергии, чем у гласных. Положение интервалов и способ учета энергии в них подлежат уточнению. Различаются два способа вычисления баланса. Первый способ: вычисляется разность энергий в двух различных полосах и нормируется на полную энергию. Второй способ: энергия в одной полосе делится на суммарную энергию в двух исследуемых полосах. Отношение амплитуды низкочастотного пика к общей энергии также может оказаться информативным для распознавания назальных звуков.

Винеровская энтропия в нескольких работах [8, 12-13] оказывается информативной для распознавания смычных звуков. В определении (1.1) энтропия вычисляется на всем частотном диапазоне. Наша задача состоит в определении эффективности её использования для распознавания кардинальных типов при вычислении в различных диапазонах частот.

Кепстральные коэффициенты являются результатом применения обратного преобразования Фурье к логарифмированному энергетическому спектру. Кепстральные коэффициенты младшего порядка широко используются в речевых задачах, поскольку они содержат информацию об общем виде спектра. Для их вычисления может использоваться энергетический спектр в частотной шкале Герц или мел. Нулевой коэффициент содержит информацию о средней энергии спектра и не используется на практике как не характеризующий форму спектра.

Отношение энергии текущего сегмента к максимальной энергии сигнала на некотором временном интервале вокруг сегмента является динамической характеристикой. Принимая во внимание физиологические свойства слуха, в частности, эффект латерального торможения, можно ожидать, что при надлежащем выборе частотной полосы и длительности временного интервала эта характеристика увеличит надежность разделения смычных, фрикативных и гласных звуков.

Все акустические параметры должны обладать инвариантностью по отношению к коэффициенту усиления канала и среднему уровню сигнала. Для этого необходимо выполнить **предварительное нормирование** сигнала к его энергии на длительном временном интервале. В работе использовалась следующая процедура нормирования. Полная энергия сигнала, взятая в децибелах, слаживалась треугольной функцией длительностью в 1 секунду. Назовём полученную после сглаживания функцию $E_{smoothA}$. Аналогично выполнялось сглаживание на более коротком интервале 200 мс, с вычислением функции $E_{smoothB}$. Затем производился поиск

пиков функции $E_{smoothB}$ на тех интервалах речи, где эта функция превышала $E_{smoothA}$. Вершины соседних пиков соединялись отрезками. Полученная непрерывная кусочно-линейная функция использовалась для нормирования каждого временного отсчета исходной спектрограммы: значение функции переводилось из децибел в энергетические единицы и выступало делителем для профиля энергетического спектра S. Для последующей работы с логарифмическим спектром $S_{\log} = \log(S+1)$ выполнялось умножение на единую для всех профилей константу, обеспечивающую работу в удобном диапазоне функции логарифма. Экспериментально установлено, что оптимальное значение этой константы примерно равно 10^4.

Мера периодичности вычисляется по автокорреляционной функции, нормированной к нулевому отсчету, как значение этой функции на смещении, равном периоду голосовых импульсов. Также для вычисления меры периодичности целесообразно использовать нормированную к накопленному среднему разностную функцию. Мера периодичности может вычисляться в различных частотных диапазонах. Для этого сигнал должен быть предварительно пропущен через полосовой фильтр. При этом можно ожидать, что мера периодичности, вычисленная в области низких частот, будет описывать наличие голосового источника: чем выше периодичность, тем вероятнее наличие огласованности. Напротив, вычисленная в области средних и высоких частот мера периодичности будет описывать наличие шумового источника: чем ниже периодичность, тем вероятнее наличие турбулентности.

6. ИССЛЕДОВАНИЕ ЭФФЕКТИВНОСТИ ПАРАМЕТРОВ

В рамках каждой из вышеописанных групп существуют различные способы вычисления акустических параметров. Например, баланс энергий можно вычислять в разных частотных диапазонах несколькими способами. Чтобы ограничить количество параметров, ограничимся рассмотрением только тех частотных полос, которые адекватны свойствам различных типов речевых звуков.

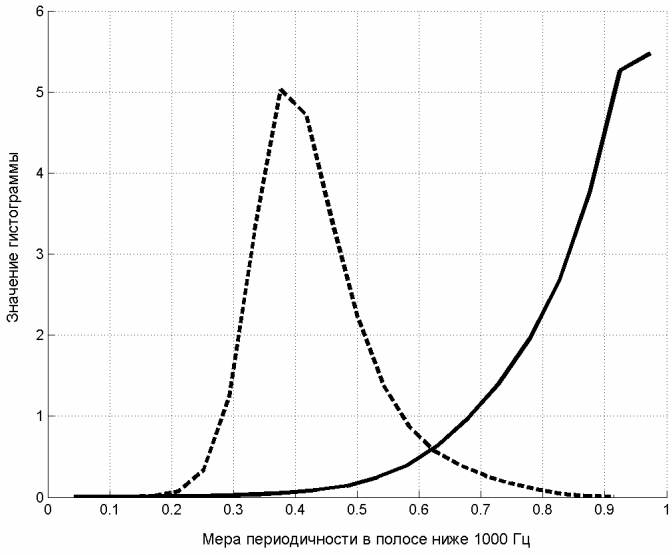

Рис. 10. Пересечение гистограмм меры периодичности в полосе ниже 1000 Гц для пары «гласные» (сплошная линия) и «фрикативные глухие» (штриховая линия). Пересекаемость гистограмм: 10%.

Эффективность различных акустических параметров для разделения кардинальных типов оценивается путем противопоставления пар различных типов. Пусть, например, требуется исследовать разделимость «гласных» и «фрикативных глухих» по выбранному акустическому параметру. Для всех сегментов, принадлежащих к первому классу («гласные») вычисляются значения этого акустического параметра, и строится гистограмма их распределения. То же проделывается для всех сегментов, принадлежащих ко второму классу («фрикативные глухие»), строится вторая гистограмма. Обе гистограммы нормируются по площади к единице. Площадь взаимного пересечения этих двух гистограмм и будет критерием разделимости этой

пары классов по данному параметру. Чем меньше площадь, тем эффективнее данный параметр разделяет эту пару классов. В идеальном случае эта площадь равна нулю. На рис. 10 показан примерный вид пересекающихся гистограмм значений акустического параметра для пары типов.

В таблице 2 для разных групп акустических параметров, эффективность которых необходимо исследовать, перечислены варианты выбора частотных полос. Всего в эксперименте участвует 54 параметра.

Таблица 2. Группы и варианты вычисления акустических параметров в зависимости от частотной полосы.

Наклон спектра	
S_{full}	полная полоса частот
S_{low}	в полосе от 0 до 2000 Гц
S_{high}	в полосе от 2000 до 6000 Гц
S_1	в полосе от 0 до 750 и от 2000 до 8000 Гц
S_2	в полосе от 0 до 750 и от 2000 до 4000 Гц
S_3	в полосе от 300 до 6000 Гц
S_4	в полосе от 300 до 750 и от 2000 до 6000 Гц
S_5	в полосе от 300 до 750 и от 2000 до 4000 Гц
S_6	в полосе от 750 до 6000 Гц

Пересечение спектра	
наклонная линия, вычисленная в полосе от 300 до 6000 Гц	
C_{11}	верхняя частота первого пика
C_{12}	нижняя частота последнего пика
C_{13}	верхняя частота последнего пика
наклонная линия, вычисленная в полосе от 750 до 6000 Гц	
C_{21}	верхняя частота первого пика
C_{22}	нижняя частота последнего пика
C_{23}	верхняя частота последнего пика
наклонная линия, вычисленная в полосе от 0 до 750 и от 2000 до 8000 Гц	
C_{31}	верхняя частота первого пика
C_{32}	нижняя частота последнего пика
C_{33}	верхняя частота последнего пика

Центры тяжести	
наклонная линия, вычисленная в полосе от 300 до 6000 Гц	
M_{11}	центр тяжести первого пика
M_{12}	центр тяжести последнего пика
наклонная линия, вычисленная в полосе от 750 до 6000 Гц	
M_{21}	центр тяжести первого пика
M_{22}	центр тяжести последнего пика
наклонная линия, вычисленная в полосе от 0 до 750 и от 2000 до 8000 Гц	
M_{31}	центр тяжести первого пика
M_{32}	центр тяжести последнего пика

Баланс энергий	
Ba_1	первый способ вычисления, полосы: 0-600, 600-1500 Гц
Ba_2	первый способ вычисления, полосы: 0-750, 750-2000 Гц
Ba_3	первый способ вычисления, полосы: 600-1500, 1500-8000 Гц
Ba_4	первый способ вычисления, полосы: 750-2000, 2000-8000 Гц
Ba_5	первый способ вычисления, полосы: 0-600, 1500-8000 Гц
Ba_6	первый способ вычисления, полосы: 0-750, 2000-8000 Гц
Bb_1	второй способ вычисления, полосы: 0-400, 400-8000 Гц
Bb_2	второй способ вычисления, полосы: 0-400, 1500-8000 Гц
Bb_3	второй способ вычисления, полосы: 0-750, 750-8000 Гц
Bb_4	второй способ вычисления, полосы: 0-750, 2000-8000 Гц
Bc_1	отношение амплитуды высшего пика в полосе 0-350 к общей энергии
Bc_2	отношение амплитуды высшего пика в полосе 0-750 к общей энергии

Винеровская энтропия	
W_{full}	полная полоса частот
W_{low}	в полосе от 0 до 2000 Гц
W_{high}	в полосе от 2000 до 6000 Гц

Кепстральные коэффициенты	
ceps1	первый коэффициент кепстра
ceps2	второй коэффициент кепстра
ceps3	третий коэффициент кепстра
ceps4	четвертый коэффициент кепстра

Отношение энергии	
E_{11}	в полосе от 0 до 350 Гц, на интервале ±250 мс
E_{12}	в полосе от 0 до 750 Гц, на интервале ±250 мс
E_{13}	в полосе от 500 до 8000 Гц, на интервале ±250 мс
E_{21}	в полосе от 0 до 350 Гц, на интервале ±500 мс
E_{22}	в полосе от 0 до 750 Гц, на интервале ±500 мс
E_{23}	в полосе от 500 до 8000 Гц, на интервале ±500 мс
E_{31}	в полосе от 0 до 350 Гц, на интервале ±1 сек
E_{32}	в полосе от 0 до 750 Гц, на интервале ±1 сек
E_{33}	в полосе от 500 до 8000 Гц, на интервале ±1 сек

Мера периодичности	
P_{full}	полная полоса частот
P_{low}	в полосе ниже 1000 Гц
P_{high}	в полосе выше 1000 Гц

В ходе эксперимента на материале речевой базы данных ИППИ были вычислены критерии эффективности для каждого параметра. Было рассмотрено 15 пар, соответствующих различным комбинациям шести кардинальных типов. В таблице 3 приведены восемь наилучших параметров для различных пар, в скобках указаны значения критерия эффективности для них. В таблице используются следующие сокращения: «Г» – гласноподобный, «Н» – назальный, «Фг» – фрикативный глухой, «Фз» – фрикативный звонкий, «Сг» – смычной глухой, «Сз» – смычной звонкий. Из таблицы видно, что хорошую разделимость имеют пары «Г»–«Фг», «Г»–«Сг», «Г»–«Сз», «Н»–«Фг», «Фг»–«Сг», «Фг»–«Сз». Плохо разделяются пары «Г»–«Н», «Г»–«Фз», «Н»–«Фз», «Н»–«Сз», «Фг»–«Фз», «Сг»–«Сз», чего следовало ожидать.

Таблица 3. Список восьми параметров, показавших наибольшую эффективность для распознавания различных пар.

пара	1	2	3	4	5	6	7	8
<Г>-<Н>	E23 (24)	Wfull (36)	Bc1 (40)	Bb1 (41)	E22 (42)	S2 (45)	S1 (45)	Ba1 (54)
<Г>-<Фг>	E22 (8)	S3 (8)	ceps1 (9)	Pfull (9)	E21 (9)	Plow (10)	S1 (11)	Ba3 (15)
<Г>-<Фз>	E22 (27)	ceps1 (30)	E21 (40)	C11 (40)	Pfull (40)	M21 (42)	M11 (48)	S3 (49)
<Г>-<Сг>	E23 (4)	E22 (6)	E21 (8)	Wfull (9)	Plow (12)	Pfull (17)	ceps1 (37)	Bc1 (43)
<Г>-<Сз>	E23 (6)	Wfull (11)	Bc1 (14)	E22 (18)	Bb1 (18)	S2 (22)	S1 (23)	Ba1 (24)
<Н>-<Фг>	S3 (4)	S1 (4)	Ba3 (6)	Bb4 (9)	ceps1 (10)	Bc1 (13)	Plow (14)	S2 (14)
<Н>-<Фз>	S3 (30)	Ba3 (32)	ceps1 (36)	S1 (42)	Bb4 (43)	Pfull (48)	Wfull (50)	Bb1 (52)
<Н>-<Сг>	Plow (16)	E21 (20)	E22 (22)	Pfull (25)	E23 (33)	Wfull (37)	ceps1 (45)	S3 (71)
<Н>-<Сз>	E23 (44)	Wfull (49)	C11 (51)	Bc1 (52)	Plow (55)	E22 (56)	Ba1 (57)	S2 (59)
<Фг>-<Фз>	Plow (20)	S1 (27)	Bc1 (30)	Ba1 (33)	E21 (33)	Bb4 (33)	S3 (36)	S2 (39)
<Фг>-<Сг>	E23 (7)	Wfull (12)	S1 (23)	S3 (24)	Ba3 (25)	Bb4 (26)	Bc1 (27)	S2 (32)
<Фг>-<Сз>	S1 (4)	Bb4 (5)	Bc1 (5)	S3 (6)	Ba3 (7)	S2 (7)	Ba1 (10)	E23 (10)
<Фз>-<Сг>	Wfull (14)	E23 (14)	Plow (22)	E21 (27)	E22 (29)	Pfull (48)	Ba3 (56)	Bc1 (56)
<Фз>-<Сз>	Wfull (18)	E23 (20)	S1 (24)	Bb4 (25)	Bb1 (25)	Bc1 (27)	Ba3 (29)	S2 (29)
<Сг>-<Сз>	E21 (40)	Plow (43)	E22 (46)	Bb1 (52)	Pfull (54)	S1 (58)	S2 (58)	Ba1 (61)

Из таблицы 3 видно, что схожие по способу вычисления варианты параметров, часто имеют близкие значения критерия. Это говорит о том, что такие параметры содержат идентичную информацию о сегментах, и использование их совместно не приведет к улучшению разделимости. Из этого следует, что по результатам тестирования эффективности каждого параметра отдельно невозможно установить для каждой пары типов набор параметров, дающий наилучшую различимость. Для этого требуется особая процедура, которая позволила бы оценить эффективность целого набора параметров сразу. Для создания такой процедуры необходимо использовать специфические методы принятия решений и моделирования плотности вероятности многомерных распределений.

7. БАЙЕСОВСКОЕ РЕШАЮЩЕЕ ПРАВИЛО

В системах принятия решений широко используется Байесовское решающее правило. Байесовский подход основан на предположении, что плотности распределения каждого из классов известны заранее. Это предположение позволяет выписать искомый алгоритм в явном аналитическом виде. Этот алгоритм является оптимальным при определенных условиях, то есть обладает минимальной вероятностью ошибочной классификации. Условия, в которых рассматривается задача распознавания кардинальных элементов, позволяют сформулировать решающее правило в виде метода максимума апостериорной вероятности:

$$a(x) = \arg\max_{y \in Y} \mathrm{P}(y \mid x) \tag{7.1}$$

Здесь $\mathrm{P}(y \mid x)$ – апостериорная вероятность класса y из набора классов Y при условии, что наблюдается вектор параметров x, $a(x)$ – принимаемое решение из набора Y.

Необходимо отметить, что распознавание кардинальных элементов в автоматическом режиме не может быть выполнено со стопроцентной надежностью. Ошибки в типе сегмента могут привести к фатальным ошибкам в решении обратной задачи (например, к неправдоподобной форме речевого тракта). Отсюда следует необходимость решения обратной задачи в предположении, что данный сегмент речевого сигнала может принадлежать любому из рассматриваемых типов. Выбор окончательного решения должен выполняться с учетом апостериорных вероятностей для всех кардинальных типов на основе сравнения полученных невязок заданных и вычисленных акустических параметров, а также по усилиям, необходимым для формирования найденного вектора артикуляторных параметров при переходе от предыдущего сегмента. Таким образом, задача распознавания кардинальных типов сводится к определению апостериорных вероятностей.

Апостериорная вероятность вычисляется по формуле

$$\mathrm{P}(y \mid x) = \frac{P_y p(x \mid y)}{\sum_{u \in Y} P_u p(x \mid u)}. \tag{7.2}$$

Здесь P_y – априорная вероятность появления класса y, $p(x \mid y)$ плотность распределения класса y, в знаменателе сумма по всем возможным классам и $u \in Y$ – индекс суммирования. В предыдущем разделе было показано, что для разделения различных пар типов сегментов эффективны различные акустические параметры. Следовательно, плотности вероятности для каждой пары должны вычисляться в различных пространствах. Это приводит к необходимости применять формулу апостериорной вероятности отдельно к каждой паре типов:

$$\mathrm{P}_b(a \mid x) = \frac{1}{1 + \dfrac{P_b p(x \mid b)}{P_a p(x \mid a)}}. \tag{7.3}$$

Здесь $\mathrm{P}_b(a \mid x)$ – вероятность того, что данный сегмент является сегментом типа a, вычисленная при противопоставлении типов a и b, x – измеренный вектор параметров,

$p(x|a)$ и $p(x|b)$ – плотности вероятности для вектора x в пространстве типов a и b соответственно, P_a и P_b – априорные вероятности появления типов.

8. МОДЕЛИРОВАНИЕ ПЛОТНОСТИ РАСПРЕДЕЛЕНИЯ ПО ВЫБОРКЕ

На практике плотности распределения классов, участвующие в вычислении апостериорной вероятности, как правило, неизвестны. Их приходится оценивать (восстанавливать) по имеющейся обучающей выборке. При этом байесовский алгоритм перестаёт быть оптимальным, так как плотность распределения можно восстановить по конечной выборке только с некоторой погрешностью. Тем не менее, в условиях рассматриваемой задачи оказалось возможным построить работоспособный алгоритм классификации при моделировании плотностей распределений акустических параметров с использованием смесей нормальных распределений:

$$p(x)=\sum_{j=1}^{k} w_j \varphi(x;\theta_j),\ \sum_{j=1}^{k} w_j = 1, \tag{8.1}$$

где $\varphi(x;\theta_j)$ – функция распределения многомерного аргумента x с параметрами θ_j, w_j – её вес, k – количество компонент в смеси. Как известно, формула для n-мерного нормального распределения записывается так:

$$\varphi(x;\theta_j) \equiv p(x|\mu_j,R_j)=\frac{1}{(2\pi)^{n/2}\left|R_j\right|^{1/2}} e^{-\frac{1}{2}(x-\mu_j)^T R_j^{-1}(x-\mu_j)},\ x\in\mathbb{R}^n. \tag{8.2}$$

Здесь $\mu_j \in \mathbb{R}^n$ – вектор математического ожидания j-й компоненты смеси, $R_j \in \mathbb{R}^{n\times n}$ – ковариационная матрица.

Для оценивания параметров смеси $\Theta=(w_1,\ldots,w_k,\mu_1,\ldots,\mu_k,R_1,\ldots,R_k)$ по известной выборке $X^m=\{x_1,\ldots,x_m\}$ используется метод максимума правдоподобия. Метод состоит в том, чтобы найти значение вектора параметров, при котором наблюдаемая выборка наиболее вероятна. На практике производится максимизация логарифма функции правдоподобия:

$$L(X^m;\Theta)=\sum_{i=1}^{m}\ln\sum_{j=1}^{k} w_j \varphi(x_i;\theta_j)\to\max_{\theta}. \tag{8.3}$$

Попытка применить принцип максимума правдоподобия «в лоб» приводит к чрезмерно громоздкой оптимизационной задаче. Обойти эту трудность позволяет алгоритм EM (expectation-maximization). Идея EM-алгоритма заключается в следующем. Вводится вспомогательный вектор скрытых переменных G, обладающий двумя замечательными свойствами. С одной стороны, он может быть вычислен, если известны значения вектора параметров Θ. С другой стороны, поиск максимума правдоподобия сильно упрощается, если известны значения скрытых переменных.

EM-алгоритм состоит из итерационного повторения двух шагов. На E-шаге вычисляется ожидаемое значение (expectation) вектора скрытых переменных G по текущему приближению вектора параметров Θ. На M-шаге решается задача максимизации правдоподобия (maximization) и находится следующее приближение вектора Θ по текущим значениям векторов G и Θ. Критерием останова служит стабилизация значения правдоподобия $L(X^m;\Theta)$.

При использовании смесей нормальных распределений возможно выписать аналитические выражения для параметров θ_j. Окончательные формулы итераций записываются так:

$$w_j^{new}=\frac{1}{m}\sum_{i=1}^{m} g_{ij}$$

$$\mu_j^{new} = \frac{\sum_{i=1}^{m} x_i g_{ij}}{\sum_{i=1}^{m} g_{ij}} \tag{8.4}$$

$$R_j^{new} = \frac{\sum_{i=1}^{m} g_{ij} \left(x_j - \mu_j^{new}\right)\left(x_j - \mu_j^{new}\right)^T}{\sum_{i=1}^{m} g_{ij}}$$

Здесь $g_{ij} \equiv \mathrm{P}\left(\theta_j \mid x_i\right)$ – апостериорная вероятность того, что обучающий объект x_i был сгенерирован j-й компонентой смеси. Заметим, что эти выражения производят оба шага EM-алгоритма одновременно, как «expectation», так и «maximization». Алгоритм использует новое приближение параметров при переходе к следующей итерации.

Классический EM-алгоритм имеет два основных недостатка. Во-первых, результат и скорость сходимости может существенно зависеть от удачности начального приближения. Во-вторых, алгоритм работает с заранее заданным числом компонент, которое может быть не оптимально для данной выборки. Рис. 11 иллюстрирует эти недостатки. Для визуализации был использован программный пакет [14]. Как видно из рисунка, упомянутые проблемы могут оказывать критическое влияние на качество моделирования плотности выборки.

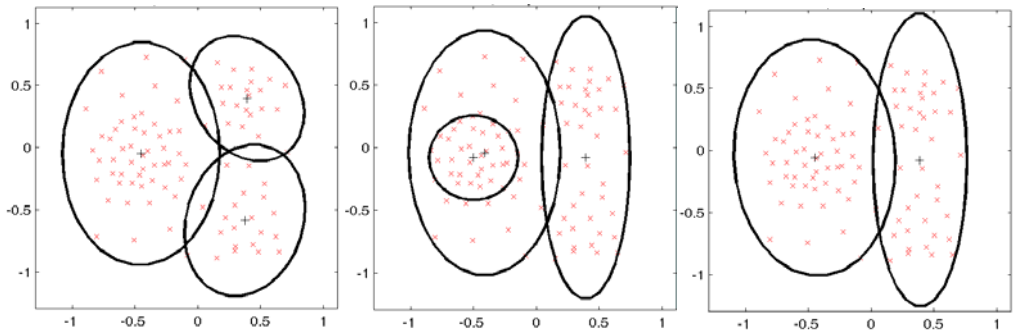

Рис. 11. (а) Правильная аппроксимация; (б) неверный выбор начального приближения; (в) неверное число компонент k.

В следующем разделе описан алгоритм последовательного разделения и добавления компонент, лишенный обоих недостатков.

9. АЛГОРИТМ ПОСЛЕДОВАТЕЛЬНОГО РАЗДЕЛЕНИЯ И ДОБАВЛЕНИЯ КОМПОНЕНТ

Идея алгоритма последовательного разделения и добавления компонент заключается в следующем. Имея некоторый набор компонент, найти компоненты, плохо описывающие принадлежащие им объекты и попытаться по очереди описать эти объекты ещё одной дополнительной компонентой. На начальном шаге алгоритма выборка аппроксимируется однокомпонентной смесью, чьи параметры находятся однозначно, что снимает проблему начального приближения.

Для реализации алгоритма требуется ввести критерий $C\left(\Theta_k\right)$ эффективности описания выборки смесью из k компонент, включающий в себя штраф на число компонент. В литературе предложено несколько таких критериев, их описание приводится, например, в [15]. В условиях рассматриваемой задачи наибольшую эффективность показал критерий ICL-BIC [16]:

$$ICL\text{-}BIC\left(\Theta\right) = -2L\left(\Theta\right) + 2E\left(\Theta\right) + \nu\left(\Theta\right)\log m\,. \tag{9.1}$$

Здесь $L(\Theta)$ – логарифм функции правдоподобия (8.3), $E(\Theta)$ – энтропия, $\nu(\Theta)$ – число свободных параметров в смеси Θ, m – число элементов в выборке. Энтропия $E(\Theta)$ определяется на основе апостериорных вероятностей g_{ij} как

$$E(\Theta) = -\sum_{i=1}^{m}\sum_{j=1}^{k} g_{ij}\log(g_{ij}) \geq 0. \tag{9.2}$$

В алгоритме также используется так называемый критерий разделимости $S(j;\Theta)$, характеризующий качество описания j-й компонентой смеси принадлежащих ей объектов. Идея разделения компонент смеси была описана в [17]. Критерий разделимости был определен там же как локальная дивергенция Кульбака:

$$S(j;\Theta) = \int f_j(x;\Theta)\ln\frac{f_j(x;\Theta)}{\varphi(x;\theta_j)}dx, \tag{9.3}$$

являющаяся расстоянием между двумя распределениями: локальной плотностью выборки $f_j(x;\Theta)$ для j-го распределения и плотностью j-го распределения, заданного текущим набором параметров Θ. Локальная плотность выборки определяется как

$$f_j(x;\Theta) = \frac{\sum_{i=1}^{m}\delta(x-x_i)g_{ij}}{\sum_{i=1}^{m} g_{ij}}, \tag{9.4}$$

где δ – дельта-функция: $\delta(x=0)=1, \delta(x\neq 0)=0$. Это модифицированное эмпирическое распределение, взвешенное апостериорными вероятностями. При этом близкие к j-му распределению объекты имеют больший вес. В [17] показано, что распределение с наибольшим $S(j;\Theta)$ имеет наихудшую оценку локальной плотности. Действительно, объекты выборки, лежащие далеко от j-го распределения, но тем не менее принадлежащие ему, будут иметь малое значение $\varphi(x;\theta_j)$ и большое значение $f_j(x;\Theta)$. Следовательно, распределение с наибольшим $S(j;\Theta)$ является первым кандидатом на разделение.

Для многомерного нормального распределения легко указать способ разделения на два распределения. Для этого можно рассмотреть геометрическую интерпретацию нормальной плотности. Линии уровня имеют форму эллипсоидов, оси которых направлены вдоль собственных векторов матрицы R. В общем случае главные оси эллипсоида R имеют разные длины. Для разбиения одного распределения смеси на два достаточно выбрать самую длинную из осей и поместить на неё центры двух новых распределений симметрично относительно центра старого распределения на расстоянии четверти длины этой оси. Вес старого распределения распределится поровну между двумя новыми распределениями. В качестве начального приближения дисперсий новых распределений берется дисперсия старого распределения. Если оси эллипсоида R имеют одинаковые длины, можно проводить разделение по любой из осей. Для работы этого способа разделения необходимо перед запуском алгоритма провести нормирование выборки для того, чтобы дисперсии по каждому признаку были равны. Без такой нормировки поиск самой длинной оси эллипсоида будет выполняться неверно – он будет зависеть от общей дисперсии выборки.

Алгоритм последовательного разделения и добавления компонент

1: **начало** – $k := 1$, параметры Θ_k найти с помощью EM-алгоритма

2: сохранить текущие параметры смеси $\Theta^* := \Theta_k$

3: для каждой компоненты смеси Θ_k вычислить $S(j;\Theta_k)$, $j=1,\ldots,k$

4: **в цикле** для всех компонентов j в порядке убывания $S(j;\Theta_k)$

5: разделить компоненту j на две, добавив $(k+1)$-ю компоненту

6: для имеющейся смеси выполнить шаги EM-алгоритма до останова и получить Θ_{k+1}

7: **если** $C(\Theta_{k+1}) > C(\Theta^*)$,

8: сохранить новые параметры смеси $\Theta^* := \Theta_{k+1}$ и выйти из цикла

9: **конец цикла** по j

10: **если** были найдены более хорошие параметры Θ^*,

11: перейти к шагу 3 с новыми параметрами смеси Θ^*, $k := k+1$

12: **иначе**

13: закончить алгоритм с параметрами Θ^*.

На шаге 1 алгоритма строится смесь, состоящая из одной компоненты $k=1$, и с помощью EM-алгоритма находятся её параметры Θ_k. На шаге 2 текущие параметры сохраняются как лучшие из найденных Θ^*. На шаге 3 начинается основной цикл. Для всех компонент смеси Θ_k вычисляется критерий разделимости $S(j;\Theta_k)$. На шагах 4-9 происходит перебор для всех компонентов j в порядке убывания $S(j;\Theta_k)$. Производится попытка разделения каждой компоненты на две, добавив в смесь $(k+1)$-ю компоненту. На шаге 6 находим параметры Θ_{k+1}, выполняя EM-алгоритм до стабилизации значения правдоподобия. Если критерий эффективности новой смеси $C(\Theta_{k+1})$ больше, чем критерий $C(\Theta^*)$, полученный для смеси до разделения, то сохраняем новые параметры смеси Θ_{k+1} в переменной Θ^* и завершаем цикл. Если по окончании цикла лучшее описание не было найдено, то алгоритм завершается с параметрами Θ^*. В ином случае, происходит возврат к шагу 3 и шаги 3-13 повторяются для смеси Θ^*.

Таким образом, рассматриваемый алгоритм позволяет, во-первых, выбрать оптимальное число компонент смеси, во-вторых, достичь глобального максимума правдоподобия, в-третьих, сократить число итераций за счет использования критерия разделимости $S(j;\Theta)$.

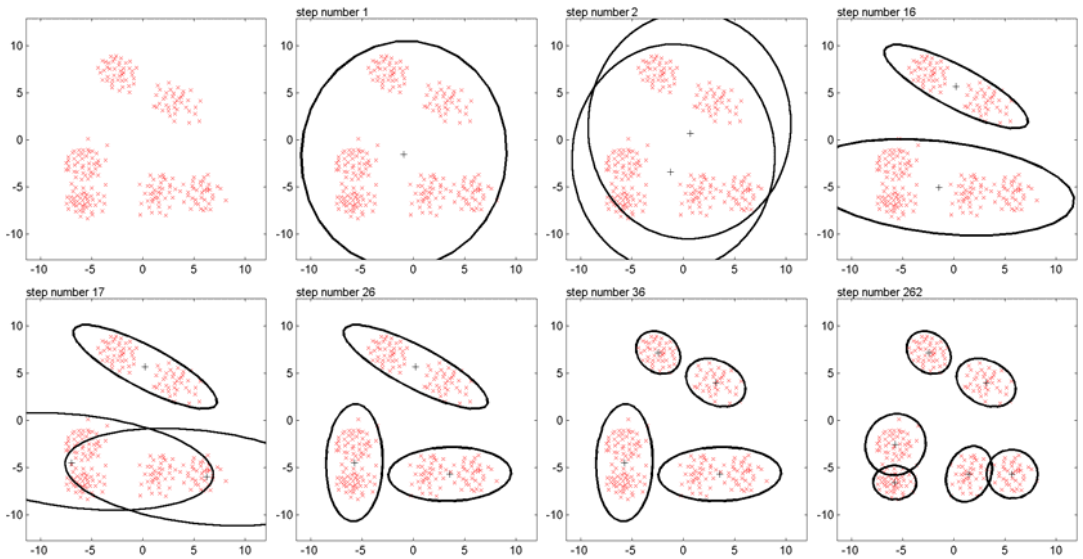

Рис. 12. Пример работы алгоритма последовательного разделения и добавления компонент на синтетической двумерной выборке.

На рис. 12 показана работа алгоритма последовательного разделения и добавления компонент на примере синтетической двумерной выборки, сгенерированной шестью нормальными распределениями. Шаг 1 – начальная аппроксимация однокомпонентной смесью. На шаге 2 происходит разделение начальной компоненты на две. Шаг 16 – окончательный вид двухкомпонентной смеси. Шаг 17 – разделение нижней компоненты на две. Шаг 26 – окончательный вид трехкомпонентной смеси, шаг 26 - четырехкомпонентной. Работа алгоритма завершается на 262 шаге. Найдено оптимальное число компонент в соответствии с критерием ICL-BIC, попытки аппроксимации смесью из семи компонент дали худшее значение критерия эффективности.

На рис. 13 показан пример моделирования плотности сложного множества – трехмерной спирали (изображены длинные оси эллипсоидов). Окончательное число компонент смеси было фиксировано и равно десяти. Видно, что спираль равномерно покрыта нормальными распределениями, длинные оси эллипсоидов параллельны центральной линии спирали. Заметим, что классический EM-алгоритм не в состоянии справиться с этой задачей.

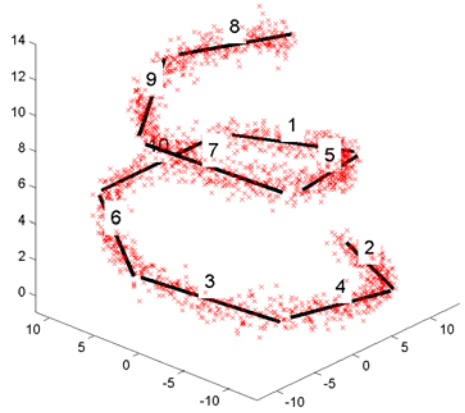

Рис. 13. Пример работы алгоритма: моделирование множества в виде трехмерной спирали.

Итак, создан алгоритм, позволяющий моделировать распределения плотности вероятности произвольных многомерных выборок. Плотность моделируется смесью нормальных распределений, чьи параметры находятся с помощью метода максимального правдоподобия. Алгоритм позволяет избежать локальных максимумов правдоподобия и выбрать оптимальное число компонент в смеси. Это делает возможным применение данного алгоритма для описания плотности вероятности многомерного распределения акустических параметров в задаче распознавания типов сегментов в речи.

10. ОПРЕДЕЛЕНИЕ РАЗДЕЛЯЮЩИХ НАБОРОВ АКУСТИЧЕСКИХ ПАРАМЕТРОВ

Выше было показано, что для разделения каждой пары кардинальных типов каждый параметр имеет свою эффективность. Однако по отдельности параметры обеспечивают слишком малую разделимость, что обуславливает необходимость выделения целого набора параметров для распознавания каждой пары типов.

Задача выбора наилучшего разделяющего набора параметров осложнена тем, что многие параметры в значительной степени коррелированны. Этот факт приводит к необходимости разработки специальной процедуры, которая выделяла бы из всего списка параметров ограниченный набор, дающий наибольшую эффективность разделения пары типов.

Процедура выделения оптимального списка параметров основывается на формуле апостериорной вероятности. За критерий оптимальности принимается площадь пересечения гистограмм апостериорных вероятностей, вычисленных для пары противопоставляемых типов в рассматриваемом подпространстве параметров. Такой выбор критерия обуславливается тем, что площадь пересечения гистограмм равна минимальной суммарной ошибке распознавания при решении задачи дихотомии для данной пары. Таким образом, процедура выбирает подпространство параметров, приводящее к наименьшей ошибке.

Процедура выделения оптимального списка параметров

1: Выбор первого параметра в наборе по минимуму пересечения гистограмм.
2: **начать**
3: **в цикле** для всех не входящих в набор параметров
4: временно включить новый параметр в набор
5: выполнить аппроксимацию выборки для первого типа из пары
6: выполнить аппроксимацию выборки для второго типа из пары
7: по всем точкам первой выборки вычислить апостериорную вероятность принадлежности этих точек к первому типу, построить первую гистограмму
8: по всем точкам второй выборки вычислить апостериорную вероятность принадлежности этих точек к первому типу, построить вторую гистограмму
9: найти значение критерия оптимальности как площадь пересечения двух гистограмм
10: **конец цикла**
11: включить в набор параметр, дающий наименьшее значение критерия по результатам цикла
12: **повторять** пока количество параметров в наборе меньше требуемого

На первом шаге происходит выбор первого параметра в наборе по минимуму пересечения одномерных гистограмм. Далее для добавления в набор еще одного параметра предпринимается следующее. Параметры, до сих пор не вошедшие в набор, по очереди временно включаются в него. В полученном подпространстве выполняется аппроксимация плотностей противопоставляемых типов. Критерий оптимальности вычисляется как пересечение гистограмм апостериорных вероятностей, вычисленных по точкам выборок. Из всех рассматриваемых параметров в набор добавляется тот, у которого значение критерия было наименьшим. Эти шаги повторяются, пока в наборе не окажется заданное количество параметров.

Эта итерационная процедура по очереди добавляет к набору параметр, наименее коррелированный с уже присутствующими в наборе, и привносящий, таким образом, наибольшее количество информации о различии противопоставляемых типов.

В ходе исследований было установлено, что для повышения эффективности распознавания целесообразно разделить сегменты на большее число типов, и искать наилучшие разделяющие наборы параметров для каждого из них отдельно. Это связано с тем, что каждый кардинальный тип сегментов объективно содержит несколько различных классов звуков, отличающихся по своим характеристикам. Наибольшие различия наблюдались для высоких и низких фрикативных, а также для высоких и низких гласноподобных звуков. В связи с этим было принято решение о поиске подпространств параметров по отдельности для «гласноподобных высоких», «гласноподобных низких», «фрикативных высоких», «фрикативных низких». Таким образом, для распознавания шести кардинальных типов сегментов следует найти разделяющие параметры для восьми вспомогательных классов, перечисленных в таблице 4.

Таблица 4. Список вспомогательных классов для поиска подпространств параметров.

название класса	краткое обозначение	некоторые члены класса
«гласноподобный низкий»	«ГН»	«У», «Ы», «А», «О», «Э»
«гласноподобный высокий»	«ГВ»	«Е», «Я», «И», «Ю»
«фрикативный высокий глухий»	«ФВг»	«С», «Ш»
«фрикативный высокий звонкий»	«ФВз»	«З», «Ж»
«фрикативный низкий»	«ФН»	«Ф», «Х»
«смычной глухой»	«Сг»	«П», «Т», «К»
«смычной звонкий»	«Сз»	«Б», «Д», «Г»
«назальный»	«Н»	«Н», «М»

В качестве итогового решающего правила использовалась формула полной вероятности

$$\mathrm{P}\left(A\,|\,x\right)=\frac{1}{N}\sum_{a}\sum_{b}\mathrm{P}_{b}\left(a\,|\,x\right). \qquad (10.1)$$

Здесь A – тип сегмента в итоговой классификации на шесть кардинальных элементов, $\mathrm{P}\left(A\,|\,x\right)$ – вероятность того, что данный сегмент является сегментом типа A при измеренном векторе параметров x, a – «свои» типы сегмента, b – все типы, противопоставляемые типам a, N – число членов в двойной сумме, $\mathrm{P}_{b}\left(a\,|\,x\right)$ – апостериорная вероятность, определенная формулой (7.3). Например, если вычисляется вероятность того, что данный сегмент является гласным, в этой формуле A = «гласноподобный», a = {«гласноподобный высокий», «гласноподобный низкий»}, b = {«назальный», «фрикативный высокий глухой», «фрикативный высокий звонкий», «фрикативный низкий», «смычной глухой», «смычной звонкий»}, $N=2\cdot 6=12$.

Оптимальные подпространства были вычислены по речевым сигналам из речевой базы данных ИППИ. Их размерность устанавливалась равной семи. Параметры, образующие подпространства, перечислены в таблице 5 в порядке добавления их в оптимальный набор.

Таблица 5. Оптимальные подпространства для распознавания вспомогательных классов.

пара	1	2	3	4	5	6	7
<ГН>-<ФВг>	S3	Pfull	ceps1	Ba3	E21	Wfull	C13
<ГН>-<ФВз>	ceps1	E22	M21	Plow	E23	C13	Bb4
<ГН>-<ФН>	E22	Plow	ceps2	Wfull	S2	E23	E21
<ГН>-<Сг>	E23	Plow	Bb4	Ba3	M11	S3	C12
<ГН>-<Сз>	E23	Bb1	Ba1	C13	M12	ceps4	Wfull
<ГН>-<Н>	E23	Ba3	Bb1	C11	Pfull	C13	M12
<ГВ>-<ФВг>	E22	Pfull	ceps1	E23	C13	Plow	Ba1
<ГВ>-<ФВз>	E22	ceps1	M12	E23	Bb4	M21	Wfull
<ГВ>-<ФН>	E21	Plow	Ba3	E22	Bb1	M11	C13
<ГВ>-<Сг>	E23	Plow	Ba3	M12	S3	C13	M21
<ГВ>-<Сз>	E23	S3	Plow	Ba3	E22	M12	C13
<ГВ>-<Н>	E23	Ba3	C11	Bb1	E22	E21	C13
<ФВг>-<ФВз>	Plow	ceps1	E23	E21	C13	ceps2	Bb1
<ФВг>-<Сг>	E23	S1	Pfull	Wfull	ceps1	C13	Bb4
<ФВг>-<Сз>	S1	Wfull	Pfull	E23	ceps1	S2	M21
<ФВг>-<Н>	S3	Wfull	Plow	ceps1	Bb1	C11	E23
<ФВз>-<ФН>	Plow	S3	E21	ceps1	Bb4	Ba1	S2
<ФВз>-<Сг>	Wfull	Plow	S3	Ba3	M12	Pfull	E23
<ФВз>-<Сз>	Wfull	S3	Plow	E22	ceps1	E23	ceps2
<ФВз>-<Н>	S3	ceps1	E23	Ba1	E22	Wfull	Bb1
<ФН>-<Сг>	E23	E22	ceps1	ceps2	S3	ceps3	E21
<ФН>-<Сз>	Bc1	Ba1	S3	Wfull	E21	E23	Plow
<ФН>-<Н>	Plow	S1	Ba1	E21	Wfull	E23	Pfull
<Сг>-<Сз>	E21	E22	Plow	M21	Wfull	Ba1	Pfull
<Сг>-<Н>	Plow	Wfull	S3	Pfull	Ba1	M12	E21
<Сз>-<Н>	E23	Plow	S2	C12	S1	ceps1	M21

11. РЕЗУЛЬТАТЫ

Тестирование точности поиска границ сегментов и качества распознавания кардинальных типов производилось на материале размеченной речевой базы данных ИППИ. База

соответствовала реальным условиям применения, содержала голоса нескольких десятков дикторов, записанные на несколько типов микрофонов и телефонных трубок. В тестировании участвовали сигналы с соотношением сигнал-шум от 12 дБ. Тестирование выполнялось без использования дополнительной информации о содержании сигнала.

При тестировании алгоритма поиска границ в качестве эталонов использовались границы артикуляторно-акустических сегментов, установленные при ручной разметке базы. Средние по всей базе характеристики для разных порогов приведены в таблице 6. Лучшее соотношение характеристик достигается при значении порога $m^* = 1.0 \cdot 10^{-2}$. Средняя погрешность определения положений границ равно 4.52 мс, среднее число пропусков границ равно 0.95%, среднее число вставок на один сегмент ручной разметки составило 1.26.

Таблица 6. Характеристики алгоритма поиска границ в зависимости от величины порога.

показатель	$m^* = 0.7 \cdot 10^{-2}$	$m^* = \mathbf{1.0 \cdot 10^{-2}}$	$m^* = 1.5 \cdot 10^{-2}$
погрешность положения границ	4.28 мс	**4.52 мс**	4.85 мс
процент пропусков границ	0.76%	**0.95%**	1.41%
число вставок на сегмент разметки	1.53	**1.26**	0.96

Точность, с которой опытный лингвист размечает сигнал, сложно измерить экспериментально. Для этого пришлось бы предлагать разным лингвистам выполнять разметку набора одинаковых сигналов, что на практике трудноосуществимо. Однако среднюю погрешность, заложенную в эталонах, можно оценить. При ручной разметке используемой базы данных использовались спектрограммы с шагом по времени 5 мс. Запись базы осуществлялась параллельно с двух микрофонов, один из которых был расположен дальше другого, и между двумя сигналами возникала задержка примерно 3 мс. Следовательно, погрешности автоматической расстановки границ, лежащие в диапазоне до 5 мс, приемлемы. Стремиться к дальнейшему уточнению положений границ не следует, поскольку это невозможно в силу содержащихся в эталонах погрешностей.

Распределение ошибок алгоритма по типам сегментов разметки показано в таблице 7 и на рис. 14 в виде гистограмм. Используются следующие обозначения: # – пауза, V – гласноподобный, N – назальные, F – фрикативный, C – смычка, B – взрыв.

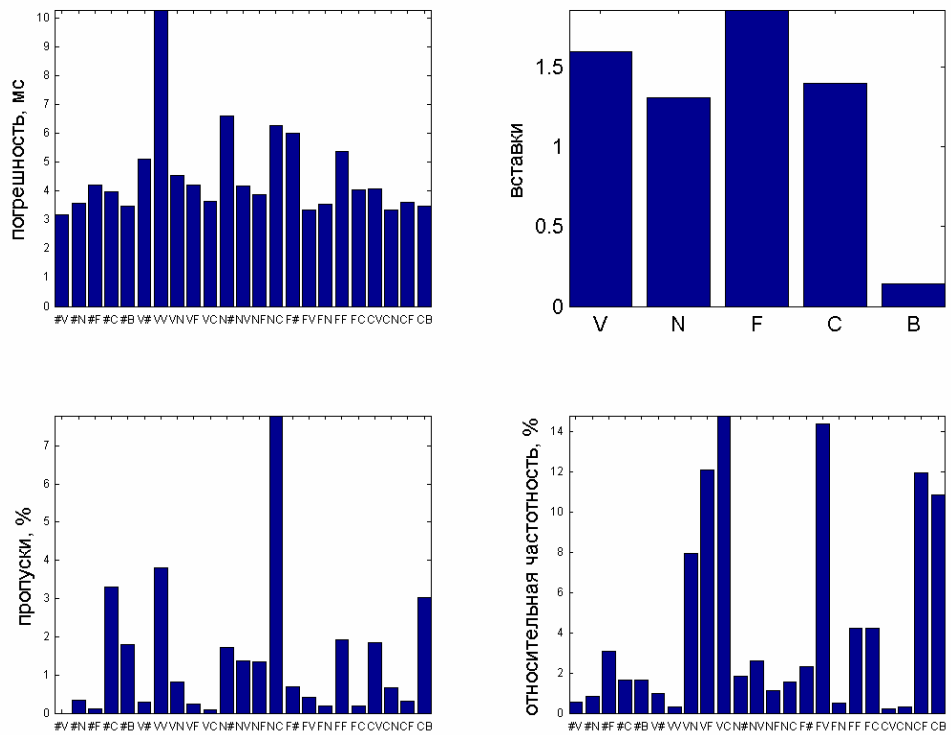

Рис. 14. Распределение ошибок алгоритма поиска границ по типам сегментов ручной разметки.

Таблица 7. Распределение ошибок алгоритма поиска границ по типам сегментов ручной разметки.

переход	погрешность, мс	пропуски, %
#V	3.2	0.00
#N	3.6	0.35
#F	4.2	0.11
#C	4.0	3.31
#B	3.5	1.81
V#	5.1	0.30
VV	10.3	3.80
VN	4.6	0.83
VF	4.2	0.25
VC	3.6	0.10
N#	6.6	1.72
NV	4.2	1.36
NF	3.9	1.35
NC	6.3	7.79
F#	6.0	0.68
FV	3.3	0.41
FN	3.6	0.18
FF	5.4	1.93
FC	4.1	0.18
CV	4.1	1.84
CN	3.4	0.66
CF	3.6	0.32
CB	3.5	3.02

класс	вставки
V	1.6
N	1.3
F	1.9
C	1.4
B	0.1

Наибольшие погрешности положений границ приходятся на переходы типа «VV». Чтобы объяснить это, следует принять во внимание, что при выполнении ручной разметки часто сложно с уверенностью указать положение границы между гласными, так как переходный процесс непрерывен. Поэтому расположение эталонных границ «VV» несет значительный отпечаток субъективного мнения разметчика, что приводит к сильному разбросу и большим погрешностям алгоритма. Кроме того, погрешности на переходах типа «VV» являются следствием объективно малого изменения характеристик сигнала.

Наибольшее число пропусков приходится на переходы от назальных к смычкам. Детальный анализ таких ошибок показывает, что в основном пропуски возникают на переходах от назальных к звонким смычкам. В этом случае справедливы все замечания, касающиеся переходов «VV»: объективная сложность создания адекватных эталонов, малое изменение спектральных характеристик сигнала.

Анализ вставок лишних сегментов показывает, что число вставок на взрывах ничтожно мало, в то время как число вставок на гласноподобных, назальных, фрикативных и смычных примерно одинаково и равно в среднем 1,5. Иными словами, каждый сегмент эталонной разметки разбит на два или три сегмента. Это легко объяснить для каждого типа сегментов. Сегменты, отмеченные разметчиком как гласные, обычно содержат три участка: переходный участок, соответствующий подходу к нужной конфигурации артикуляторов; квазистационарный участок, соответствующий целевой конфигурации; переходный участок, соответствующий движению артикуляторов к следующему состоянию. На рис. 15 приведена иллюстрация этого явления для слова «ОДИН». Видно, что первый звук, безударное «а», фактически состоит из двух участков: квазистационарного участка «а» и переходного процесса, который приводит форманты к состоянию, по звучанию близкому к звуку «и». Движение формант обусловлено наличием последующей мягкой смычки «Д», и в теории речи известно как явление коартикуляции. Это же явление объясняет появление первой границы внутри ударного «И». Вторая граница внутри «И» соответствует опусканию нёбной занавески и началу назализации.

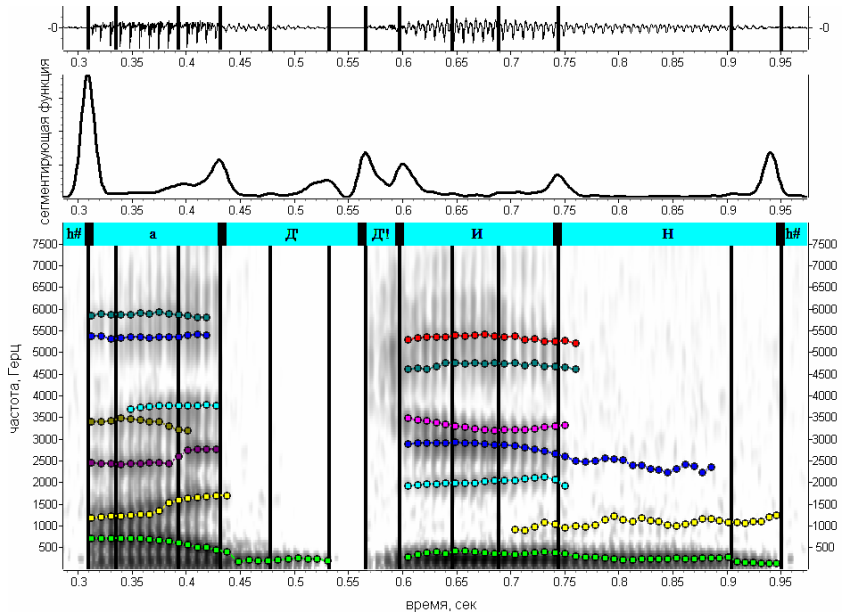

Рис. 15. Иллюстрация переходных процессов и явления коартикуляции на примере слова «ОДИН».

В случае со смычными появление лишних границ вызвано, как правило, затуханием голосовых колебаний к концу звонкой смычки вследствие выравнивания давления в подсвязочной области и речевом тракте (рис. 15, звонкая мягкая смычка «Д»). Кроме того, появление лишних границ на смычных и на назальных может быть вызвано различными нестационарными шумами, как внешними, так и связанными с речевым аппаратом диктора, поскольку эти типы сегментов имеют низкое соотношение сигнал-шум по сравнению с гласноподобными и фрикативными, что повышает влияние шумов на анализ сигнала.

На фрикативных лишние границы возникают вследствие наличия турбулентного источника возбуждения акустических колебаний, имеющих изначально нестационарный характер. Кроме того, на фрикативных часто возникает ситуация, когда часть звука огласована, а другая часть не содержит голосовых импульсов. На рис. 16 изображен пример такой ситуации для звукосочетания «АЖА». Средняя часть «Ж» является глухим фрикативным участком.

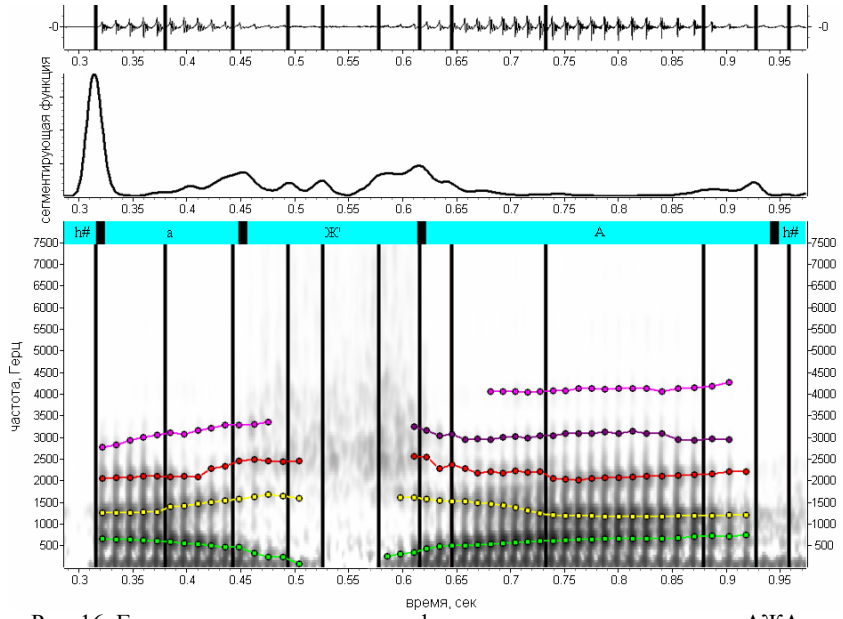

Рис. 16. Глухие и звонкие участки на фрикативных, звукосочетание «АЖА».

Сравнение точности нахождения границ с описанными в литературе результатами позволяет говорить о превосходстве предложенного в текущей работе алгоритма. Например, в [6] погрешность положения границ составила примерно 5.79 мс, однако тестирование производилось на базе данных для одного диктора, во фразах преобладали слоги «согласный – гласный», что формирует значительно более благоприятные условия по сравнению с условиями

базы данных ИППИ. В статье [18] производится тестирование трех алгоритмов сегментации: MFC-SVF, RASTA-SVF и FBDYN-SVF на материале англоязычной телефонной базы данных OGI_TS при соотношении сигнал-шум 20 дБ. В таблице 8 приведены результаты тестирования этих алгоритмов в сравнении с предложенным алгоритмом. Видно, что предложенный в текущей работе алгоритм значительно превосходит конкурирующие.

Таблица 8. Сравнение характеристик предложенного алгоритма поиска границ с алгоритмами MFC-SVF, RASTA-SVF и FBDYN-SVF при соотношении сигнал-шум 20 дБ.

показатель	MFC-SVF	RASTA-SVF	FBDYN-SVF	предложенный алгоритм
погрешность положения границ	8.45	15.6	7.5	**4.52 мс**
процент пропусков границ	6.06%	12.41%	3.08%	**0.95%**
число вставок на сегмент разметки	1.80	1.03	2.41	**1.26**

Тестирование качества распознавания кардинальных типов на материале речевой базы данных ИППИ показало, что в 85% случаев правильный тип имел наибольшее значение апостериорной вероятности $\mathrm{P}(A \mid x)$, а в 96,3% случаев правильный тип находился в первой двойке. Это позволяет сделать вывод о том, что выбранные акустические параметры и способ их анализа позволяют адекватно описывать и распознавать кардинальные элементы речевого сигнала.

Как было сказано выше, для решения практических задач не требуется обязательного принятия окончательного решения о типе сегмента. Однако представляет интерес мера близости кардинальных типов в пространствах исследованных акустических параметров. Анализ ошибок распознавания показал, что чаще всего вероятность правильного типа оказывается не на первом месте вследствие незначительного превосходства значения вероятности парного типа, объективно близкого по акустическим свойствам. Была обнаружена близость следующих пар: «смычной глухой»–«смычной звонкий» и «смычной звонкий»–«назальный». В меньшей степени близки следующий пары: «гласноподобный»–«назальный» и «фрикативный глухой»–«фрикативный звонкий». В таблице 9 приведена матрица перепутывания кардинальных типов, построенная по принципу максимального правдоподобия. Эта матрица подтверждает сделанные наблюдения о близости указанных пар типов. Видно также, что лучше всего отделены от прочих типов «гласноподобный» и «фрикативный глухой».

Таблица 9. Матрица перепутывания кардинальных типов в пространствах акустических параметров.

		распознано как					
		гласно-подобный	назальный	фрик. гл.	фрик. зв.	смыч. гл.	смыч. зв.
предъявлено	гласно-подобный	88	7	0	4	0	0
	назальный	8	71	0	2	3	16
	фрик. гл.	0	0	87	11	1	0
	фрик. зв.	6	4	5	82	0	2
	смыч. гл.	0	2	1	1	80	16
	смыч. зв.	0	15	0	1	13	71

По матрице перепутывания также видно, что точность распознавания по принципу максимального правдоподобия в случае «назальных» и «смычных звонких» звуков далека от точности, с которой эти типы распознает человек [19]. Это, по-видимому, является следствием того, что человек использует более тонкие динамические признаки по сравнению с описанными в данной работе параметрами группы «Отношение энергии». Выявление и исследование этих признаков представляет интерес и является одним из дальнейших направлений работы.

В целом результаты тестирования позволяют заключить, что предложенные в данной работе методы сегментации речевого сигнала на квазистационарные и переходные участки, а также распознавания кардинальных типов сегментов обладают высокой точностью и устойчивостью, что делает возможным их использование в реальных задачах речевых технологий.

12. СЕГМЕНТАЦИЯ В ЗАДАЧАХ РЕЧЕВЫХ ТЕХНОЛОГИЙ

В обратной задаче происходит расчет формы речевого тракта по акустическим параметрам речевого сигнала. Из работы [20], посвященной обратной задаче, известно, что для каждого типа сегментов должны быть использованы разные параметры. Например, для гласных и назальных это формантные частоты, для фрикативных – характерные частоты спектра, для звонких смычных – частота радиального резонанса. Таким образом, для решения обратной задачи необходимо разделить речевой сигнал на однотипные сегменты, определить тип каждого сегмента и соответствующие этому типу акустические параметры. Представленные в данной работе алгоритмы позволяют полностью обеспечить информационную поддержку решения обратной задачи. На рис. 17 показан обработанный сигнал, поступающий на вход обратной задачи. Сверху вниз: осциллограмма, основной тон, сегменты и вероятности типов.

Тестирование выполнялось на материале речевой базы данных Westbury [21], содержащей измерения на микролучевой рентгеноскопической установке и включающей в себя образцы речи и артикуляции примерно полусотни дикторов – носителей американского английского языка. При использовании предложенных алгоритмов сегментации и вычисления акустических параметров для решения обратной задачи показано, что погрешность определения формы речевого тракта для гласных составила 6%, для фрикативных – 3%, что сопоставимо с погрешностью измерения. Субъективное тестирование артикуляторного ресинтеза показало, что на слух различие между оригинальным и ресинтезированным сигналами ничтожно мало. Это свидетельствует о том, что описанные в работе алгоритмы являются достаточно надежными и точными для использования в решении обратной задачи.

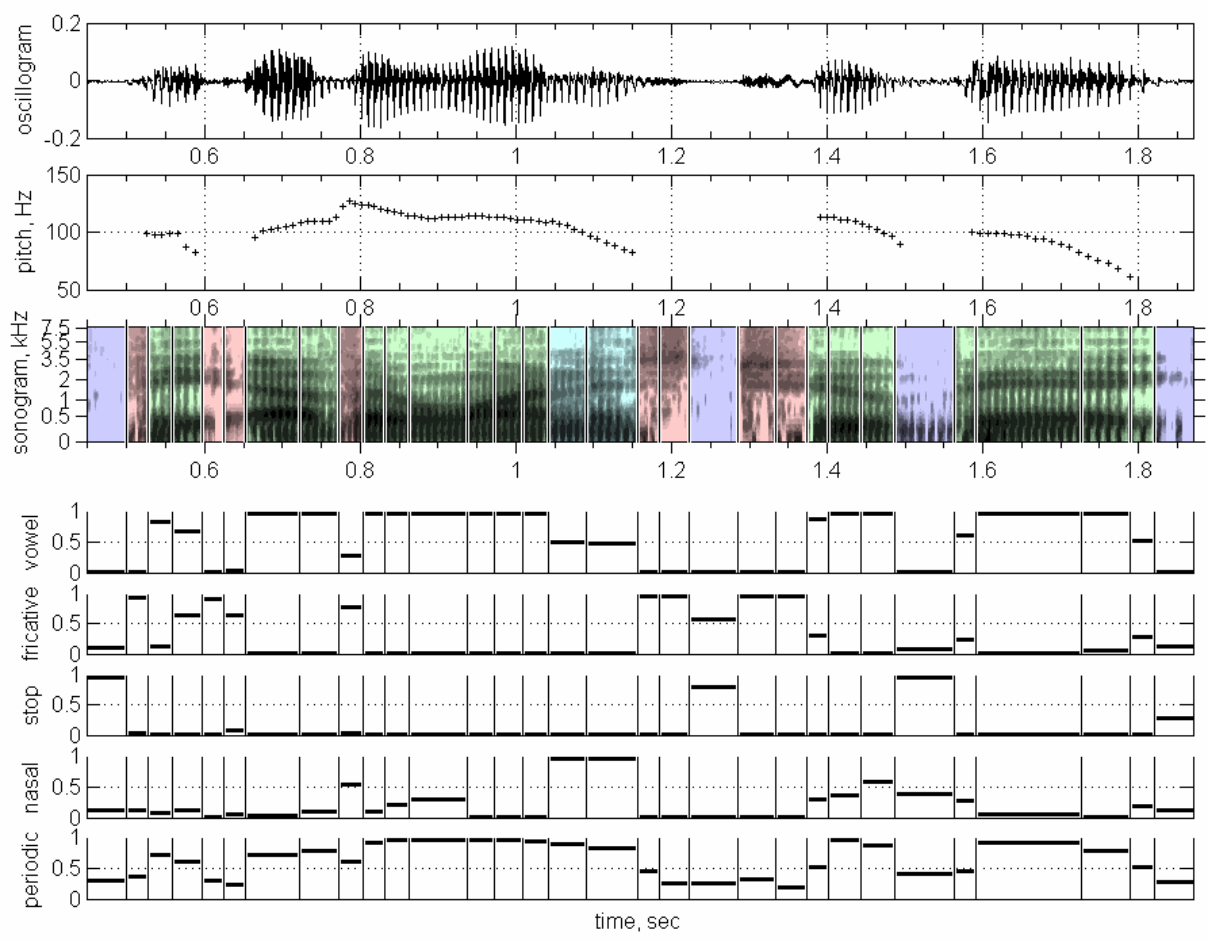

Рис. 17. Обработанный сигнал, поступающий на вход обратной задачи

В некоторых речевых задачах доступна дополнительная информации о речевом сигнале. Например, в задаче голосового набора номера ограничен словарь распознаваемых слов. В таких задачах возможно использование метода динамической трансформации шкалы времени (Dynamic Time-Warping, DTW) для распознавания фраз и поиска фонетических сегментов на сигнале [22-24]. По размеченной базе речевых данных производится сбор эталонов, включающих в себя различные акустические параметры сигналов и положения установленных вручную границ сегментов. С помощью метода DTW выполняется выравнивание границ и строится мера сходства между эталонами и пришедшим сигналом. Выбирается один наиболее сходный эталон и соответствующее ему положение границ из ручной разметки.

Эффективность работы такого алгоритма зависит от способа описания эталонов. Как правило, эталоны содержат данные о сигнале, полученные из сонограммы, подвергнутой примитивной обработке. Распространенным вариантом является использование младших коэффициентов кепстра.

В литературе описаны попытки использования DTW для распознавания речи [25-27], однако точность при этом не поднимается выше 95% в хороших условиях (высокое соотношение сигнал-шум, один микрофон), а в реальных условиях опускается ниже 80%.

Было проведено тестирование влияния информации о принадлежности сегментов к кардинальным типам на точность распознавания изолированных цифр русского языка. При сборе эталонов по базе данных выполнялась сегментация каждого сигнала, и вычислялись апостериорные вероятности принадлежности каждого сегмента к одному из шести кардинальных типов. Эти вероятности сохранялись в эталонах наряду с данными, полученными по сонограмме. В качестве меры сходства отсчетов сигнала и эталона, используемой в DTW применительно к вероятностной части эталонов, использовалась метрика L_2.

Распознавание проводилось независимо от диктора и микрофона по речевой базе ИППИ с минимальным соотношение сигнал-шум 12 дБ (реальные условия). По результатам тестирования количество ошибок распознавания составило 2%, что значительно превышает результаты обычных систем, использующих DTW.

Для сравнения была рассмотрена система распознавания, использующая вариант дискретного динамического программирования, основанного на акустических детекторах артикуляторных событий [28]. В этой системе различалось 120 типов сегментов, на которые фактически происходила сегментация сигнала при распознавании. Тестирование этой системы на той же базе данных показала, что количество ошибок распознавания составляет 12%, что в 6 раз превышает ошибку, полученную на системе, использующей вероятности шести кардинальных типов.

Можно сделать вывод, что использование вероятностей кардинальных элементов, а также уменьшение количества типов сегментов, на которые происходит обучение, в условиях ограниченного словаря существенно повышает точность распознавания слов.

Метод DTW также может быть использован в задаче верификации диктора по голосу с контекстно-ограниченным паролем. В этом случае заранее известно содержание произнесения. При голосовой верификации голос диктора характеризуется как акустическими параметрами (формантные частоты гласных, характеристические частоты фрикативных), так и временными (длительность слова, длительность ударного гласного, длительности пар сегментов). Для извлечения этих параметров из сигнала необходимо сначала указать границы значимых сегментов. Например, для вычисления формантных частот на ядре ударного гласного требуется найти положение слова на сигнале и ударного гласного на слове.

Для этого в рассматриваемой системе верификации использовался метод DTW с эталонами, содержащими информацию о типе сегмента. Метод DTW выполняет поиск наилучшего согласующегося с сигналом эталона, и расставляет на сигнале границы в соответствии с содержащимися в эталоне границами из ручной разметки. Следовательно, при такой сегментации число потерь границ сегментов, число вставок и число неверно распознанных в терминах разметки сегментов практически равны нулю (они зависят от профессионализма разметчика базы данных). При этом происходит некоторое увеличение погрешности положений границ. В экспериментах эта погрешность составила в среднем 16 мс, что не оказало существенного влияния на качество верификации.

Тестирование системы верификации проводилось на материале специальной базы речевых данных, содержащей голоса 130 дикторов, произнесших каждое слово не менее 40 раз. Средняя ошибка верификации для паролей, состоящих из десяти слов, составила 0.04%. Это говорит о

том, что комбинированный алгоритм сегментации (DTW + вероятности типов) позволяет точно и устойчиво находить положения характерных фонетических сегментов.

13. ЗАКЛЮЧЕНИЕ

В работе показано, что использование динамических и статических свойств позволяет эффективно производить анализ речевого сигнала.

Предложены алгоритмы сегментации речевого сигнала на квазистационарные и переходные участки на основе корреляции спектров и распознавания кардинальных типов речевых сегментов. Их эффективность оценена на материале базы речевых данных русского языка для 47 человек и нескольких типов телефонных трубок и микрофонов с ручной разметкой на 127 типов артикуляторно-акустических сегментов по сигналам с соотношением сигнал-шум от 12 дБ.

При сегментации речевого сигнала выполнялся поиск границ квазистационарных и переходных участков. На имеющейся базе данных алгоритм сегментации определил положения границ со средней погрешностью 4,52 мс, было пропущено 0,95% границ, среднее число вставок было равно 1,26 на один сегмент ручной разметки. Показано, что основную часть пропущенных границ составляли слабовыраженные переходы, и что основные погрешности положений границ и вставки обусловлены объективными свойствами сигналов и субъективностью разметки эталонов.

Распознавание кардинальных типов речевых сегментов производилось в подпространствах акустических параметров, установленных специальной оптимизационной процедурой. При тестировании правильный тип в 85% случаев имел наибольшее значение апостериорной вероятности, в 96,3% входил в первую двойку.

СПИСОК ЛИТЕРАТУРЫ

1. *Сорокин В.Н., Цыплихин А.И.* Сегментация и распознавание гласных // Информационные процессы, 2004, Т. 4, № 2, С. 202-220.
2. *Леонов А.С., Макаров И.С., Сорокин В.Н., Цыплихин А.И.* Артикуляторный ресинтез фрикативных // Информационные процессы, 2004, Т. 4, № 2, С. 141-159.
3. *Mermelstein P.* Automatic segmentation of speech into syllable units // J. Acoust. Soc. Amer., 1975, V. 58, N. 4, P. 880-883.
4. *Ali A.M.A., Spiegel J.V.* Acoustic-phonetic features for the automatic classification of fricatives // J. Acoust. Soc. Am., 2001, V. 109, N. 5, Pt. 1, P. 2217-2235.
5. *van Hemert J. P.* Automatic segmentation of speech // IEEE Transactions on Signal Processing, 1991, V. 39, P. 1008-1012.
6. *Wu Y.J., Kawai H., Ni J., Wang R.H.* Discriminative training and explicit duration modeling for HMM-based automatic segmentation // Speech Communication, 2005, V. 47, N. 4, P. 397-410.
7. *Liu S.* Landmark detection for distinctive feature based speech recognition // J. Acoust. Soc. Amer., 1996, V. 100, P. 3417-3430.
8. *Niyogy P., Sondhi M.M.* Detecting stop consonants in continuous speech // J. Acoust. Soc. Amer., 2002, V. 111, P. 1063-1076.
9. *Pellom B.L., Hansen J.H.L.* Automatic segmentation and labeling of speech recorded in unknown noisy channel environments // Speech Communication, 1998, V. 25, P. 97-116.
10. *Amit Y., Koloydenko A., Niyogi P.* Robust acoustic object detection // J. Acoust. Soc. Amer., 2005, V. 118, N. 4, P. 2634-2648.
11. *Сорокин В.Н., Чепелев Д.Н.* Первичный анализ речевых сигналов // Акустический ж., 2005, Т. 51, № 4, С. 536-542.
12. *Toh A.M., Togneri R., Nordholm S.* Spectral Entropy as Speech Features for Speech Recognition // Proceedings of PEECS2005, Perth, 2005, P. 22-25
13. *Niyogi P., Ramesh P.* The Voicing Feature for Stop Consonants: Recognition Experiments with Continuously Spoken Alphabets // Speech Communication, 2003, V. 41, P. 349-367.
14. *Franc V., Hlavac V.* Statistical Pattern Recognition Toolbox // Czech Technical University Prague, 2005, http://cmp.felk.cvut.cz .

15. *Kuo B.C., Landgrebe D.* Improved Statistics Estimation And Feature Extraction For Hyperspectral Data Classification: PhD Thesis and School of Electrical & Computer Engineering Technical Report TR-ECE 01-6, December 2001.
16. *McLachlan G., Peel D.* // Finite Mixture Models, 2000, New York: John Wiley & Sons Inc.
17. *Naonori U., Ryohei N., Ghahramani Z., Hinton G.E.* SMEM Algorithm for Mixture Models // Neural Computation, 2000, V. 12, N. 9, P. 2109-2128.
18. *Bojan P., Anderson O., Dalsgaard P.* On the robust automatic segmentation of spontaneous speech // In Proceedings of the International Conference on Spoken Language Processing (ICSLP'96) , Philadelphia, 1996, P. 913-916.
19. *Сорокин В.Н.* Теория речеобразования. – М.: Радио и связь, 1985. – 313 с.
20. *Леонов А.С., Макаров И.С., Сорокин В.Н., Цыплихин А.И.* Кодовая книга для речевых обратных задач // Информационные процессы, 2005, Т. 5, № 2, С. 101-119.
21. *Westbury J.* X-ray Microbeam Speech Production Database User's Handbook, Version 1.0 (June 1994) // University of Wisconsin, 1994.
22. *Stöber K., Hess W.* Additional Use of Phoneme Duration Hypotheses in Automatic Speech Segmentation // Proceeding of the ICSLP´98, Sydney, 1998, Paper number 239.
23. *Kominek J., Bennett C., Black A.W.* Evaluating and Correcting Phoneme Segmentation for Unit Selection Synthesis // in Proceedings ESCA Eurospeech'03, 2003.
24. *Gomez J.A., Castro M.J.* Automatic Segmentation of Speech at the Phonetic Level // Proceedings in Springer-Verlag LNCS, 2002, V. 2396 (T. Caelli et al. eds.), P. 672-680.
25. *Itakura F.* Minimum prediction residual principle applied to speech recognition // IEEE Trans Acoustics Speech Signal Process, 1975, V. 23, P. 52–72.
26. *Myers C., Rabiner L., Roseneberg A.* Performance tradeoffs in dynamic time warping algorithms for isolated word recognition // IEEE Trans Acoustics Speech Signal Process, 1980, V. 28, P. 623–635.
27. *Rabiner L., Juang B.* Fundamentals of speech recognition // Prentice, Englewood Cliffs, NJ, 1993.
28. *Сорокин В.Н., Ижнин А.Н., Цыплихин А.И., Чепелев Д.Н.* Артикуляторно-ориентированная система распознавания речи // Труды Международного семинара «Диалог – 2003», 2003, С. 657-662.