Информационные процессы, Том 4 , № 2, стр. 202-220

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

Сегментация и распознавание гласных

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

Поступила в редколлегию 05.2004

Аннотация—Для сегментации речевого сигнала на квазистационарные участки использовалась математическая модель восприятия речи. Детектирование гласных звуков выполнялось синхронно с импульсами источника голосового возбуждения. Анализ формантных частот производился на основе механизма латерального торможения в спектрально-временной области. Статистический анализ результатов сегментации выполнялся на базе речевых данных для 47 человек и нескольких типов телефонных трубок и микрофонов с ручной разметкой на 127 типов артикуляторно-акустических сегментов. Границы 85% сегментов были найдены правильно независимо от типа сегмента. Среди них было правильно детектировано около 80% гласных.

1. ВВЕДЕНИЕ

Сегментация непрерывной речи в соответствии с фонетической транскрипцией является фундаментальной задачей любой голосовой системы и необходима в большинстве задач речевой технологии. Точность сегментации в значительной степени определяет надежность автоматического распознавания речи. При верификации диктора необходимо сначала сегментировать речевой сигнал на характерные элементы, определить тип сегмента, а затем проводить сравнение по различительным признакам. Тип сегмента – гласный, назальный, фрикативный или смычка, определяет методы решения обратной задачи относительно формы речевого тракта.

Ручная сегментация требует значительных затрат сил и времени, и она подвержена ошибкам. Кроме того, практически невозможно воспроизвести результаты ручной сегментации вследствие изменчивости человеческого зрительного и слухового восприятия. Автоматическая сегментация не безошибочна, однако она непротиворечива по своей сути, и её результаты воспроизводимы. В идеале хотелось бы иметь систему автоматической сегментации, способную работать с любыми языками и дикторами.

Существует два основных типа алгоритмов сегментации речи. К первому типу относятся алгоритмы, которые производят сегментацию речи при условии, что известна последовательность фонем данной фразы [1, 2]. Другой тип алгоритмов не использует априорной информации о фразе, и при этом границы сегментов определяются по степени изменения акустических характеристик сигнала [3, 4]. Существуют и другой тип алгоритмов, которые принимают решение как на основе априорной информации, так и на основе изменения акустических характеристик [5].

При сегментации желательно использовать только общие характеристики различных элементов речевого кода, поскольку на этом этапе обычно нет информации о конкретном содержании речевого высказывания.

В [6] было установлено, что стационарные и переходные участки звука вызывают активность в различных отделах слуховой коры головного мозга. Это подтверждает практически единодушное мнение исследователей о том, что сегментация речевого сигнала должна проводиться с использованием как динамических, так и статических его свойств. Динамические характеристики устойчивы к стационарным искажениям канала связи и помехам [7], тогда как анализ стационарных состояний позволяет использовать временное накопление, повышая тем самым устойчивость к помехам и искажениям. Стационарные сегменты в речи

появляются на смычках, фрикативных звуках, а также на ударных или конечных гласных. Поэтому распознавание этих элементов можно выполнять в два этапа: сначала выделить из речевого сигнала все участки со стационарными спектральными характеристиками, а затем произвести их классификацию.

Сегментация на два класса – «динамический/стационарный», по-видимому, является наиболее общим случаем.

2. ПЕРВИЧНАЯ ОБРАБОТКА РЕЧЕВОГО СИГНАЛА

На этапе первичной обработки речевого сигнала вычислялась его сонограмма, и производилась её нормировка. Для вычисления сонограммы (двумерной функции времени и частоты) использовалось быстрое преобразование Фурье по 256 отсчетам, взвешенным функцией окна Лапласа (так называемой колокольной функцией):

$$w_{lap}(t) = e^{-\alpha^2 t^2}, \tag{1}$$

где α – величина порядка 400. Спектр этого окна известен:

$$S_{lap}(\omega) = \frac{\sqrt{\pi}}{\alpha} e^{-\omega^2/4\alpha^2}. \tag{2}$$

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

Окно анализа сдвигалось через 1 мс. Для каждого спектрального профиля шкала частот преобразовывалась в шкалу мелов по формуле

$$M(f) = 1125 \cdot \ln(1 + f/700) \tag{3}$$

где f – частота в герцах, M – частота в мелах. После преобразования частотной шкалы спектральные профили содержали 129 неравномерно расположенных отсчетов. Затем выполнялась интерполяция кубическими сплайнами с представлением спектра в равномерных отсчетах через каждые 25 мел. Таким образом, в шкале мелов спектр состоял из 114 равномерных отсчетов.

Известно, что в слуховом анализаторе происходит локальное взвешенное интегрирование в частотной области [8]. Поэтому профиль сглаживался функцией, имеющей вид неравнобедренного треугольника (рис. 1). На разных этапах сегментации длины отрезков a и b на шкале мелов выбирались равными 100 и 250, или 50 и 125 мелам соответственно. Высота треугольника h выбиралась с тем условием, чтобы площадь треугольника была равна единице.

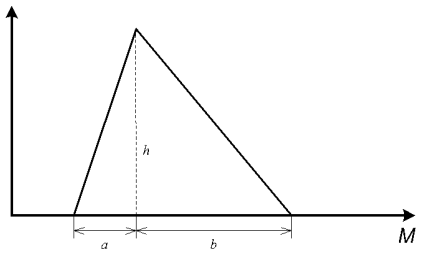

Рис. 1. Весовая функция сглаживания в виде неравнобедренного треугольника.

Каноническое представление динамического спектра речевого сигнала $S(\omega, t)$ в виде так называемой сонограммы сопровождается его логарифмированием по амплитуде. С одной стороны эта операция соответствует свойствам восприятия, а, с другой стороны, уменьшает динамический диапазон и облегчает визуальный анализ.

Механизм латерального торможения состоит в подчеркивании локальных выпуклостей спектрально-временной функции. В нашей модели этот механизм реализуется в виде логарифма отношения функции $S(\omega, t)$, усредненной на области $[t - T_1, t + T_1; \omega - \Omega_1, \omega + \Omega_1]$ к ее значению, усредненному на области $[t - T_2, t + T_2; \omega - \Omega_2, \omega + \Omega_2]$, где t и ω – центр интервала нормирования, время и частота в точке анализа:

$$D_0(\omega,t)=\log\left(\frac{\Omega_2\mathrm{T}_2\int\limits_{t-\mathrm{T}_1}^{t+\mathrm{T}_1}\int\limits_{\omega-\Omega_1}^{\omega+\Omega_1}S(w,\tau)\cdot dw\cdot d\tau}{\Omega_1\mathrm{T}_1\int\limits_{t-\mathrm{T}_2}^{t+\mathrm{T}_2}\int\limits_{\omega-\Omega_2}^{\omega+\Omega_2}S(w,\tau)\cdot dw\cdot d\tau}\right), \quad (4)$$

Здесь Ω_1,Ω_2 – половины длительностей частотных интервалов, по которым производится нормировка, $\mathrm{T}_1,\mathrm{T}_2$ – аналогичные длительности временных интервалов.

Если среднее в области $2\mathrm{T}_1\times 2\Omega_1$ равно среднему в области $2\mathrm{T}_2\times 2\Omega_2$, то $D(\omega,t)=0$. Это соответствует участкам спектра со строго линейной функцией от частоты, в частности, с постоянным уровнем. $D(\omega,t)>0$ только для выпуклых участков локально нормированного спектра. Для вогнутых участков нормированный спектр $D(\omega,t)<0$. Если ввести ограничение на отрицательные значения $D(\omega,t)$, то вогнутые участки спектра будут «невидимы». Тем самым в спектре выделяются локальные максимумы, определенные на различных частотных интервалах в различные моменты времени. Для избежания ситуации деления на нуль к знаменателю (4) добавляется некая малая константа.

В некоторых случаях необходимо сохранить вогнутые участки спектра. Для этого в (4) к выражению под логарифмом добавляется единица.

$$D_1(\omega,t)=\log\left(\frac{\Omega_2\mathrm{T}_2\int\limits_{t-\mathrm{T}_1}^{t+\mathrm{T}_1}\int\limits_{\omega-\Omega_1}^{\omega+\Omega_1}S(w,\tau)\cdot dw\cdot d\tau}{\Omega_1\mathrm{T}_1\int\limits_{t-\mathrm{T}_2}^{t+\mathrm{T}_2}\int\limits_{\omega-\Omega_2}^{\omega+\Omega_2}S(w,\tau)\cdot dw\cdot d\tau}+1\right), \quad (5)$$

Для моделирования латерального торможения только в частотной области интегрирование по времени не выполняется:

$$G_0(\omega,t)=\log\left(\frac{\Omega_2\int\limits_{\omega-\Omega_1}^{\omega+\Omega_1}S(w,t)\cdot dw}{\Omega_1\int\limits_{\omega-\Omega_2}^{\omega+\Omega_2}S(w,t)\cdot dw}\right), \quad (6)$$

$$G_1(\omega,t)=\log\left(\frac{\Omega_2\int\limits_{\omega-\Omega_1}^{\omega+\Omega_1}S(w,t)\cdot dw}{\Omega_1\int\limits_{\omega-\Omega_2}^{\omega+\Omega_2}S(w,t)\cdot dw}+1\right), \quad (7)$$

На рис. 2 показаны обычная сонограмма и сонограммы, преобразованные механизмом латерального торможения.

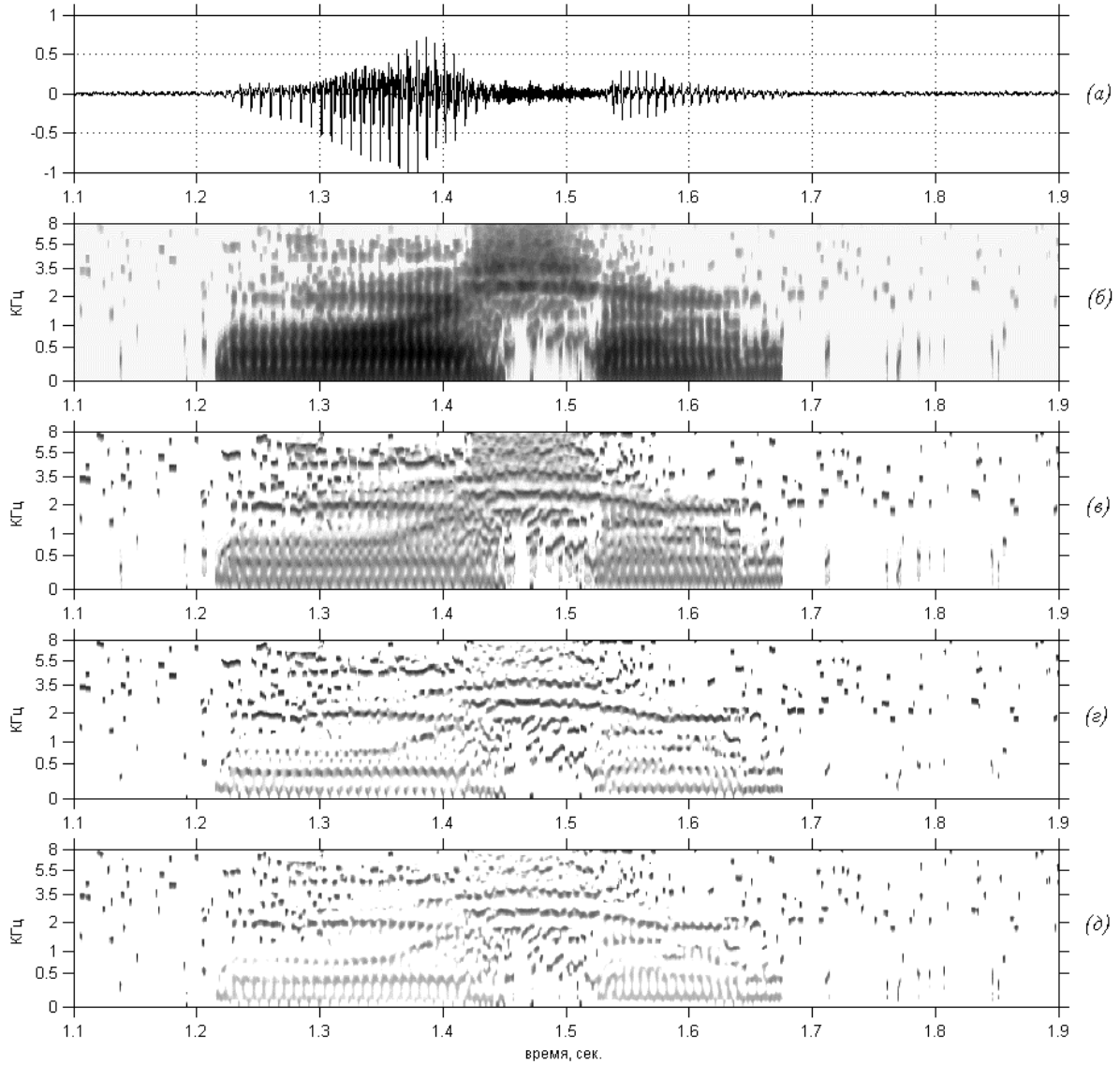

Рис. 2. Осциллограмма (а), сонограмма (б) и нормированные сонограммы (в) – $G_1(\omega,t)$, (г) – $G_0(\omega,t)$, (д) – $D_0(\omega,t)$ для слова «восемь»

3. СЕГМЕНТАЦИЯ КВАЗИСТАЦИОНАРНЫХ УЧАСТКОВ

Нормированная сонограмма слабо зависит от коэффициента усиления (уровня громкости) и наклона спектрального профиля по оси частот. Это позволяет использовать её для оценки степени однородности спектрального состава речевого сигнала безотносительно к изменениям его уровня. С этой целью исследовались различные метрики, в которых измерялось расстояние между соседними спектральными профилями. Наиболее эффективными оказались меры L_2 (средний квадрат разности) и L_4, записанные ниже одной формулой в виде парадигмы L_p:

$$m_{\mathrm{L}_p}\left(\hat{\varphi}_1^i,\hat{\varphi}_2^i\right)=\sum_i\left|\hat{\varphi}_1^i-\hat{\varphi}_2^i\right|^p,\ \text{где}\ \hat{\varphi}_k^i=\varphi_k^i\Big/\left(\sum_i\left|\varphi_k^i\right|^p\right)^{\frac{1}{p}}. \qquad (8)$$

Здесь φ_1^i и φ_2^i – сравниваемые функции, $i=\overline{1\ldots n}$. Для задачи сегментации была выбрана мера L_2, оказавшаяся в ходе экспериментов наиболее эффективной.

Нормировка расстояний может привести к большим скачкам меры близости на участках с отсутствием речи (т.е. с шумом). Для того чтобы избежать этого, к каждой из сравниваемых функций добавлялась некоторая константа c_m:

$$\tilde{\varphi}_k^i = \hat{\varphi}_k^i + c_m;\ \forall k,i \tag{9}$$

Значение этой константы должно быть достаточно большим, чтобы сгладить колебания сонограммы на участках шума.

Алгоритм автоматической сегментации состоит в следующем. Предположим, что момент начала текущего сегмента известен. Он совпадает с концом предыдущего сегмента или с началом сигнала. Продвигаясь вдоль оси времени по сонограмме от начала сегмента, рекуррентно вычисляем в каждый момент времени средний спектр текущего сегмента.:

$$\overline{S}_i(\omega) = \frac{\overline{S}_{i-1}(\omega)\cdot(i-1) + S(\omega,t_i)}{i},\ \overline{S}_0(\omega) = S(\omega,t_0). \tag{10}$$

Здесь $\overline{S}_i(\omega)$ – средний спектр на интервале $[t_0,t_i]$, $S(\omega,t_i)$ – профиль сонограммы в момент времени t_i.

Момент времени t является концом сегмента, если мера близости среднего спектра сегмента и каждого профиля спектра на интервале $[t,t+\tau_{delay}]$ превышает порог Δ_m. В этом случае накопленный средний спектр обнуляется, и возобновляется поиск конца следующего сегмента.

Известно, что спектральный состав речевого сигнала может сильно меняться от профиля к профилю. Спектры фрикативных звуков нестабильны в силу шумовой природы источника возбуждения. Поэтому для сегментов с шумовым возбуждением необходима частотно-временная фильтрация для сглаживания быстрых изменений спектра. В простейшем случае, такая фильтрация выполняется с помощью n-точечной пирамидальной весовой функции.

Изменения спектра на интервале одного периода основного тона связаны со свойствами вынужденных и свободных акустических колебаний в речевом тракте. Было найдено, что наилучший момент для отсчета спектрального профиля находится между максимумами импульса голосового возбуждения. Для поиска этого максимума использовалась нормированная сонограмма $D_0(\omega,t)$, вычисленная с параметрами $T_1 = 0$ мс, $T_2 = 1$ мс, $\Omega_1 = 5$ мел, $\Omega_2 = 150$ мел, $a = 50$ мел, $b = 125$ мел. По ней для каждого временного отсчета вычислялась средняя амплитуда спектра в некой частотной полосе $[\omega_1,\omega_2]$, и полученная одномерная функция времени $P(t)$ сглаживалась n_{pitch}-точечной пирамидальной весовой функцией. За моменты импульсов основного тона принимались положения максимумов результирующей функции. Исследования проводились в частотных полосах [0, 600]; [700, 2000]; [1800, 3500] Гц, n_{pitch} принимал значения от 1 до 7. Вид спектра $D_0(\omega,t)$ и функции $P(t)$ показан на рис. 3.

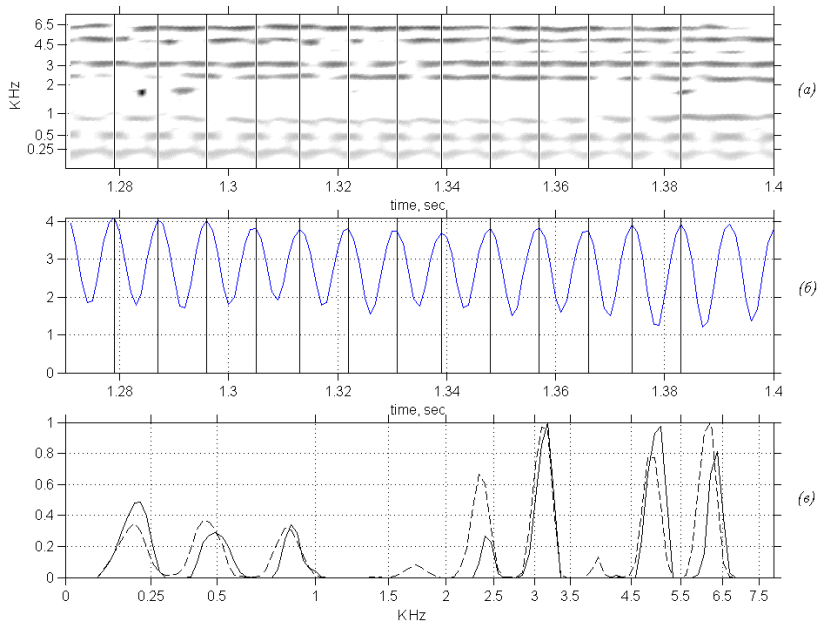

Рис. 3. Вид спектра $D_0(\omega,t)$ *(а)*, функции $P(t)$ *(б)* и среднего спектра *(в)* сегмента звука «О», вычисленного синхронно со импульсами основного тона (сплошная линия на *(в)*) и асинхронно (прерывистая линия).

Такой алгоритм поиска максимумов голосового возбуждения требует аккуратного выбора частотной полосы $[\omega_1,\omega_2]$ и коэффициента сглаживания n_{pitch}. Для оценки эффективности алгоритма был введен критерий периодичности, представляющий собой дробь, в числителе которой стоит среднеквадратичное отклонение значений периода основного тона на периоде $\sigma=\sqrt{\frac{1}{N}\sum_{i=1}^{N}T_i^2-\left(\frac{1}{N}\sum_{i=1}^{N}T_i\right)^2}$, а в знаменателе – средний период основного тона $\mu=\frac{1}{N}\sum_{i=1}^{N}T_i$:

$$C_{period}=\frac{\sigma}{\mu}. \qquad (11)$$

Очевидно, что чем устойчивее длительность периода T_{pitch}, тем ближе к нулю C_{period}, и наоборот.

Были опробованы два варианта алгоритма. В первой оставался неизменным интервал $[\omega_1,\omega_2]$, а коэффициент n_{pitch} принимал несколько значений. Во второй коэффициент n_{pitch} оставался неизменным, а интервал $[\omega_1,\omega_2]$ изменялся. Для каждого сегмента речевого сигнала выбиралась такая комбинация $[\omega_1,\omega_2]$ и n_{pitch}, при которой значение C_{period} было минимально.

Эксперименты показали, что наилучший результат достигается при использовании одного частотного интервала [0, 600] Гц и значений n_{pitch} от 1 до 7. При этом разделимость мужских и женских голосов по частоте основного тона составляла 40%. В другой схеме использовался один интервал [0, 600] Гц и одно значение n_{pitch}=3. В этом случае разделимость мужских и женских голосов составляла 48%.

Основным свойством алгоритма сегментации является его чувствительность к фонетически значимым изменениям спектра и устойчивость к случайным вариациям. Алгоритм должен быть достаточно чувствительным, чтобы точно определять положение переходов между сегментами. Но не слишком чувствительным, чтобы не реагировать на незначительные переходы внутри одного фонетического сегмента.

Чувствительностью алгоритма можно управлять с помощью набора параметров. Часть этих параметров относится к первичной обработке сигнала: используемый тип нормировки ($G_1(\omega,t)$, $G_0(\omega,t)$, или $D_0(\omega,t)$), параметры нормировки T_1, T_2, Ω_1 и Ω_2, параметры

спектрального сглаживания треугольником a и b, добавочная константа в мере L_p (8), (9) c_m, число точек в пирамидальной весовой функции временного сглаживания n_{smooth}. Часть относится непосредственно к алгоритму сегментации: порог близости профилей Δ_m, длительность интервала сравнения τ_{delay}.

Качество работы алгоритма можно оценить, сравнивая разметку речевых сигналов, выполненную вручную, с автоматической разметкой. При этом следует учитывать, что на протяжении некоторых фонетических и артикуляторных сегментов спектр сигнала может значительно изменяться. Например, в слове «один» между первым гласным звуком «А» и звонким смычным «Д» присутствует гласный сегмент, похожий по спектральному составу на «И». Это связано с особенностями артикуляции. При автоматической сегментации такие сегменты могут быть разбиты на несколько частей. По этой причине сравнение ручной и автоматической разметки позволяет оценить эффективность алгоритма лишь приблизительно.

Для оценки эффективности следует использовать набор показателей, характеризующих как точность нахождения границ сегментов, так и процент пропусков переходов. Достаточными являются, например, следующие три показателя: Π_{lost} – процент пропущенных квазистационарных сегментов; $\Pi_{accurate}$ – процент границ, положение которых найдено с точностью не менее 15 мс; Π_{found} – отношение числа автоматически найденных сегментов к числу указанных разметчиком вручную. Следует заметить, что равенство Π_{found} единице не влечёт за собой равенство нулю Π_{lost}, так как одному указанному вручную сегменту может быть поставлено в соответствие несколько найденных автоматически. Каждому автоматически найденному сегменту ставится в соответствие сегмент разметки, имеющий с ним наибольшую пересекаемость.

Оптимизация параметров алгоритма сегментации должна была бы заключаться в подборе таких значений параметров, при которых показатели качества принимают наилучшие значения. То есть такого типа нормировки, параметров T_1, T_2, Ω_1, Ω_2, a, b, c_m, n_{smooth}, Δ_t и Δ_m, при которых Π_{lost} минимально, $\Pi_{accurate}$ максимально, а Π_{found} наиболее близко к единице.

Несмотря на то, что набор возможных значений параметров относительно невелик, число их сочетаний велико, что делает задачу оптимизации сложной. Так как наилучший порог Δ_m зависел от остальных параметров, то для нахождения наилучшей комбинации значений мы задавались неким набором первичных параметров, и для каждого из таких наборов перебирали несколько значений Δ_m.

Оптимизация выполнялась, исходя из следующих соображений. Предполагается, что большинство границ между сегментами речи будет найдено с помощью детекторов переходных процессов, описанных в [7]. Поэтому была задана допустимая доля пропуска границ квазистационарных сегментов в размере 15% от переходов, зарегистрированных в базе данных с ручной разметкой (т.е. Π_{lost}=0,15). Для этого значения Π_{lost} был сконструирован одномерный критерий

$$C_{segm} = \Pi_{accurate} - 0,2\left(\Pi_{found} - 1\right), \qquad (12)$$

согласно которому и проводилось сравнение результатов сегментации.

Равенство Π_{lost}=0,15 достигалось за счет выбора порога Δ_m и интерполяции полученных результатов. Кроме того, детальное исследование показало, что большинство пропущенных переходов либо относятся к очень коротким сегментам, либо выражены слабо. Как видно из таблицы 1, более половины пропущенных границ связано именно с такими переходами.

Таблица 1. Список пропущенных переходов при оптимальных значениях параметров алгоритма.

переход		процент от числа пропущенных	процент от числа переходов в разметке
с	на		
Т	Т!	5.20	1.56
С'	\/	4.80	1.44
Д	Д!	3.64	1.09
Д'	Д'!	3.64	1.09
Р'	Р'!	3.57	1.07
ъ/	Р'	3.56	1.07
Д'	Дh'	2.90	0.87
Т	Т!h	2.88	0.87
ъ/	В	2.68	0.80
Н	Н!	2.25	0.68
С	\/	2.19	0.66
Т'	Th'	1.97	0.59
Ш	\/	1.73	0.52
П'	П'!	1.72	0.52
М'	М'!	1.58	0.47
Ы	Р'	1.53	0.46
В'	ь	1.44	0.43
Р'	И	1.39	0.42
В	А	1.05	0.31
Р'	и	1.05	0.31
С	Т	1.03	0.31
ь	Т	0.93	0.28
ь	М'	0.89	0.27
Р'!	И	0.88	0.27
ь	М	0.84	0.25

В соответствии с выбранным критерием наилучшие результаты получены при следующих значениях параметров: тип нормировки $G_1(\omega,t)$, $\Omega_1 = 5$, $\Omega_2 = 150$, $a = 50$, $b = 125$, $c_m = 10$, $\Delta_t = 15$, $\Delta_m = 1.3 \cdot 10^{-4}$, $n_{smooth} = 30$. При этом общее отношение числа найденных границ к числу границ по разметке $\Pi_{found} = 1.36$, а доля достаточно точных границ по времени $\Pi_{accurate} = 0.61$. Учитывая тот факт, что обычно по крайней мере одна из границ между квазистационарными участками попадает между символами ручной разметки, такой показатель точности детектирования границ кажется приемлемым.

На рис. 4 показаны варианты процесса оптимизации при оптимальных значениях параметров алгоритма.

Пример результата работы алгоритма показан на рис. 5. На нем изображена осциллограмма речевого сигнала для числительного «61» с ручной разметкой и нормированная сонограмма $G_1(\omega,t)$ с автоматически найденными границами сегментов.

Таким образом, автоматическая сегментация на квазистационарные участки речевого сигнала, с одной стороны, демонстрирует достаточно близкое соответствие ручной разметке, и, с другой стороны, не требует никаких предварительных указаний о составе речевого сигнала.

Исследования проводились на материале 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. Они представлены в таблице 2. Частота дискретизации сигналов составляла 16 КГц.

Таблица 2. Артикуляторные и фонетические элементы алфавита разметки.

Тип сегмента	Символ
Гласные ударные:	А, Э, О, У, Ы, И
Гласные предударные:	а, э, о, у, ы, и
Гласные безударные (редуцированные):	ъ, ь, уу
Сегмент последнего безударного гласного:	ъ_, ь_, уу_
Сегмент последнего ударного гласного:	А_, Э_, О_, У_, Ы_, И_, Я_,Е_,Ё_,Ю_
Огласованный сегмент:	ъ/, ь/, у/
Полугласный язычный звонкий:	Й
Полугласный язычный глухой:	й
Дифтонги ударные:	Я, Е, Ё, Ю
Дифтонги безударные:	я, е, ё, ю
Согласные твердые:	Б, В, Г, Д, Ж, З, К, Л, М, Н, П, Р, С, Т, Ф, Х, Хг, Ц, Ш
Согласные мягкие:	Б', В', Г', Д', Ж', З', К', Л', М', Н', П', Р', С', Т', Ф', Х', Ц', Ч, Ш', Хг'
Смычка аспиративная твердая:	Бв, Дз, Дж, Пф, Кх, Рш, Тс
Смычка аспиративная мягкая:	Бв', Дз', Дж', Пф', Кх', Рш', Тс'
Взрыв твердый:	Б!, Д!, Г!, П!, Т!, К!, Л!, Р!, М!, Н!
Взрыв мягкий:	Б'!,Д'!,Г'!,П'!,Т'!,К'!, Л'!, Р'!, М'!, Н'!
Взрыв аспиративный твердый:	Б!h, Д!h, Г!h, П!h, Т!h, К!h, Л!h, Р!h
Взрыв аспиративный мягкий:	Бh', Дh', Гh', Пh', Тh', Кh', Лh', Рh'
Короткая пауза (провал после фрикативных, взрывов и т.д. "epenthetic"):	\/
Фарингальный взрыв:	gb; перед начальными гласными
Аспиративный сегмент:	vh; придыхание перед началом слова или по окончании слова
Неречевые сегменты:	
Не речь (шум канала) :	h#
Пауза (между раздельными словами):	z#
Короткая пауза ("провал") между словами в слитном тексте):	][
Неизвестный:	?
Чмок:	ch

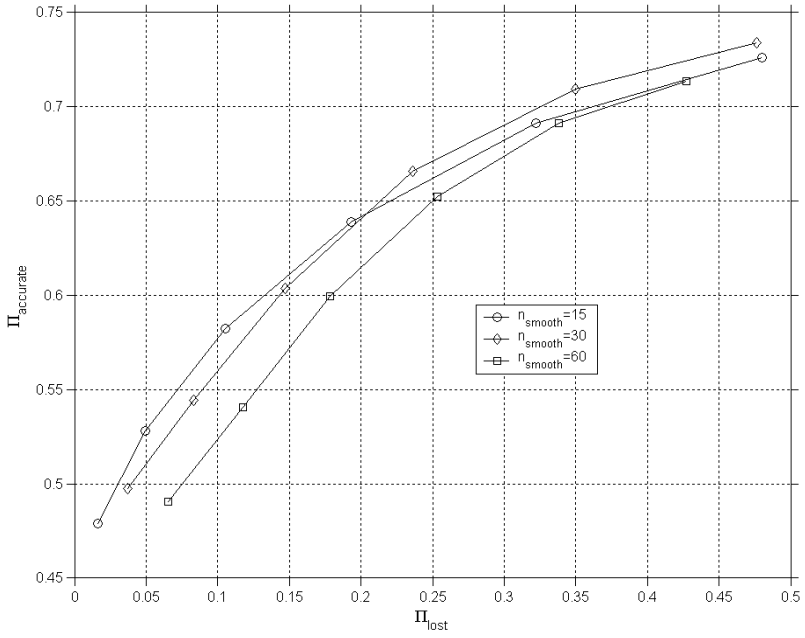

Рис. 4. Графики зависимости $\Pi_{accurate}$ от Π_{lost} для различных Δ_m при оптимальных параметрах алгоритма для трёх возможных значений n_{smooth}.

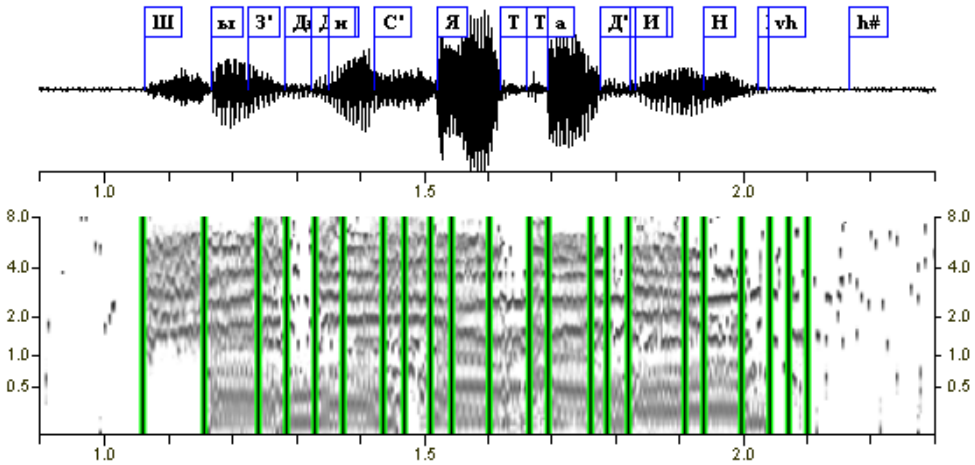

Рис. 5. Пример автоматической сегментации речевого сигнала.

4. СЕГМЕНТАЦИЯ И АНАЛИЗ ГЛАСНЫХ

Гласные звуки играют важную роль в речевом коде. Они обладают наибольшей энергией, что соответствует лучшему отношению сигнал/шум. Гласные характеризуются, в основном, резонансными частотами речевого тракта, которые в спектральной области обычно выглядят как локальные максимумы энергии, называемые формантами. Наличие более или менее длительного стационарного участка на гласных позволяет использовать накопление с целью более надежного определения формантных частот. Большинство переходных процессов к согласным и от них проявляется в виде изменения амплитуды и частотного состава гласных.

Определение формантных частот гласных звуков в спектральной области требует синхронизации момента отсчета спектра с импульсами голосового возбуждения. Анализ устойчивости определения формантных частот показал, что момент отсчета спектра должен быть сдвинут по времени вперед относительно максимума энергии функции $P(t)$, описанной в предыдущем разделе, примерно на 40% от интервала между последовательными максимумами.

По результатам различных исследований принято считать, что формантные частоты русских гласных распределены так, как это показано в таблице 3.

Таблица 3. Канонические значения формантных частот гласных звуков, герц.

звук	F_1	F_2	F_3
А	600	1200	2400
Э	450	1700	2500
О	450	800	2500
У	350	750	2200
Ы	300	1800	2450
И	300	2200	3000
Е	400	1850	2600
Я	450	1500	2450

Однако в каждой конкретной реализации гласного какие-то из этих частот могут быть по разным причинам либо плохо определяться, либо появляются дополнительные пики в спектре. Причинами пропадания пиков могут быть либо потери энергии в речевом тракте, либо особенности амплитудно-частотной характеристики канала речевой связи, либо высокий уровень шумов. По этим же причинам могут появляться и дополнительные, "ложные", пики. Кроме того, дополнительные пики в спектре могут появиться и при назализации при не полностью поднятой небной занавеске, или за счет резонансов подсвязочной области. При решении обратной задачи важно точно определить, какие формантные частоты определяются формой речевого тракта. При классификации гласных в задачах распознавания речи или верификации диктора также нужно установить, какие из частот имеют «фонетический» смысл.

Нормирование спектра по (4 – 7) позволяет найти все локальные экстремумы спектра, но, поскольку при этом теряется информация об амплитуде, даже мелкие неоднородности спектра могут претендовать на форманты. Эта ситуация иллюстрируется на рис. 6, где видны пики

энергии, не соответствующие установившемуся представлению о фонетически значимых формантных частотах гласного /О/.

При построении алгоритма определения положения формант на спектральном профиле мы исходим из предположения, что каждой «фонетической» форманте соответствует один пик профиля нормированной сонограммы $G_0(\omega,t)$. Остальные пики следует считать ложными. Задача состоит в выделении формантных пиков и фильтрации ложных. Для решения этой задачи необходим перебор всех возможных комбинаций формантных пиков, среди которых будет выбрана наиболее правдоподобная.

Возможной комбинацией формантных пиков будем считать набор пиков спектра с частотами, расположенными строго по возрастанию, т.е. $F_i < F_{i+1}$, и лежащими в интервалах соответствующих формант. Кроме того, частота первой форманты должна быть строго больше частного основного тона F_0. Таким образом, ограничительное условие можно записать так:

$$\begin{cases} F_i < F_{i+1}, i = 0\ldots 3, \\ F_1 \in [150, 900], \\ F_2 \in [700, 2700], \\ F_3 \in [1800, 4000]. \end{cases} \tag{13}$$

Здесь частоты указаны в герцах. Частотные интервалы выбраны с учетом всех возможных положений формант для мужского и женского голосов.

В качестве критерия $C_{formants}$ для осуществления выбора наилучшей комбинации пиков мы использовали сумму амплитуд спектрального профиля S, взятых на частотах положения пиков. Пиками считались все локальные максимумы профиля D_0^{ω}, имеющие амплитуды больше порога 0,02. Этот порог нужен для фильтрации очень малых пиков. На рис. 6 таковые расположены на частотах 0,62 и 1,4 КГц.

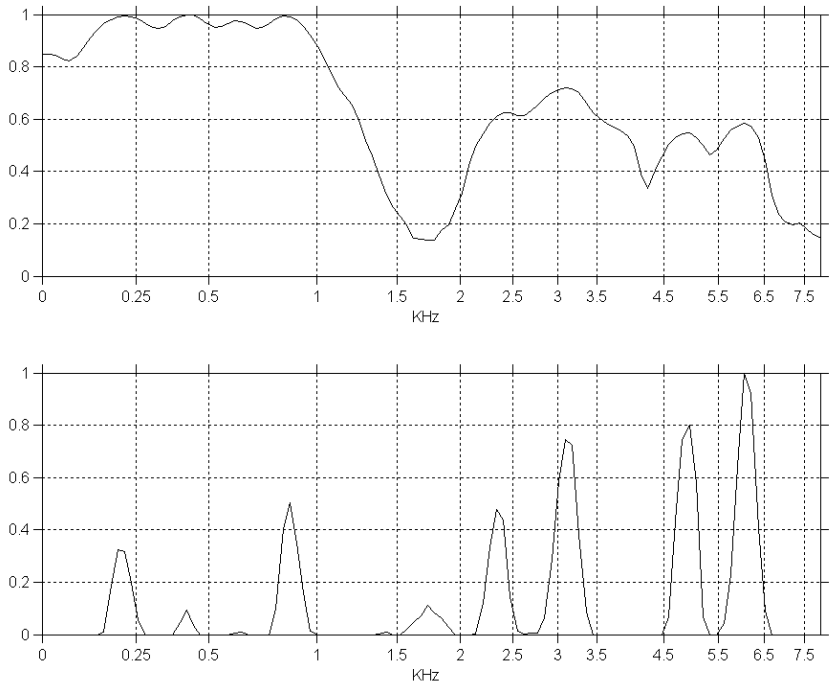

Рис. 6. Профили сонограммы S (сверху) и D_0^{ω} (снизу) для звука «О». Профили нормированы к единице по максимальной амплитуде.

Практика показала, что более точный результат получается, если искать не первые три, а первые четыре форманты, отбрасывая затем последнюю. При этом можно ограничить по частоте интервал поиска четвертой форманты снизу тремя килогерцами.

В таблице 4 показано значение критерия $C_{formants}$ и комбинации первых 4-х формантных частот для различных гипотез о нумерации формант для профилей на рис. 6. Строки отсортированы по убыванию $C_{formants}$. Все частоты указаны в герцах.

Таблица 4. Значение критерия $C_{formants}$ для различных гипотез о нумерации 4-х формант.

$C_{formants}$	F_1	F_2	F_3	F_4
0.8300	416	858	2334	3089
0.8277	194	858	2334	3089
0.8231	416	858	3089	6053
0.8208	194	858	3089	6053
0.8143	416	858	3089	4953
0.8120	194	858	3089	4953
0.7958	416	858	2334	6053
0.7935	194	858	2334	6053
0.7869	416	858	2334	4953
0.7846	194	858	2334	4953
0.7279	416	2334	3089	6053
0.7256	194	2334	3089	6053

Результаты работы описанного алгоритма для ударных гласных, представленных в нашей базе данных, приведены в таблице 5 отдельно для мужчин и женщин.

Таблица 5. Средние значения найденных формантных частот ударных гласных.

мужские голоса

звук	F_1	F_2	F_3
А	500	1180	2380
Э	350	1690	2400
О	380	1090	2420
У	340	980	2540
Ы	300	1810	2510
И	290	1850	2630
Е	320	1850	2620
Я	410	1600	2460

женские голоса

звук	F_1	F_2	F_3
А	540	1140	2650
Э	400	1980	2770
О	430	1200	2790
У	410	1120	2780
Ы	390	2010	2800
И	380	1960	2860
Е	410	2040	2920
Я	470	1570	2520

Как видно из этих таблиц, формантные частоты примерно соответствуют среднестатистическим значениям, указанным в таблице 3. Некоторое повышение значений частот для женских голосов, очевидно, связано с более короткими речевыми трактами женщин. Вместе с тем, для различных произнесений и дикторов на графиках рассеивания этих частот наблюдается большое разнообразие. Например, на рис. 7 показаны распределения частот для гласного /Э/ (мужские голоса).

Более подробный статистический анализ структуры распределения формантных частот был выполнен с использованием аппроксимации смесью нормальных распределений

$$p(x,\Theta)=\sum_{m=1}^{K}\alpha_m p_m(x,\theta_m). \qquad (14)$$

Здесь α_m – веса компонент смеси, удовлетворяющие условиям $\alpha_m \ge 0$ и $\sum_{m=1}^{K}\alpha_m = 1$, $p_m(x,\theta_m)$ – многомерная плотность вероятности m-й компоненты, $\Theta=\{(\alpha_m,\theta_m), m=1,\ldots,K\}$ – набор параметров смеси.

Аппроксимация распределения частот выполнялась с помощью EM-алгоритма [9-13], позволяющего найти наиболее правдоподобную комбинацию взвешенных многомерных нормальных распределений, породившую имеющуюся выборку.

На рис. 8 показаны результаты аппроксимации для ударного гласного /Э/. Здесь и далее эллипсы задаются ковариационной матрицей выборки (расстояние одной дисперсии).

Вес и частоты центров нормальных распределений для ударных гласных приведены в таблице 6.

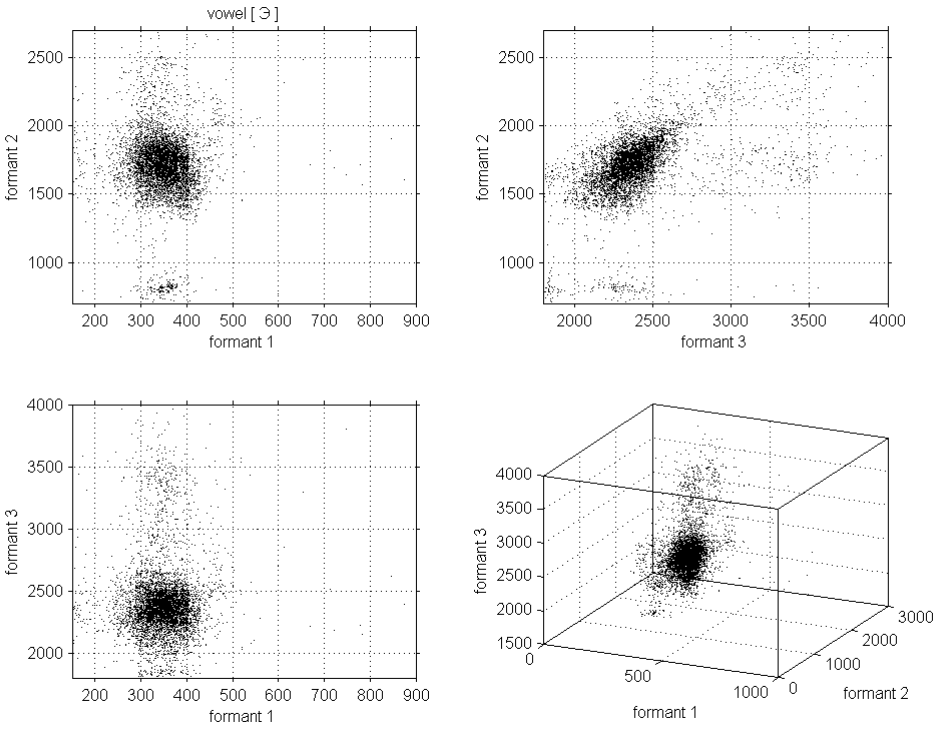

Рис. 6. Распределение формантных частот для гласного /Э/ (мужские голоса).

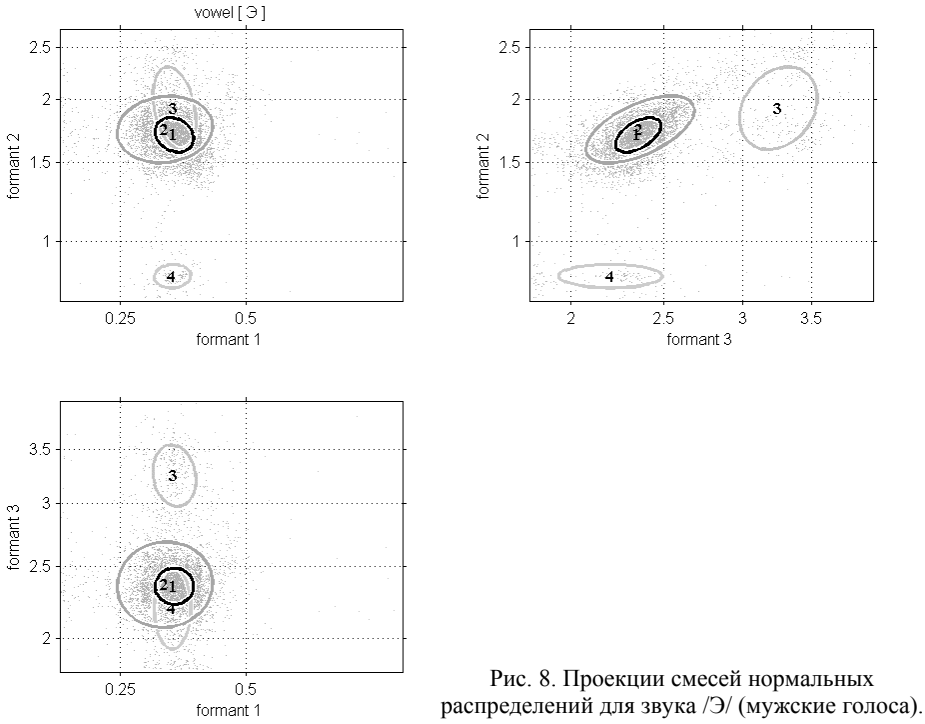

Рис. 8. Проекции смесей нормальных распределений для звука /Э/ (мужские голоса).

Таблица 6. Веса и частоты центров нормальных распределений для ударных гласных.

звук «Э»

вес	F_1	F_2	F_3
0.748	350	1706	2352
0.161	332	1750	2364
0.062	350	1925	3250
0.029	347	818	2202

звук «О»

вес	F_1	F_2	F_3
0.695	381	1116	2287
0.150	381	1027	3218
0.110	360	811	2210
0.044	369	1923	2676

звук «А»

вес	F_1	F_2	F_3
0.372	536	1269	2369
0.344	536	1119	2201
0.197	381	1134	2344
0.087	501	1139	3356

звук «У»

вес	F_1	F_2	F_3
0.934	336	923	2512
0.066	476	2109	3018
–			
–			

звук «Ы»

вес	F_1	F_2	F_3
0.517	298	1816	2386
0.297	305	1866	2451
0.127	288	2124	3395
0.058	331	1036	2220

звук «И»

вес	F_1	F_2	F_3
0.704	287	1929	2538
0.171	279	2119	3371
0.076	351	1429	2308
0.049	296	803	2282

звук «Е»

вес	F_1	F_2	F_3
0.459	315	1942	2621
0.312	308	1768	2465
0.159	309	2106	3182
0.071	376	1152	2230

звук «Я»

вес	F_1	F_2	F_3
0.430	432	1510	2397
0.270	392	1763	2439
0.178	398	1698	2421
0.123	410	1423	2766

5. КЛАСТЕРИЗАЦИЯ В ПРОСТРАНСТВЕ ФОРМАНТНЫХ ЧАСТОТ

Определим степень неопределенности (пересекаемость) в классификации гласных звуков как общий объем под многомерными функциями плотностей вероятности «чужих» элементов («не-класса») и «своих» («класса»). То есть объем под функцией плотности вероятности «не-класса» вычисленный в области, где плотность вероятности «класса» превышает плотности вероятности «не-класса», плюс объем под функцией плотности вероятности «класса» вычисленный в области, где плотность вероятности «не-класса» превышает плотности вероятности «класса».

Пересекаемость трехмерных распределений формантных частот характеризует разделимость гласных в пространстве формант. Взаимные пересекаемости пар соответствующих смесей нормальных распределений приведены в таблице 7.

Таблица 7. Взаимные пересекаемости пар смесей нормальных распределений, %.

звук	А	Э	О	У	Ы	И	Е	Я
А	100	6	40	18	7	8	7	17
Э	6	100	10	6	45	38	42	45
О	40	10	100	46	9	12	9	16
У	18	6	46	100	6	10	7	8
Ы	7	45	9	6	100	64	51	25
И	8	38	12	10	64	100	65	24
Е	7	42	9	7	51	65	100	33
Я	17	45	16	8	25	24	33	100

Пересекаемости смеси нормальных распределений для каждого гласного с равновзвешенной комбинацией остальных смесей («свои» против «чужих») приведены в таблице 8.

Таблица 8. Взаимные пересекаемости смесей нормальных распределений, % («свои» против «чужих»).

звук	А	Э	О	У	Ы	И	Е	Я
%	23	39	32	19	40	45	45	37

Ручная разметка речевых сигналов очень трудоемка и несет на себе отпечаток субъективного мнения разметчика. Поэтому была предпринята попытка определения «объективно» существующих классов в пространстве формантных частот. С этой целью формантные частоты всех ударных гласных были объединены. Затем с помощью EM-алгоритма был выполнен кластерный анализ, т. е. найдены центры сгущений в трехмерном распределении.

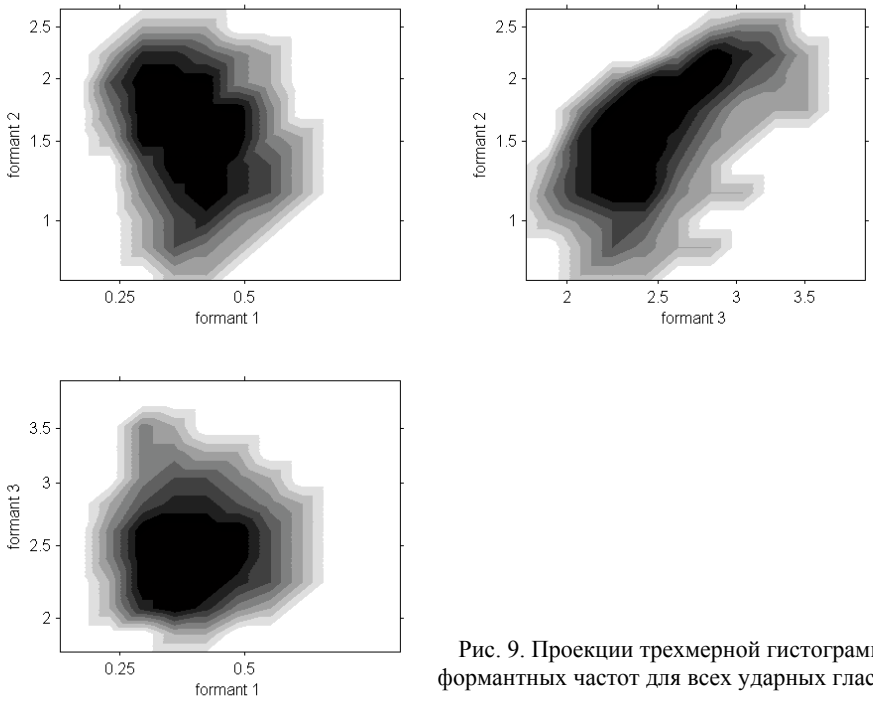

Рис. 9. Проекции трехмерной гистограммы формантных частот для всех ударных гласных.

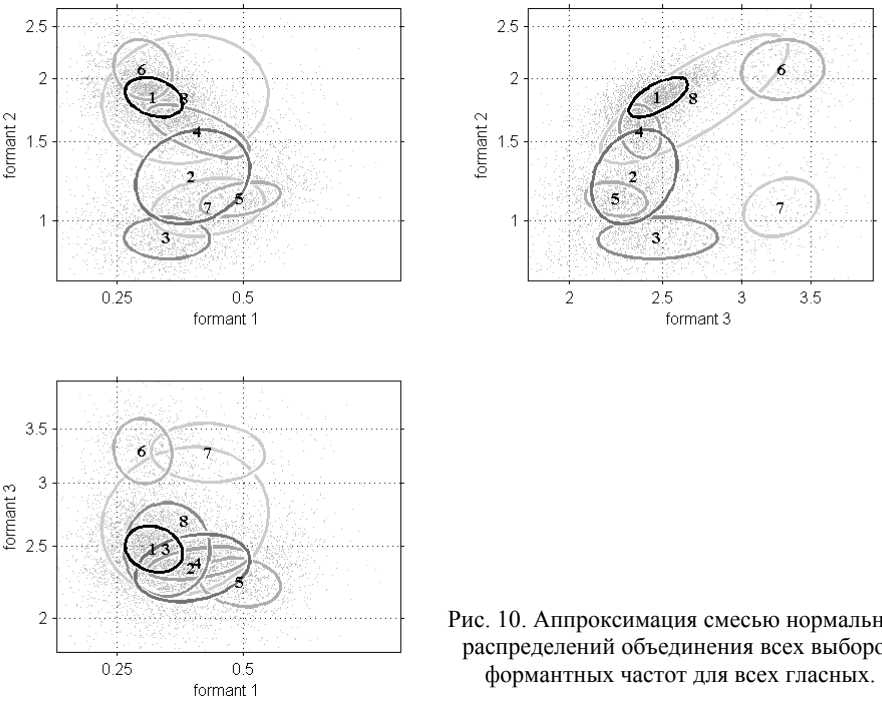

Рис. 10. Аппроксимация смесью нормальных распределений объединения всех выборок формантных частот для всех гласных.

На Рис. 9 изображены проекции трехмерной гистограммы объединения всех выборок формантных частот для всех гласных. Представительности выборок были предварительно выровнены.

На рис. 10 показаны результаты кластеризации, а в таблице 9 приведены веса и частоты первых 8 кластеров. Количество кластеров было выбрано так, чтобы оно равнялось количеству гласных, участвующих в кластеризации. В таблице 10 показаны принадлежности гласных к компонентам смеси. В таблице 11 показаны взаимные пересекаемости пар компонент смеси нормальных распределений. В таблице 12 показаны взаимные пересекаемости нормальных распределений («свои» против «чужих»).

Таблица 9. Веса и частоты центров нормальных распределений для объединения всех выборок.

№	вес	F_1	F_2	F_3
1	0.365	317	1842	2475
2	0.181	392	1262	2339
3	0.138	342	907	2471
4	0.103	403	1576	2373
5	0.082	492	1124	2239
6	0.071	297	2082	3287
7	0.032	423	1076	3273
8	0.029	377	1827	2690

Таблица 10. Принадлежность гласных к компонентам смеси, %.

	1	2	3	4	5	6	7	8
А	0	11	6	15	56	0	9	3
Э	52	5	3	29	0	6	0	5
О	1	24	26	3	29	2	13	2
У	1	4	67	0	8	2	14	4
Ы	64	3	3	13	0	14	0	3
И	54	3	6	7	0	22	0	7
Е	65	2	3	8	0	15	0	7
Я	22	9	2	52	3	3	2	9

Таблица 11. Взаимные пересекаемости пар нормальных распределений, %.

Звук	1	2	3	4	5	6	7	8
1	99	12	0	37	0	6	0	18
2	12	99	28	30	30	2	4	38
3	0	28	99	1	15	0	16	5
4	37	30	1	99	7	1	0	20
5	0	30	15	7	99	0	1	8
6	6	2	0	1	0	99	1	23
7	0	4	16	0	1	1	99	2
8	18	38	5	20	8	23	2	99

Таблица 12. Взаимные пересекаемости нормальных распределений, % («свои» против «чужих»).

звук	1	2	3	4	5	6	7	8
%	15	45	16	20	13	7	6	38

Как видно, пересекаемость кластеров в ряде случаев значительно меньше, чем пересекаемость гласных, и, в среднем, пересекаемость уменьшилась почти вдвое.

6. ДЕТЕКТИРОВАНИЕ ГЛАСНЫХ

Найденные центры кластеров могут использоваться для определения, является ли данный речевой сегмент гласным звуком. Поскольку гласные характеризуются не только резонансными частотами, но и затуханием резонансных колебаний и отношением амплитуд формант, то в качестве признака принадлежности к гласному сегменту было выбрано средневзвешенное расстояние d от данного спектрального профиля сегмента до характерных спектральных профилей основных гласных (то есть до тех профилей, чьи резонансные частоты были наиболее близки к центрам кластеров).

$$d = 1 - \sum_{i=1}^{K} \alpha_i \left(1 - m_{\mathrm{L}\,p}\left(S_{current}, S_i\right)\right), \tag{15}$$

где α_i – вес i-й компоненты аппроксимирующей смеси нормальных распределений, K – количество компонент смеси, $m_{\mathrm{L}\,p}$ – мера близости (8), $S_{current}$ – средний спектр данного сегмента, S_i – характерный спектр i-й компоненты смеси.

Исследовались метрики L_2 и L_4. Для каждой метрики вычислялись распределения расстояний для гласных и негласных звуков. В качестве гласных учитывались не только ударные, но и безударные звуки.

Гласные звуки образуются с участием голосового источника. Поэтому в качестве второго признака гласных использовался показатель периодичности возбуждение на интервале сегмента C_{period} (описан в разделе 3).

Мерой близости сегмента к гласному звуку считается апостериорная вероятность

$$P\left(V \mid d, p\right) = \frac{1}{1 + \dfrac{P\left(d \mid \overline{V}\right)}{P\left(d \mid V\right)} \cdot \dfrac{P\left(p \mid \overline{V}\right)}{P\left(p \mid V\right)} \cdot \dfrac{P\left(\overline{V}\right)}{P\left(V\right)}} \tag{16}$$

Здесь обозначим:
V – гласный, $\overline{V}$ – негласный, d – расстояние, p – периодичность, т.е. C_{pitch}; $P\left(V \mid d, p\right)$ – вероятность того, что данный сегмент является гласным звуком при данном расстоянии d и периодичности p; $P\left(d \mid V\right)$ и $P\left(d \mid \overline{V}\right)$ – распределение расстояний для гласных и для негласных; $P\left(p \mid V\right)$ и $P\left(p \mid \overline{V}\right)$ – распределение периодичности для гласных и для негласных; $P\left(V\right)$ и $P\left(\overline{V}\right)$ – априорные вероятности появления гласного и негласного звуков. Для русского языка можно принять $P\left(V\right) = P\left(\overline{V}\right)$.

Пример распределения двумерных апостериорных вероятностей на плоскости (*d,p*) показаны на рис. 11.

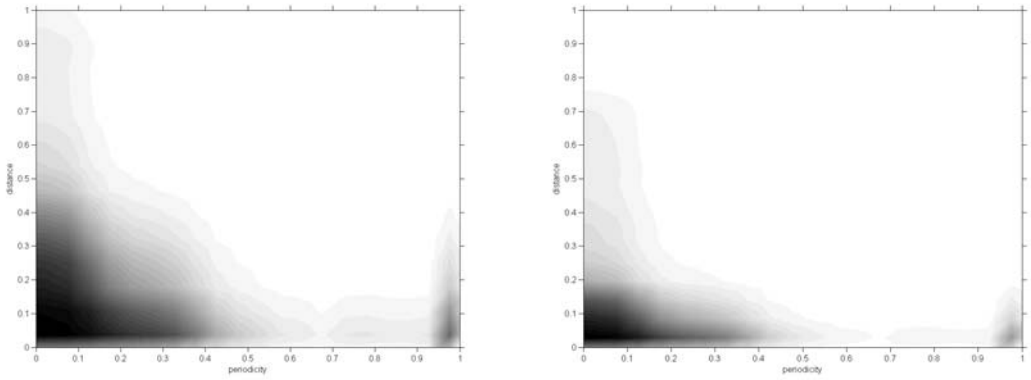

Рис. 11. Двумерные апостериорные вероятности $P\left(V \mid d, p\right)$ для меры L_2 (слева) и L_4 (справа).

На основе полученного двумерного распределения $P\left(V \mid d, p\right)$ для всех сегментов базы речевых данных были вычислены апостериорные вероятности того, что данный сегмент является гласным. На рис. 12 сверху изображены распределения апостериорных вероятностей $P\left(V \mid d, p\right)$ для гласных (сплошная линия) и негласных (прерывистая линия). Снизу по оси абсцисс отложены значения пороговой вероятности. Если $P\left(V \mid d, p\right)$ выше этого порога, считаем звук гласным, и наоборот. По оси ординат отложено относительное количество реальных гласных, распознанных при данном пороге как негласные (сплошная линия), и

относительное количество реальных негласных, распознанных при данном пороге как гласные (прерывистая линия). В идеальном случае оба эти относительных количества равны нулю. На рисунках видно, что оптимальное значение порога находится в районе 0,5. При этом правильно распознаются 77% гласных и 78% негласных при использовании меры L_2, и 70% гласных и 79% негласных при использовании меры L_4.

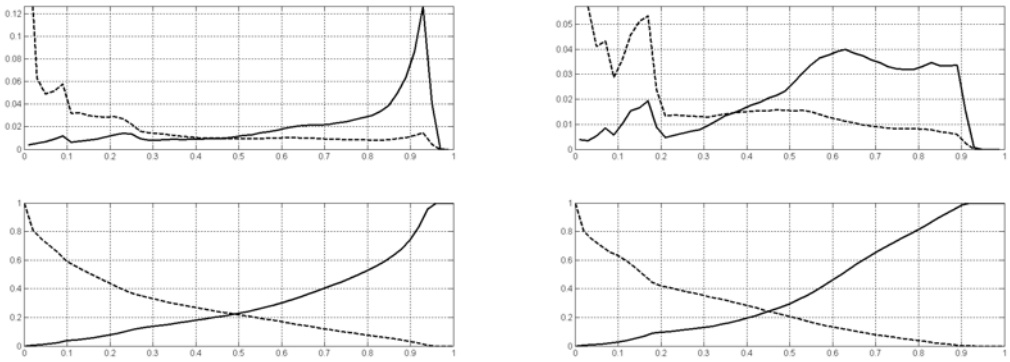

Рис. 12. Распределения апостериорных вероятностей $P(V \mid d, p)$ (сверху) и доля распознанных неверно сегментов в зависимости от порога (снизу). Для меры L_2 (слева) и L_4 (справа). Гласные (сплошная линия) и негласные звуки (прерывистая линия).

В таблице 13 приведен список неправильно распознанных гласных и негласных сегментов при значении порога 0,5 для меры L_2. Для L_4 список почти идентичен. Видно, что среди гласных распознаваемость мала для коротких (плохая периодичность) и слабовыраженных, редуцированных (большое расстояние в пространстве профилей спектров) сегментов. В свою очередь, среди негласных звуков за гласные приняты, во-первых, назальные, во-вторых, звонкие звуки (хорошая периодичность). На принятых за гласные глухих звуках в большинстве случаев имела место реверберация, из-за чего сигнал определялся как периодичный, а профили были близки к характерным для гласных профилям.

Таблица 13. Список неправильно распознанных гласных и негласных сегментов для меры L_2.

гласные				негласные		
тип	процент от числа пропущенных	процент от общего числа		тип	процент от числа пропущенных	процент от общего числа
ь	28.11%	6.44%		Н	12.71%	2.78%
и	13.82%	3.16%		ъ/	9.97%	2.19%
ъ	13.75%	3.15%		В	9.44%	2.07%
О	12.52%	2.87%		М'	9.25%	2.03%
И	8.85%	2.03%		В'	5.17%	1.13%
Е	7.81%	1.79%		Л'	4.51%	0.99%
Ы	4.13%	0.95%		Т	4.48%	0.98%
А	2.38%	0.55%		М	4.26%	0.93%
Я	2.35%	0.54%		С'	3.95%	0.87%
Э	1.73%	0.40%		Д'	3.93%	0.86%
ы	1.70%	0.39%		Т!	2.64%	0.58%
а	1.38%	0.32%		Т!h	2.43%	0.53%

7. ЗАКЛЮЧЕНИЕ

В данной работе показано, что использование математической модели восприятия (шкала частот в мелах, локальное взвешенное интегрирование в частотной области, механизм латерального торможения в частотно-временной области) позволяет эффективно производить анализ речевого сигнала.

Предложены алгоритмы сегментации речевого сигнала на квазистационарные участки на основе латерального торможения, детектирования и распознавания гласных звуков. 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.

При сегментации речевого сигнала границами сегментов считались участки, на которых значительно менялась спектральная структура сигнала. На вокализованных сегментах отсчеты спектра брались синхронно с импульсами голосового источника. На имеющейся базе данных алгоритм сегментации определил положение 85% границ сегментов, в том числе 61% с точностью не хуже 15 мс. При этом было найдено на 36% больше границ, чем отмечено при ручной разметке. Показано, что основную часть пропущенных границ составляли слабовыраженные переходы, либо переходы между короткими сегментами.

Распознавание гласных звуков производилось на основе формантной структуры спектра. Показано соответствие полученных формантных частот каноническим частотам. Аппроксимация выборок формантных частот осуществлялась смесью нормальных распределений.

Определение вероятности принадлежности сегмента к гласным звукам выполнялось на основе коэффициента периодичности и меры близости спектров данного сегмента и спектров гласных звуков. Правильно распознано около 80% звуков.

Работа выполнена при поддержке РФФИ, грант № 03-01-00116.

СПИСОК ЛИТЕРАТУРЫ

1. Rabiner L.R., Rosenberg A.E., Wilpon J.G. and Zampini T.M. A bootstrapping training technique for obtaining demisyllable reference patterns. *JASA*, 1982, vol. 71, no. 6, pp. 1588-1595.
2. Ganapathiraju A., Hamaker J., Picone J., Doddington G.R. and Ordowski M. Syllable-Based Large Vocabulary Continuous Speech Recognition. *IEEE Transactions on Speech and Audio Processing*, 2001, vol. 9, no. 4, pp. 358-366.
3. Wilpon J.G., Juang B-H. and Rabiner L.R. An investigation on the use of acoustic sub-word units for automatic speech recognition. *Proc. Int. Conf. Acous., Speech, and Sig. Processing*, Dallas, TX, 1987, pp. 821-824.
4. Kamakshi Prasad, Nagarajan, Hema Murthy Automatic segmentation of continuous speech using minimum phase group delay functions. *Speech* Communication, 2004, vol. 42, pp. 429-446.
5. Van Hemert J.P. Automatic segmentation of speech. *IEEE Transactions on Signal Processing*, 1991, vol. 39, pp. 1008-12.
6. Seifritz E., Esposito F., Hennel F., Mustofic H., Neuhoff J.G., Bilecen D., Tedeschi G., Scheffler K., Di Salle F. Spatiotemporal pattern of neural processing in the human auditory cortex. *Science*, 2002, vol. 297, no. 5587, pp. 1706-1708.
7. Сорокин В.Н. Модель многослойного первичного анализа речевых сигналов. *Труды 13-й сессии Российского акустического общества*, 2003, стр. 11-16.
8. Чистович Л.А., Венцов А.И., Гранстрем М.П. и др. *Физиология речи. Восприятие речи человеком.* М.: Наука, 1976, стр. 388.
9. Dempster A.P., Laird N.M. and Rubin D.B. Maximum-likelihood from incomplete data via the EM algorithm. *J. Royal Statist. Soc. Ser. B.*, 1977, vol. 39.
10. Wu C.F.J. On the convergence properties of the EM algorithm. *The Annals of Statistics*, 1983, vol. 11, no. 1, pp. 95-103.
11. Xu L. and Jordan M.I. On the convergence properties of the EM algorithm for Gaussian mixtures. *Neural Computation*, 1996, vol. 8, pp. 129-151.
12. Naonori U., Ryohei N., Ghahramani Z., Hinton G.E. SMEM Algorithm for Mixture Models. *Neural Computation*, 2000, vol. 12, no. 9, pp. 2109-2128.
13. Vernaz Y., Langrognet F., Biernacki C. et al. Mixture Modelling Software MIXMOD User's Guide. *http://www-math.univ-fcomte.fr/MIXMOD/userguide/*, 1999