МЕТОДЫ ОБРАБОТКИ СИМВОЛЬНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ И СИГНАЛОВ
(Вычислительные системы)
1989 год

УДК 621.391:534.78

ЛИНГВИСТИЧЕСКИЙ ПОДХОД К ПРОБЛЕМЕ АВТОМАТИЧЕСКОЙ СЕГМЕНТАЦИИ РЕЧИ

Н.В. Зиновьева

§1. Проблема сегментации при восприятии и распознавании речи

Проблема сегментации речевого сигнала имеет многоаспектный характер, что подтверждается достаточно богатой историей ее изучения. Она является центральной при исследовании связи между континуальной речевой волной и ее дискретным символьным отражением в ментальном представлении человека. Для решения этой проблемы необходимо ответить на следующие вопросы: а) следует ли строить процедуру распознавания как модель вос - приятия речи человеком? б) является ли сегментация необходимым этапом в процессе восприятия речи человеком? в) каковы антропоморфные механизмы сегментации речи?

Ответ на первый вопрос может быть двояким. До сих пор большинство практических разработок в области автоматического распознавания речи шло путем наиболее простых технических решений вне ориентации на моделирование процессов восприятия.Это было оправдано тем более, чем менее мы были осведомлены о перцептивных механизмах в их глобальном стратегическом описании и частных особенностях функционирования в различных ситуациях речевого общения. Однако решение сложных задач распознавания (таких, как распознавание больших словарей слитной речи без под-

стройки под диктора), с которыми легко справляется человек, при подобных чисто технических подходах наталкивается на практически непреодолимые трудности (они хорошо известны). Это заставляет вновь обратиться к исследованию перцептивных механизмов, обеспечивающих быструю и эффективную процедуру восприятия речи человеком. Поэтому мы склоняемся к положительному ответу на первый вопрос и считаем, что даже в условиях отсутствия общей модели речевосприятия необходимо использовать в автоматическом распознавании те частные сведения о перцептивных процессах, которые уже накоплены в науке, пусть даже и в неточном или неполном виде.

Второй вопрос является к настоящему времени центральным для решения ряда проблем, в том числе проблемы базовых единиц общей стратегии речевосприятия.

Действительно, имеет ли место сегментация речевой волны или же дискретизация сигнала происходит в результате лингвистического абстрагирования и не затрагивает собственно физических характеристик речи? Если сегментация существует,то единицы какой размерности вычленяются из потока речи: фонемной,слоговой, слоговой или же некоторой несоотносимой с лингвистическими единицами, но характеризующейся высокой степенью внутренней однородности и устойчивости? И наконец, весь ли сигнал членится на единицы одинаковой размерности (скажем, фонемной) или возможны участки, которые обрабатываются иными способами?

Прямых свидетельств в пользу того или иного решения перечисленных вопросов как будто нет, что обусловливает наличие различных, порою прямо противоположных точек зрения. Но есть некоторые косвенные факторы, которыми мы и воспользуемся при обосновании нашей точки зрения.

Так, уже накоплено огромное количество экспериментальных данных, связанных с перцептивной релевантностью длительности сегментов фонемной размерности для их категориальной интерпре-

тации. В частности, относительная длительность гласных, несомненно, используется для определения ударности-безударности и соответственно ритмической структуры слова, длительность согласных и примыкающих к ним гласных используется при определении глухости-звонкости и места образования согласных и т.д. Использование же фактора длительности возможно тогда, когда дискретизация осуществляется не только на уровне лингвистической абстракции, но и соотносима с определенными участками речевой волны.

С другой стороны, как будто небезосновательным является утверждение о том, что "речевой поток ни акустически, ни артикуляторно не членится на отрезки, которые соответствовали бы фонемам" [1, с.145]. Это утверждение опирается на большое количество исследований, в результате которых так и не удалось разработать абсолютно надежную процедуру сегментации. Отсюда возникают попытки выделять в речевом сигнале те участки, которые поддаются достаточно надежному и устойчивому выделению, но могут не соотноситься с лингвистическими единицами. Эти попытки понятны, но при таком чисто физикалистском подходе невозможно использовать информацию о длительностях фонемных или соотносимых с ними сегментов.

Наиболее разумным в этой ситуации можно считать подход, в соответствии с которым речевой поток может члениться на отрезки фонемной размерности, но не всегда. В дальнейшем мы укажем те известные нам случаи, когда это невозможно, в остальном же будем исходить из того, что существуют способы лингвистически ориентированной дискретизации речевой волны, обнаружение которых может служить основой построения соответствующих алгоритмов сегментации.

Процитированное нами выше высказывание связано еще и с тем, что сегменты фонемной размерности не обеспечивают полной признаковой спецификации фонемы, так как многие признаки (в

частности, признаки места образования и - для русского языка - твердости-мягкости) частично реализуются на соседних участках фонем. Это означает только то, что в речевом сигнале мы не обнаружили своеобразных "акустических букв",так как фонемная интерпретация - процесс контекстно-зависимый и относительно полное признаковое описание возможно только на основании контекстной информации. Но даже сам факт использования контекстных данных, связанных с соседними сегментами, предполагает уже осуществление некоторых сегментирующих операций, так как для разумного использования контекстной информации надо отделить анализируемый сегмент от окружения.

Все это, а также ряд экспериментальных наблюдений над спектральными характеристиками речевого сигнала свидетельствуют о том, что сегменты разделяются в первую очередь по способу образования, в то время как спецификация места образования осуществляется с использованием контекста. Поэтому и сегментация должна быть ориентирована главным образом на выделение акустических признаков способа образования.

То, что членение речевого потока на сегменты в качестве базовых использует признаки способа образования,как будто подтверждается и экспериментально.

Так, в работе нидерландских исследователей Рингелинга и Ээфтинга [2] приводятся результаты эксперимента,в ходе которого аудиторам предъявлялись фразы, согласные сегменты которых заменялись двояким образом: а) все согласные подвергались заменам на звуки, отличающиеся местом образования (например, [п] на [к]),и б) все согласные подвергались заменам на звуки, отличающиеся по способу образования (например, [п] на [ф]).Тестовый материал подавался на аудирование в шумах. Результаты эксперимента показали, что при заменах по способу образования уровень словесной разборчивости (независимо от избыточности ис-

ходного речевого материала) составил 2-4%, в то время как при заменах по месту образования этот показатель варьировал от 52% (в условиях избыточности) до 8-10% в остальных случаях. Это одно из немногих известных нам доказательств того, что признаки способа и места образования играют разную роль при восприятии речи на слух, что очень существенно.

Все остальные исследования в этой области связаны с анализом речевых характеристик,объективированных с помощью различных спектроанализирующих приборов. Для ответа на третий из перечисленных нами в начале статьи вопросов обратимся к некото - рым из этих исследований.

Итак, какими могут быть механизмы сегментации речевой волны при восприятии речи человеком? Как мы уже отмечали, сегментация, по всей видимости, должна быть ориентирована на анализ признаков способа образования. Но сейчас уже очевидно,что вряд ли возможно обнаружить некоторую единую сегментирующую функцию, которая членила бы речевой поток на соответствующие сегменты. Для выделения звуковых сегментов, отличающихся способом образования, применяются различные процедуры, как-то: процедура выделения фрикативных, гласных, смычных, сонантов и т.д. Так, например, некоторые эксперименты,в частности так называемые опыты по "синхронизации" [3], свидетельствуют о том,что человек довольно надежно выделяет участки гласных, причем при этом происходит определение признака фонемного класса, в дан - ном случае - класса гласных,опознание которого определяет границы звука. То есть практически в данном случае можно гово - рить о сегментации через частичное распознавание звука по признаку способа образования.

К выводу об отсутствии единой сегментирующей процедуры пришли практически все исследователи (и мы в том числе), занимавшиеся так называемым "чтением сонаграмм" на материале раз-

личных языков. Наш опыт в этом отношении свидетельствует о том, что для эксперта минимальное затруднение вызывает сегментация, осуществляемая им с высокой степенью надежности (около 98% правильно выделенных сегментов для так называемой "лабораторной речи"), хотя никаких специальных сегментирующих процедур он не использует (в случае с сонаграммами это вряд ли и возможно). Эксперт же просто исходит из того, что способ образования характеризуется наличием - отсутствием и определенным порядком следования некоторых существенных акустических событий (для сонаграмм можно говорить о "видимых объектах"), границы которых и определяют границы сегментов.

Мы не будем здесь подробно останавливаться на описании процедуры сегментации речевых спектров, представленных на сонаграммах, так как она имеет опосредованное отношение к моделированию сегментации применительно к цифровому представлению речевого сигнала. Вместе с тем подчеркнем, что именно экспертные знания являются основой для создания лингвистически ориентированных алгоритмов сегментации. Это тем более примечательно, что практически все исследователи, работавшие с сонаграммами, в дальнейшем пришли к одной и той же идеологии при членении речевых сигналов автоматическим путем, независимо от того, с каким языком они работали. Обратимся к некоторым из этих работ.

Так, группа французских исследователей из Нанси в ряде работ [4,5] последовательно проводит идею сегментации через распознавание, подчеркивая, что первичная сегментация выполняется на основе сильного фонетического критерия. Этот критерий связан с поиском в речевой волне соответствий ядрам вокальных, фрикативных и взрывных участков. Сегментация является неопределенной , так как границы в дальнейшем могут быть модифицированы (неокончательность решений является очень важным моментом в вопросах сегментации).

Один из основоположников экспертного исследования сона - грамм Виктор Зу [6,7] предлагает членить речевой сигнал на единицы, соответствующие так называемым широким фонетическим классам, в которые он включает слабые фрикативные,сильные фрикативные, смычные, гласные, сонанты (плюс интервокальные носовые) и краткие звонкие смычки. Как видно из этого перечня, выделяемые классы соотносимы с признаками способа образования. Подчеркивается, что определение этих широких фонетических классов является мощным средством для выделения гипотез слов,слов-кандидатов, которые в словаре кодируются таким же способом,чтобы затем осуществить окончательное принятие решения о слове с помощью более тонкого анализа.

Р.де Мори и др. [8] подчеркивают, что сегментация непрерывной речи осуществляется одновременно с интерпретацией выделяемых сегментов и проходит через ряд этапов, в ходе которых постепенно улучшается ее качество. Основными элементами словаря сегментов являются гласные, взрывные, фрикативные, носовые и плавные.

Остановимся теперь на вопросе выбора наиболее удобного первичного описания речевого сигнала для осуществления процедуры сегментации. Оказывается, что количество параметров,необходимых для осуществления первичной сегментации,не очень велико. Так, В.Зу [7] в качестве рабочего набора параметров использует значение энергии в различных частотных полосах (не больше шести полос), количество переходов через ноль, а также суммарную огибающую интенсивности. К сожалению, мы не знаем границ частотных полос, которые используются в его алгоритмах.

Французские исследователи разработали, как мы уже отмечали выше, три алгоритма для выделения трех типов сегментов на базе спектрально-полосовых признаков: алгоритм NOVOCA обнаруживает вокализованные участки вместе с их границами путем анализа энергетических вариаций внутри частотной полосы

250-2350 Гц; алгоритм PLOSIV обнаруживает участки молчания или очень низкого уровня энергии в частотной полосе от 700 до 6000 Гц; в алгоритме FRIC анализируется частотный диапазон от 250 до 6000 Гц, при этом особое внимание уделяется полосе 400-6000 Гц (см. [5]). Р.де Мори [8] осуществляет первичную сегментацию на базе анализа огибающей интенсивности и спектральной информации в следующих частотных полосах: 250-650 Гц, 650-1300 Гц, 1300-2200 Гц, 2200-3100 Гц, 3100-4300 Гц.

Мы так подробно остановились на проблеме выделения полос потому, что именно адекватный их выбор определяет степень диктороне зависимости алгоритма. Мы полагаем (об этом более подробно сказано ниже), что абсолютной дикторонезависимости при таком небольшом количестве параметров достичь невозможно, однако можно разработать дополнительную процедуру адаптации полос к голосу диктора.

Выскажем несколько общих соображений относительно возможных границ полос. К.Стивенс [9] отмечает, что в слуховой системе звук может обрабатываться по-разному в разных диапазонах частот. Границы этих диапазонов до сих пор хорошо не определены, может быть, существует даже небольшое их перекрытие. Самый низкочастотный диапазон ограничен сверху частотой примерно 800 Гц и обычно включает в себя область варьирования форманты F1 для взрослых (вероятно, мужчин. - Н.З.). Следующий, среднечастотный диапазон включает область варьирования формант F2 и F3 и ограничен сверху частотой 3000-3500 Гц. И наконец, высокочастотному диапазону соответствует область выше 3500 Гц. Эти диапазоны различаются частотной разрешимостью внутри диапазона, определяемой шириной критических полос (около 100 Гц в низкочастотном диапазоне и все возрастающей с увеличением частоты), и связанной с ней временной разрешимостью. Очевидно, что если мы хотим ориентировать распознавание на моделирование вос-

приятия речи, эти диапазоны должны учитываться при выборе полос анализа.

С другой стороны, было бы опрометчивым ориентироваться только на характеристики слуховой системы. Необходимо учитывать и акустические особенности звуков, обусловленные процессом речепорождения и связанные, в частности, с качеством голоса диктора. Нам кажется разумным такой подход, при котором адаптация к голосу диктора осуществляется не столько алгоритмическим способом, сколько путем разбиения частотного диапазона на полосы таким образом, чтобы спектрально-энергетический баланс наилучшим образом отражал лингвистические реалии в произнесении данного диктора.

§2. Основные принципы построения алгоритма на базе спектрально-энергетического баланса

Нами разработаны два алгоритма сегментации речевого сигнала на базе спектрально-энергетического баланса, которые отличаются друг от друга главным образом исходным параметрическим представлением.

В предыдущем разделе мы попытались показать, что разбиение спектрального диапазона на полосы может быть различным, но оно должно в той или иной мере удовлетворять двум требованиям: а) соответствовать основным диапазонам обработки речевого сигнала в слуховой системе человека; б) отражать лингвистически значимое распределение энергии в голосе конкретного диктора. По-видимому, первое требование является общим для всех видов первичного представления, второе - предполагает их частичную адаптацию. Прежде чем приступить к описанию алгоритмов, необходимо сказать, что полученное нами первичное представление в обоих случаях удовлетворяло второму требованию, так как каждое из них отражало характеристики одного (в каждом случае - своего) диктора. Вопросы адаптации алгоритмов к другим голосам на-

ми пока детально не рассматривались, однако основные принципы построения процедур сегментации могут быть продемонстрированы и на ограниченном дикторском материале.

Разработка алгоритмов осуществлялась экспертным методом, что означает следующее: эксперты по чтению сонаграмм получали цифрограммы речевых сигналов в соответствующем первичном представлении, адаптировали свои знания о спектральных характеристиках речи к данному представлению, отрабатывали приемы сегментации, которые затем формулировали в виде правил. В созданных по этой схеме алгоритмах работают в той или иной мере такие общие принципы анализа сонаграмм, как: а) сегментация речевого сигнала через частичное распознавание по признакам способа образования; б) опора на некоторое специальное описание речевого сигнала в виде так называемых "акустических ключей"; в) широкое использование контекстной информации; г) активный характер анализа (более подробно об этих принципах см. [10]).

К указанным принципам можно добавить еще один, который был нами осознан и сформулирован именно в ходе разработки алгоритмов сегментации, но, по-видимому, может быть отнесен ко всему процессу анализа и восприятия речи - это принцип различения и соответственно использования разных способов обработки центра и периферии любых лингвистических объектов: сегментов, слогов, слов, синтагм и т.д. Центр, как правило, характеризуется наибольшей (максимальной) выделенностью характерных для данного объекта признаков и наибольшей информативностью. Поэтому определение центра в достаточной мере контекстонезависимо, в то же время именно центр обеспечивает контекст для опознания периферии, которая гораздо более аморфна и поэтому может по-разному оцениваться в различных условиях.

Различение центра и периферии позволяет смоделировать такую удивительную способность человека, как умение по-разному оценивать абсолютно одинаковые участки речевой волны и одина -

ково оценивать очень различающиеся фрагменты. Этот эффект достигается благодаря тому, что периферия оценивается не вообще, а относительно ближайшего центра, информация о котором определяет правила интерпретации периферийных участков.Как этот принцип работает на уровне сегментов, будет видно из описания алгоритмов, здесь же отметим следующее: на уровне слов центральные участки предполагают наличие наиболее надежно вычленимых сегментов (так называемые "острова надежности"), в то время как периферийные (например, заударные суффиксально-флексийные комплексы типа "-ия", "-ние" и др.) могут не члениться вовсе и оцениваться принципиально по-другому. Поэтому, как уже отмечалось выше, не следует ожидать абсолютной сегментируемости речевой волны на участки, соответствующие в обыденном языковом сознании буквам.

В связи с различной надежностью выделения разных сегментов строится и общий сценарий алгоритмов: на первом этапе определяются центры и границы наиболее контрастных участков: вокальных, фрикативных, смычных,- а затем внутри вокальных осуществляется поиск сонорных. Наиболее проблематичным является разделение плавных сонорных и слабых (безударных) гласных, так что, может быть, на этом уровне разумнее такие участки также считать неделимыми, как и упоминавшиеся выше заударные суффиксально-флексийные комплексы.

Прежде чем приступить к описанию алгоритмов, следует отметить еще одно обстоятельство, обусловившее некоторое различие в формулировке правил первого и второго вариантов: в первом случае исходное параметрическое представление, с которым работали эксперты, было получено с использованием специальных процедур нормирования энергетического диапазона, что позволило нам вводить в правила некоторые пороговые значения; во втором случае нормировка не осуществлялась, что исключало использование энергетических порогов , но зато в этом варианты мы имели

дополнительную информацию о количестве переходов через ноль (ρ). Все эти особенности отражены в правилах, к описанию которых мы приступаем.

АЛГОРИТМ 1

Сегментация по уровням энергии в четырех полосах частот (A1-A4)

Используемые признаки:	Функциональная нагрузка:
Энергия в полосе A1 (от 0 до 300 Гц)	- определяет наличие - отсутствие форманты F0 и частично F1;
Энергия в полосе A2 (от 200 до 900 Гц)	- определяет наличие - отсутствие форманты F1 (частично F0 и F2); а также низкочастотного шума ("х-шума");
Энергия в полосе A3 (от 500 Гц до 3,5 кГц)	- определяет наличие - отсутствие формант F2, F3, F4,а также среднечастотного шума;
Энергия в полосе A4 (от 3,5 до 7 кГц)	- определяет наличие - отсутствие высокочастотных шумов ("с-шум" и "ш-шум"), а также высокочас - тотных формант гласных.

ПРАВИЛА СЕГМЕНТАЦИИ

1.1. Определение центров "с-шума". Отыскиваются отсчеты, соответствующие локальным максимумам значений энергии в четвертой полосе, т.е. отсчеты, удовлетворяющие условиям: $(A4)_t > (A4)_{t-\Delta t}$ и $(A4)_t > (A4)_{t+\Delta t}$, где t - момент времени, характеризующий положение локального максимума, Δt - шаг дискретизации (равный 12 мсек в нашем конкретном случае). Для каждого выделенного отсчета проверяются следующие условия: $A4 > 10$, $A4 > A2 + A3$ и $A3 < 8$ (индекс t здесь и далее опускается). Если эти условия на данном отсчете удовлетворяются, то он маркируется как центр "с-шума".

1.2. Определение границ "с-шума". Маркер "с-шум" распространяется направо и налево от каждого выделенного центра на все отсчеты, удовлетворяющие условиям: $A4 > A2 + A3$ и $A2 \leq 4$. Эта процедура определяет все зоны наличия "с-шума".

2.1. Определение центров "ш-шума". Отыскиваются локальные максимумы в полосе A3, проверяются следующие условия: $A3 > 8$ и не далее чем на 1-2 отсчета от каждого максимума существует максимум в полосе A4 такой, что $A4 \geq 10$; $A4 > A2$.

2.2. Определение границ "ш-шума". Шум длится до тех пор, пока $A3 = 0$ и $A4 \geq A2$ и $A2 \leq 4$.

3.1. Определение центров сильного вокального. Отыскиваются моменты времени, которым соответствуют локальные максимумы суммы (A2+A3). Проверяются следующие условия: $A2 > A4$; $A1 \neq 0$ и $A3 \neq 0$; не далее чем на один отсчет от каждого выделенного максимума (A2+A3) расположен локальный максимум, соответствующий всем четырем полосам, т.е. (A1+A2+A3+A4), который должен быть не меньше 40.

3.2. Определение границ сильного вокального. Сильный вокальный длится до тех пор, пока $A2 \geq A4$ или $A2 + A3 \geq A4$ и A1, A2, A3 не равны нулю.

4.1. Определение центров слабого вокального. Отыскиваются так же, как центры сильных гласных, но при снижении порога для суммарной энергии до 20. Проверяются условия $A2 + A3 > A4$ и A1, A2, A3, A4 не равны нулю и $A2 \geq 4$.

4.2. Определение границ слабого вокального. Движемся направо и налево от центра до тех пор, пока $A2 + A3 \geq A4$ и A1, A2, A3 не равны нулю.

5.1. Определение центров пауз. Выделяем все отсчеты, в которых A2, A3 и A4 равны нулю.

5.2. Определение границ паузы. Справа она находится там, где кончается центр; граница паузы слева - там, где нарушается условие $A2 + A3 + A4 < 5$ и $A2 < 3$ и $A3 < 3$.

5.3. Определение слабых пауз. Пауза может не иметь цент - ра. Тогда она называется слабой и ищется только среди отсчетов, не обозначенных другими маркерами. Слабая пауза фиксируется при выполнении условия левой границы зоны паузы, т.е. когда $A2+A3+A4 < 5$ и $A2 < 3$ и $A3 < 3$.

6.1. Определение взрывов. Выявление паузы служит сигналом для поиска справа от нее слабых шумов (взрывов). Так,если справа от паузы есть несколько отсчетов, удовлетворяющих условию $A4 \geq A2+A3$ или $A4 \geq A2$, то они идентифицируются как взрыв.

6.2. Взрыв может идентифицироваться и на участке, который ранее был помечен как вокальный (в соответствии с правилами 3.2, 4.2). Если в этой конфликтной ситуации $A1 \geq A4$, то сохраняем маркировку соответствующих отсчетов как вокальных; если $A4 > A1$, то меняем маркировку и идентифицируем соответствующие отсчеты как взрывные.

6.3. Определение слабого шума. Среди неопознанных сегментов отыскиваем те, которые удовлетворяют критерию взрыва,но не стоят после паузы, и маркируем их как "слабый шум".

6.4. Взрыв может идентифицироваться и на участке, который ранее был опознан как шум. В этом случае сохраняем маркировку соответствующего шума. Однако если взрыв определяется на уча - стке между паузой и шумом и получается так, что сегменты, маркированные как взрывные, переходят в сегменты, маркированные как шумовые, то участок взрыва переименовывается в шум.

7.1. Определение центров "х-шума". Отыскиваются моменты времени, которым соответствуют локальные максимумы суммы $A2+A3$, проверяются условия: $A3 > A4$ и $A3 > A2$ и $A1 = 0$.

7.2. Определение границ "х-шума". Каждая зона "х-шума" охватывает все отсчеты справа и слева от центра, удовлетворяющие условиям A3 > A4 и A3 > A2*).

ПОИСК СОНАНТОВ

8. Обнаружение "р-просветов".

8.1. Если два вокальных сегмента разделяются одним или двумя отсчетами, которые либо никак не обозначены, либо обозначены как пауза или слабая пауза, то это может быть "р-просвет". Проверяем этот отсчет или оба отсчета по критерию совпадения минимумов во всех четырех полосах частот. Если на интересующих нас отсчетах совпадают локальные минимумы энергии A1, A2, A3 и A4, то этот сегмент маркируется как "р-просвет".

8.2. В любом вокальном сегменте отыскиваются локальные минимумы A1, A2, A3 и A4, и если все эти минимумы приходятся на один отсчет, то этот отсчет из вокального переименовывается в "р-просвет".

9.1. Обнаружение центров носовых сонантов. Внутри вокальных сегментов, длительность которых больше 3-х отсчетов, отыскиваются отсчеты, удовлетворяющие условиям: A4 = 0, A1 > A2 и A1 > A3.

9.2. Границы носовых сонантов определяются по невыполнению условий: A4 < A3 и A1 > A2, A1 > A3.

10.1. Обнаружение центров сонантов. Внутри вокальных участков, длительность которых больше 3-х отсчетов, определяются отсчеты, удовлетворяющие условиям: A4 = 0, A1+A2+A3 < 20.

10.2. Определение границ сонантов. Маркер "сонант" распространяется на все отсчеты, расположенные слева и справа

---

*) Правила выделения "х-шума" отнесены нами в конец не потому, что этот шум плохо вычленяется из потока речи, а просто потому, что у нас было недостаточно материала для проверки надежности работы правил.

от центра и удовлетворяющие условиям: A4 < 4 и A1+A2+A3 < < 20.

ОГРАНИЧЕНИЯ ПРИ ОПРЕДЕЛЕНИИ СОНАНТОВ

1. Распространение сонантов за пределы вокальных участков допускается только в случае, если окружающие участки пока никак не маркированы, либо маркированы как слабые паузы. Распространение на другие сегменты не допускается.

2. Запрещается распространение сонантов на центры гласных.

Поиск сонантов не ограничивается теми правилами, которые изложены выше. Сонанты иногда так вокализуются, что выделить их по правилам не представляется возможным. В этом случае начинает работать поиск по расстояниям и их производным (или первым разностям), отражающим скорость изменения параметров. Мы не будем здесь подробно останавливаться на описании этих процедур, так как они лучше знакомы исследователям, занимавшимся вопросами сегментации. Отметим лишь следующее обстоятельство. Поиск по расстояниям и производным осуществляется только внутри вокальных участков. Отдельно разработаны процедуры поиска перед центром, между центрами (так как очень часто вокальные участки, содержащие сонанты и гласные, характеризуются наличием нескольких центров, что уже само по себе является сигналом к поиску сонантов) и после центра. Такая организация выделения сонантов позволяет более осмысленно учитывать различные коартикуляционные влияния гласных на сонанты и наоборот.

Для завершения краткого описания первого алгоритма следует сказать, что в него включены правила объединения некоторых субфонетических сегментов (таких, как паузы и взрывы или паузы и шумы, р-просветы и участки краткой вокальности) в сегменты фонемной размерности с определением признаков способа их образования, а также некоторый набор правил определения глухо-

сти - звонкости шумных согласных. Мы не будем их детализировать. Для примера приведем лишь описание процедуры поиска сонантов перед центром вокального участка.

Для всех вокальных отрезков, длительность которых превышает 4 отсчета, вычисляются расстояния между соседними отсчетами по формуле:

$$S_j = \sum_{i=1}^{4} (A_i^j - A_i^{j-1}),$$

где i - номер полосы, j - номер отсчета, а также первые разности $S'_j = S_j - S_{j-1}$.

Далее:

1. Определяется максимум S_j. Если он расположен перед центром и не на пограничном отсчете, то анализируется S'_j.

2. Если на этом же отсчете зафиксирован максимум S'_j и он больше 10, то за этим отсчетом проходит граница между сонантом и гласным.

3. Если на этом отсчете зафиксирован максимум S'_j и он меньше 10, то определяется удельный прирост S_j/S_{j-1}; если он больше 0,2, то после этого отсчета проходит граница между сонантом и гласным. Если удельный прирост меньше 0,2, то граница отсутствует и сонант в этом вокальном не выделяется.

4. Для участка, выделенного в качестве потенциального сонанта, проверяется выполнение условия $A4 \leq 4$. Если это условие не удовлетворяется, то прежнее решение отменяется и сонант в данном сегменте не выделяется.

АЛГОРИТМ П

Сегментация по энергии в четырех полосах (A1-A4), суммарной энергии (AC) и числу переходов через ноль (ρ)

Используемые признаки:	Функциональная нагрузка:
Энергия в полосе A1 (от 200 до 800 Гц)	- определяет наличие - отсутствие форманты F1 (частично также F0 и F2);
Энергия в полосе A2 (от 800 до 1200 Гц)	- определяет наличие - отсутствие форманты F1 для непередних гласных и низкочастотного шума ("х-шума");
Энергия в полосе A3 (от 1200 до 2400 Гц)	- определяет наличие - отсутствие форманты F2 для незадних гласных и (частично) среднечастотного шума ("ш-шума");
Энергия в полосе A4 (от 2400 Гц до 5000 Гц)	- определяет наличие - отсутствие формант F3, F4 для гласных,среднечастотного шума ("ш-шума") и частично - высокочастотного шума ("с-шума");
AC - суммарная энергия в полосе от 200 до 5000 Гц	- определяет наличие - отсутствие и уровень энергии в указанной полосе частот;
ρ - количество переходов через ноль	- определяет наличие - отсутствие шумовых или вокальных участков речевой волны.

ПРАВИЛА СЕГМЕНТАЦИИ

1.1. Определение центров вокальных участков. Отыскиваются локальные максимумы значений AC при наличии не далее чем на один отсчет локального максимума A1+A3. Для каждого максимума проверяются следующие условия: $0 < \rho < 10$, $A1 > A2$, $A1 > A3$, $A1 > A4$, $A3 > 0$, $A2+A4 > 0$, $A1 > A2+A4$, $A3 > A4$. Если эти условия на выделенном отсчете выполняются, то он маркируется как центр вокального участка.

1.2. Определение границы вокального участка. Маркер "вокальный участок" распространяется на все отсчеты справа и слева от центра до тех пор, пока на них удовлетворяются все вышеприведенные условия, кроме двух: A3 может равняться 0 и A3 не обязательно строго больше A4, т.е. достаточно выполнения не - строгого неравенства $A3 \geq A4$.

1.3. Дополнительные правила. Они работают внутри вокальных участков и позволяют частично распределять эти участки внутри признакового континуума: "передний - задний" гласный. Эти правила применяются не к сегменту в целом, а к каждому отсчету отдельно, так как некоторые гласные в силу влияния кон - текста могут содержать участки более переднего характера наряду с непередними и наоборот (как, например, в словах "нюх", "быль" и др.). Приведем эти правила:

1.3.1. Если $A3 = 0$ и $A1 > A2 \geq A3 \geq A4$, то это задний гласный. Правило может быть переформулировано в более мягкой форме: если $A1 > A2+A3+A4$ и $A1 > A2 \geq A3 \geq A4$, то это задний гласный.

1.3.2. Если $A2 = 0$ или $A4 > A2$, то это передний гласный.

1.3.3. Если $A1 > A3$ и $A2 > A4$, а A2 и A3 близки по значению (т.е. разница между ними не превышает 5), то это сред - ний гласный.

2.1. Определение центров высокочастотного шума. Отыскиваются отсчеты с локально максимальными значениями ρ при ус - ловии, что $\rho > 10$, и не далее чем на один отсчет от каждого выделенного отмечается локальный максимум энергии (A3+A4). Проверяется также условие: $A4 > A2$.

2.2. Определение границ высокочастотного шума. Шум длится до тех пор, пока $\rho \geq 10$, и $A4 \geq A2$.

2.3. Дополнительные правила. Внутри шумовых участков работают свои дополнительные правила, которые определяют относи-

тельную высоту шума, т.е. различают собственно высокочастотный "с-шум", среднечастотный "ш-шум" и слабый среднечастотный "й-шум". Перечислим эти правила:

2.3.1. Если A4 > A3 и $\rho > 30$ и $\rho > A3+A4$, то это "с-шум".

2.3.2. Если $\rho < 30$ и ρ сопоставимо с A3+A4, то это "ш-шум".

2.3.3. Если $\rho < 30$ и ρ значительно превышает A3+A4, то этой "й-шум".

2.4. Сегментация переходных участков. Между вокальным участком и участком "й-шума" возможно появление сегмента, на ко - тором не выполняются до конца ни условия вокальных, ни условия шумовых сегментов (например, $\rho < 10$, что не позволяет при - писать этому сегменту значения шума, и A4 > A3, что не позволяет считать его вокальным). Решение о том, куда отнести такой. участок, требует дополнительных исследований. Возможно,его следует оставить в качестве переходного, не относя ни к гласному, ни к согласному.

3.1. Определение центров низкочастотного шума ("х-шума"). Отыскиваются локальные максимумы значений АС при условии, что не далее чем на один отсчет от каждого имеется локальный мак - симум A2+A3. Проверяются следующие условия: A2 > A1, A1+A3 > 0, A3 > A4. Если они удовлетворяются,то выделенному отсчету приписывается маркер "центр х-шума".

3.2. Определение границ "х-шума". Низкочастотный шум длится до тех пор,пока удовлетворяются условия выделения центра,но в менее жестком варианте, т.е. при замене строгих неравенств нестрогими.

4.1. Определение центров пауз. Маркер "центра паузы" приписывается всем отсчетам, на которых выполняется условие A2+A3+A4 = 0 при наличии локального минимума по ρ .

4.2. Определение границ паузы. Правая граница паузы совпадает с границей центра. Это означает,что как только перестает удовлетворяться условие A2+A3+A4 = 0, считаем,что пауза закончилась. Левая граница определяется по более мягкому критерию, т.е. при движении влево от центра относим к паузе все отсчеты, удовлетворяющие условию A2+A3+A4 < 10 и $\rho < 10$.

4.3. Неопознанные участки. Они могут возникать справа от паузы,должны, скорее всего,соответствовать взрыву смычного согласного и оцениваться по критериям,определенным для периферийных участков шумов, т.е. для границ шумов (правило 2.2, 3.2).

5. Носовые сонанты определяются без центра.Для этого все нешумовые участки (т.е. неопознанные и вокальные) оцениваются по следующим критериям: AC-(A1+A3) < 3 и $\rho < 3$. Все отсчеты, удовлетворяющие этим условиям,маркируются как носовые сонанты.

Так же, как и в первом алгоритме, здесь используются процедуры поиска сонантов по скорости изменения параметров внутри вокалических участков, а также правила определения глухости - звонкости шумных согласных и объединения субфонетических сегментов в сегменты фонемной размерности.

Для примера приведем часть этих правил.

Правило 1. Если длительность $t(П) < t(Ш)$,то данный сегмент относится к группе глухих смычных.

Правило 2. Если $t(П) \sim t(Ш)$,то данный сегмент может соответствовать глухим смычным аффрицированным (чаще всего мягким переднеязычным или твердым и мягким в позиции абсолютного конца слова-фразы) или стечениям глухих смычных с глухими фрикативными (разного места образования).

Правило 3. Если $t(П) < t(Ш)$,то данный сегмент относится к группе аффрикат.

Последовательности типа пауза-шум и шум-звонкий также оцениваются с помощью временных акустических ключей второго уровня, в результате чего выделяются сегменты, соответствующие вибрантам. Для этого используется

Правило 4. Если длительность t(П) (или t(П зв.)) - минимальная и не превышает 25-30 мс, то данный сегмент соответствует вибрантам; чаще всего этот сегмент объединяется с прилегающими к нему слева и справа вокализованными участками, если длительность последних не превышает 40-50 мс. В противном случае считаем, что вибранту соответствует только участок минимальной по длительности паузы.

Сравнение двух приведенных выше алгоритмов показывает, что они различаются, главным образом, постольку, поскольку в их основу положены различные параметрические представления, полученные в результате первичной обработки речевого сигнала. Вместе с тем общая идеология их создания свидетельствует о перспективности применения экспертных знаний к информации о спектральном балансе, благодаря чему достигается лингвистически осмысленная сегментация речевой волны. Следует еще раз подчеркнуть, что эта сегментация ни в коем случае не является окончательной, но она дает основу для дальнейшего использования экспертных знаний и последовательного уточнения границ сегментов. Как свидетельствует опыт зарубежных исследователей [4-8], но, к сожалению, не наш собственный, процедуры сегментации, основанные на экспертных правилах анализа спектрального баланса, являются достаточно устойчивыми и дикторонезависимыми, что позволяет использовать их при разработке дикторонезависимых систем распознавания.

К выделенным в результате лингвистически ориентированных процедур сегментации участкам речевой волны можно применять процедуры эталонизации и сравнения, что уже сейчас позволило бы приспособить существующие алгоритмы для распознавания слитной речи. Таким образом, лингвистический подход к проблеме сегментации открывает путь к решению таких сложных задач, как распознавание слитной речи и создание дикторонезависи -

мых систем распознавания речи (или систем с минимальной подстройкой к диктору).

Литература

1. БОНДАРКО Л.В., ЗИНДЕР Л.Р. Исследования фонетики //Основы теории речевой деятельности. - М., 1974.

2. RINGELING J.C.T., EEFTING W. A case for global listening strategies //Proceedings XIth ICPhS.- 1987. -Vol. II.

3. Физиология речи. Восприятие речи человеком.-Л., 1976.

4. CARBONELL N., FOHR D., HATON J.P. et al. An expert system for the automatic reading of french spectrograms //Proceedings ICASSP-84.- 1984. - Vol. III.

5. CARBONELL N., DAMESTOY J.-P., FOHR D. et al. APHODEX, Design and implementation of an acoustic-phonetic decoding expert system //Proceedings ICASSP-86. - 1986. -Vol. II.

6. ЦЗУЭ (Зу) В.В. Лингвистический подход к автоматическому распознаванию речевых сигналов //ТИИЭР. "Речевая связь с машинами". - 1985. -Т. 73, № 11.

7. CHEN F.R., ZUE V.W. Application of allophonic and lexical constraints in continuous digit recognition //Proceedings ICASSP-84.- 1984. -Vol. III.

8. De MORI R., GILLOUX M., MERCIER G. et al. Integration of acoustic, phonetic, prosodic and lexical knowledge in an expert system for speech understanding //Proceedings ICASSP-84. - 1984. - Vol. III.

9. STEVENS K.N. Relational properties as perceptual correlates of phonetic features //Proceedings XIth ICPhS. - 1987. Vol. IV.

10. ЗИНОВЬЕВА Н.В. Механизмы извлечения лингвистической информации из спектрального представления речевого сигнала: Автореф. дис ... канд. филолог. наук: 10.02.21. - 16 с. - М., 1986.

Поступила в ред.-изд.отд.
20 сентября 1989 года