УДК 004.934

О ПРОБЛЕМЕ АВТОМАТИЧЕСКОЙ СЕГМЕНТАЦИИ РЕЧЕВОГО СИГНАЛА НА ФОНЕТИЧЕСКИЕ ЭЛЕМЕНТЫ

Алёшина Н.Д., аспирант**,** E-mail: natalia91@bk.ru
Фёдоров В.Б., к.т.н., доцент, E-mail: fdorov@mail.ru
МГТУ МИРЭА, Москва, Россия

Аннотация. Дается краткий обзор известных методов решения задачи автоматической сегментации речевого сигнала на фонетические элементы, обсуждаются их достоинства и недостатки и предлагается новый подход к решению данной проблемы, основанный на использовании нелинейных процедур построения оптимальных базисов для данного речевого сигнала, наилучшим образом учитывающие локальные временные особенности этого сигнала.
Ключевые слова: распознавание речи, речевой сигнал, фонемы, сегментация, оконное переобразование Фурье, вейвлеты, локальные косинусные базисы

ON THE PROBLEM OF THE AUTOMATIC PHONETIC SEGMENTATION OF THE SPOKEN LANGUAGE RECORDINGS

Aleshina N.B.: graduate student, E-mail: natalia91@bk.ru
Feodorov V. B.: Ph.D., Associate Professor, E-mail: fdorov@mail.ru
MSTU MIREA, Moscow, Russia

Abstract. In this paper a short review of the state of the art methods in the field of the automatic phonetic segmentation of the spoken language recordings is given, a number of advantages and disadvantages of the methods are discussed; the authors suggest a new approach to the solution of the problem considering nonlinear construction of optimal bases for the given utterance providing the best emphasis of the local temporal features.
Keywords: speech recognition, speech signal, phonemes, automatic segmentation, window transform Fourier, wavelets, local cosine bases

Введение

Проблема автоматического определения границ между фонетическими элементами человеческой речи является весьма актуальной. Она важнадля верификации, идентификации диктора, автоматического распознавания речи. Так, если бы удалось добиться высокого качества автоматической сегментации, то задача построения системы компьютерного распознавания речи была бы значительно упрощена [1]. Однако, несмотря на многочисленные попытки справиться с задачей автоматической сегментации, и несмотря на определенные успехи в этом направлении [1-5], существующие системы все еще далеки от совершенства. Необходимо отметить, что эта задача объективно является весьма трудной, даже ручная разметка речи на

фонетические составляющие для достижения хорошего качества требует от оператора специальной лингвистической подготовки и немалого опыта.

В данной работе, после обсуждения некоторых из ранее предложенных алгоритмов сегментации, рассматривается возможность реализации еще одного перспективного, по мнению авторов, подхода к решению этой проблемы.

Методы сегментации

В настоящее время известны различные подходы к проблеме автоматической сегментации речи [1-6]. В работах [1,2] представлено обширное исследование и предложены алгоритмы частично решающие данную проблему.

В основном все известные [1-5] подходы к сегментации основываются на выделении в речи элементов каких-либо широкие фонетических классов (ШФК), при этом разные авторы определяют различные ШФК. В ряде работ [1-3] различительные признаки ШФК формируются на основе кратковременного Фурье-анализа. Другие исследования основываются на использовании вейвлетов, например, [5,6]. Имеются также исследования, в которых границы фонем детектируются на основе отслеживания изменений полной вариации последовательных сегментов сигнала [4,6].

Остановимся более подробно на обсуждении метода, предложенного в работе [3], как достаточно «прозрачного» с физической точки зрения. Предложенный в[3] алгоритм реализует сегментацию речевого сигнала на следующие ШФК: вокализованные звуки (гласные, полугласные и звонкие согласные), невокализованные (глухие согласные), сонорные, взрывные. Шум в паузах между участками речи рассматривается как еще один класс звуков и, таким образом, данный алгоритм, в какой-то мере, может выполнять функции детектора голосовой активности. При этом, поскольку сонорные и взрывные звуки являются непересекающимися подклассами класса невокализованных звуков, то некоторые из рассматриваемых ШФК находятся в иерархическом отношении. Соответственно, рассматриваемый алгоритм сегментации также имеет иерархический характер: сначала определяются границы звуков, относящихся к более широкому фонетическому классу (ФК), а уже затем, внутри найденных границ, определяются границы элементов более узких ФК.

Для этого сначала поток отсчетов исходного сигнала покрывается скользящими весовыми окнами Хеннинга длительностью 6 мс и с шагом смещения, равным 1 мс. Затем, при условии, что частота дискретизации сигнала не ниже 16 кГц, длину каждого сегмента сигнала, взвешенного таким окном, увеличивают до 512 отсчетов путем дополнительного присоединения в его конец необходимого числа нулей. Это делается с

целью обеспечения гарантированного попадания в рассматриваемые далеечастотные диапазоны хотя бы по 1 спектральному отсчету (после перехода в частотную область).

Затем выполняется переход в частотную область путем ДПФ каждого взвешенного сегмента, и в получаемом частотном сегменте оставляется только первых 256 отсчетов (при вещественном исходном векторе, обе половины получаемого частотного спектра находятся в отношении комплексного сопряжения, и поэтому всю полезную информацию несет любая из них). При этом каждый частотный отсчет сегмента заменяется квадратом своего модуля.

После чего выполняется усреднение полученных частотных сегментов на последовательных интервалах времени длительностью 20 мс со сдвигом в 1 мс, т.е. вычисляется среднее по 1…20-ому сегментам, затем – по 2…21-ому сегментам, и т.д. до конца. Такое усреднение спектрограммы дает сглаживание пульсацией голосовых связок и флуктуаций шума, при этом длительность 20 мс обычно считается равной периоду квазистационарности речевого сигнала.

Полученную таким образом последовательность частотных сегментов будем называть усредненной спектрограммой, а её элементы – усредненными частотными сегментами или просто – частотными сегментами. При этом временная ось усреднённой спектрограммы оказывается смещенной вправо на 10 мс по отношению к исходной не усредненной спектрограмме.

В каждом частотном сегменте спектрограммы 1-й отсчет соответствует 0-частоте, а 256-й отсчет – частоте равной 8 кГц (при частоте дискретизации 16 кГц). Далее весь рассматриваемый частотный диапазон от 0 до 8 кГц подразделяется на 6 поддиапазонов (полос), представленных в таблице 1. Такое подразделение может быть оправдано тем, что звуки различных рассматриваемых ШФК имеют существенно отличающиеся свойства в выделенных частотных полосах.

Частотные диапазоны Таблица 1.

№ диапазона	Диапазоны частот
1	0-0.4 кГц
2	0.8-1.5 кГц
3	1.2-2.0 кГц
4	2.0-3.5 кГц
5	3.5-5.0 кГц
6	5.0-8.0 кГц

Первый диапазон 0…0.4 кГц позволяет определять границы всех вокализованных звуков, причем верхняя граничная частота этого диапазона выбрана так, чтобы отсекать случайные выбросы низкочастотной составляющей усредненной

спектрограммы. Границы сонорных звуков определяются с использованием сразу 2-5-го частотных диапазонов. При этом 2-ой и 3-ий диапазоны имеют небольшое пересечение, что оправдано особой важностью этих диапазонов, а именно тем, что на границах двух различных вокализованных звуков сильные изменения частотного спектра часто происходят именно на интервале частот от 0.8 кГц до 2 кГц. Последний 6-ой диапазон 5.0…8.0 кГц, охватывающий всю область высоких частот, используется для определения границ пауз между участками речи.

Фактически использование только 6-ти выделенных частотных поддиапазонов приводит к «сжатию» усредненной спектрограммы по оси частот путем интегрирования (суммирования) ее частотных сегментов по каждому из поддиапазонов, таким образом, что вместо 256 частотных отсчетов получится только 6. Далее будем рассматривать именно сжатую усредненную спектрограмму.

Физический смысл частотных компонентов сжатой усредненной спектрограммы – это парциальные энергии данного сегмента, соответствующие выделенным частотным диапазонам. Можно было бы предположить, что границам звуков соответствуют локальные минимумы некоторых из этих парциальных энергий. Однако авторы [3] вместо собственно парциальных энергий предлагают использовать их 1-ые разности и искать моменты, в которые эти разности достигают экстремальных значений. Т.е. предполагается, что границам между звуками соответствуют моменты наиболее быстрых изменений некоторых парциальных энергий. Причем предполагается, что, например, началу вокализованного участка соответствует положительныйлокальный экстремум разности энергий 1-ой частотной полосы, а окончанию – отрицательный локальный экстремум. И подобное предположение касается не только вокализованных участков, но и всех других определяемых ШФК.

Первые разности вычисляются на интервалах дискретизации в 50 мс, смещая этот интервал для получения каждой очередной разности на 1 мс. Интервал длительностью именно 50 мс выбирается исходя из требования полного захвата всех частей любой фонемы, без исключения её начала или конца. Для исключения слишком большого числа обнаружения ложных локальных экстремумов, не отвечающих действительным границам звуков, используется порог: абсолютная величина локального экстремума рассматриваемых разностей должна превышать 9 дБ, в противном случае он игнорируется.

Всё описанное выше можно назвать «грубым» режимом определения положения границ звуков. В алгоритме используется также «тонкий» режим, представляющий

собой туже процедуру, но с другими параметрами. Сравнение численных значений параметров для обоих режимов представлено в таблице 2.

Таблица 2

Сравнение численных значений параметров алгоритма сегментации для 2-х режимов его работы

Параметры алгоритма	«Грубый» режим	«Тонкий» режим
Ширина окна усреднения	20 мс	10 мс
Шаг для вычисления 1-ой разности	50 мс	26 мс
Порог для экстремумов 1-ой разности 1-го частотного диапазона	9 дБ	6 дБ
Порог для экстремумов 1-х разностей 2-6-го частотных диапазонов	9 дБ	9 дБ

Положение границ каждого вокализованного участка речи определяются с помощью «грубого» режима поиска положения пары локальных экстремумов («плюс-пика» и «минус-пика»), при этом дополнительно требуется, чтобы изменение энергии на интервале между парой этих «пиков» было не менее 20 дБ. В этом случае принимается, что соответствующий промежуток времени отвечает вокализованному звуку, в противном случае найденная пара («плюс-пик», «минус-пик») игнорируется.

Далее, внутри обнаруженных вокализованных участков речи ищутся границы сонорных звуков с помощью той же самой процедуры, но с использованием «тонкого» режима, и в частотных полосах со 2-ой по 5-ую. Найденные в каждом частотном диапазоне пары («плюс-пик», «минус-пик») сопоставляются и, и из всего набора «плюс-пиков» тоже выбирается наибольший по модулю, и таким образом формируется окончательная пара («плюс-пик», «минус-пик»), которая затем подвергается 2-м тестам. Первый тест заключается в проверке постоянства во времени всех составляющих спектра в диапазоне частот 0-0.6 кГц. Второй тест состоит в поиске существенных (не менее 5 дБ) изменений энергии в диапазоне 1.3 - 8 кГц. Участки, не прошедшие данные тесты, с большой вероятностью являются полугласными.

Положения взрывных согласных звуков ищутся вне пределов участков вокализованных фонем. На этих участках сначала ищутся пары («плюс-пик», «минус-пик») в диапазонах с 3-го по 6-ой. Затем в односторонних окрестностях пиков (у положительного - слева, у отрицательного - справа) ищутся границы речевых пауз.

Следует отметить, что некоторые шаги алгоритма требуют еще уточнения с целью повышения его эффективности. Так, при поиске экстремумов может возникнуть ситуация, когда соседствуют «пики» одного знака (например, в случае, когда промежуточный «пик» противоположного знака оказался ниже порога). Также

необходимо уточнить, каким образом лучше осуществлять вставку недостающего «пика» какой-либо полярности, или наоборот заменять группу следующих друг за другом однополярных «пиков» одним. Детектирование взрывных фонем также требует некоторой оптимизации.

В целом, по утверждению автора [3], данный алгоритм при отношении сигнал-шум 30 дБ обнаруживает более 90% процентов всех границ звуков и при этом не менее 25% обнаруженных границ оказываются ложными. Алгоритм, предложенный в [1], по утверждению его авторов, даже несколько превышает данные показатели, однако остается открытым вопрос о возможности сегментации речевого сигнала при более низких отношениях сигнал-шум, и при воздействии помех различного типа.

Все известные алгоритмы сегментации для выделения локальных особенностей речевых сигналов, соответствующих их фонетическим элементам используют разложение сигнала по некоторому фиксированному базису, например, используют кратковременный базис Фурье или какой-либо вейвлет-базис. При этом сами базисы являются универсальными и не зависят от локальных особенностей сигнала. Однако возможна и другая постановка задачи: для конкретного сигнала надо подобрать оптимальный базис, который бы в максимальной степени учитывал локальные особенности данного сигнала.

В [7] описан способ построения таких базисов, называемых локальными косинусными базисами. Локальный косинусный базис строится на основе перекрывающихся проекторов - индикаторов множества отрезков с концами $\hat{o}a\tilde{o}$. Каждому проектору соответствует отрезок

$I = {}_{x}a - \eta_{.}, a + \eta_{.z}$

где $\hat{o}^{\eta}\tilde{o}$ - множество отрезков перекрывания (фактически боковых частей окна). Для нашего случая можно взять одинаковые η для всех окон.

Проектор имеет форму окна вида

$$g(p,j)(t) = \{\blacksquare(\beta(\eta(-1)(t - a(p,j)))), \quad \text{если } t \in [a(p,j) - \eta, a(p,j) + \eta], \quad @1,$$

где $\beta(t)$ - профиль «фронтов» окна, определяется рекурсивно:

$$\beta_{-}EtF = \beta_{E}\sin_{E}\frac{\pi t}{2}FF,$$

обеспечивая требуемый порядок гладкости k, j – индекс некоторого подбазиса, определяемого ниже.

Для разложения сигнала по локальному косинусному базису используется специальное двоичное дерево, дающее разбиение носителя сигнала (части временной оси) на отрезки требуемой величины. Параметр j далее будет обозначать уровень узла

дерева, $1 \le j \le J$, *J*- высота дерева. Базис в подпространстве, соответствующем узлу дерева на глубине *j* в позиции *p*, имеет вид

$$B = {}_{\hat{o}}g\, EtF\hat{u}\frac{2}{2} \cos {}_{X}\pi\, Ek + 0.5\frac{F}{2}\, Et - aF_{Z\hat{o}},\ k \in \mathbb{Z}.$$

Для более точной аппроксимации сигнала в таком базисе следует брать окна меньшего размера и получать более детальное дробление носителя сигнала. Однако, при наличии в сигнале стационарных участков разной продолжительности целесообразно на разных таких участках использовать окна соответствующих им размеров, т.е. подстраивать под такие участки размер окна. Основная идея такого алгоритма состоит в поиске наилучшего базиса, минимизирующего некоторую выпуклую функцию стоимости, например, вогнутую сумму Шура. Реализация алгоритма основывается на идее динамического программирования, и поэтому достаточно эффективна. В оптимальном базисе меньшие по временной протяженности окна локализуют кратковременные особенности сигнала, а стационарным участкам большей продолжительности соответствуют окна большего размера. Таким образом, оптимальный для данного сигнала локальный косинусный базис, локализует особенности сигнала самых различных протяженностей, и поэтому его можно пытаться использовать для определения границ фонетических элементов речи.

Список литературы

1. Цыплихин А.И., Сорокин В.Н. Сегментация речи на кардинальные элементы // Информационные процессы. – 2006. – Т. 6, №3. – с.177-207.
2. Сорокин В.Н., Цыплихин А.И. Сегментация и распознавание гласных// Информационные процессы. – 2004. – Т. 4, №2. – с.202-220.
3. Liu S.Landmark detection for distinctive feature based speech recognition // J. Acoust. Soc. Amer., 1996, V. 100, P. 3417-3430.
4. Шелепов В.Ю., Ниценко А.В. Структурная классификация слов русского языка. Новые алгоритмы сегментации речевого сигнала, распознавания фонем и их классов // Искусственный интеллект. - 2005. - №4. - с.679-690.
5. Ермоленко Т.В., ЛащенкоА.В. Применение вейвлет-анализа для определения границ речи в зашумленном сигнале //Штучнийінтелект. -2009. -№ 1. – с. 35-40.
6. Алёшина Н.Д., Фёдоров В.Б. Об автоматической сегментации слитной речи // 62 научно-техническая конференция МИРЭА. Сборник трудов, часть 2. Физико-математические науки. М.: 2013
7. Малла С. Вейвлеты в обработке сигналов. М.: Мир, 2005.