УДК 004.934.2

ОБ ОДНОМ МЕТОДЕ АВТОМАТИЧЕСКОЙ СЕГМЕНТАЦИИ РЕЧЕВЫХ СИГНАЛОВ

Т.В. Шарий

Донецкий национальный университет, г. Донецк, Украина, tsphere@mail.ru

Исследованы изменения коэффициентов MFCC речевого сигнала на границах фонем. Введена мера кепстральной гладкости сигнала, учет которой, наряду с мерой спектрального перехода, позволил повысить точность фонемной сегментации.

СЕГМЕНТАЦИЯ, КЕПСТР, МЕРА СПЕКТРАЛЬНОГО ПЕРЕХОДА, МЕРА КЕПСТРАЛЬНОЙ ГЛАДКОСТИ, ФОНЕМА

Введение

Сегментация является важным этапом обработки речевого сигнала в научных исследованиях и работе автоматических систем распознавания речи (АСРР). От точности определения границ между фонемными сегментами зависит эффективность всей системы распознавания. На сегодняшний день наиболее точных результатов позволяет добиться ручная сегментация. В то же время, ручная сегментация является дорогостоящей операцией, требующей значительных затрат времени и участия эксперта. Большинство методов *автоматической* сегментации речевых сигналов [1, 2] основываются преимущественно на результатах распознавания фонем и фонетической транскрипции исходного речевого сигнала (контролируемая сегментация, supervised segmentation). Однако результаты распознавания часто ненадежны, а наличие транскрипции возможно только на этапе обучения лексических моделей.

Для определения лишь границ фонем, без необходимости распознавания («слепая» сегментация, «blind» segmentation), существуют простые методы, основанные на величине и скорости изменения определенных акустических характеристик. Две наиболее распространенные характеристики – это коэффициент перехода уровня сигнала через ноль (Zero Cross Rate, ZCR) [3] и мера спектрального перехода (Spectral Transition Measure, STM) [4]. Тем не менее, эксперименты показывают, что для надежной сегментации этих величин недостаточно. На основании результатов проведенных автором экспериментов в работе вводится новая характеристика речевого сигнала – мера кепстральной гладкости (Cepstral Smoothness Measure, CSM). Учет этой величины позволяет лучше выделять границы фонем при их коартикуляции в непрерывной речи и, таким образом, повысить точность автоматической сегментации речевых сигналов.

В общем случае, если в качестве минимальной единицы речи выбрана фонема, сегментация речи может производиться на трех уровнях: фонемном, слоговом и словесном. В статье исследуется самый важный и сложный уровень – определение фонемных границ. Для повышения точности фонемной сегментации в работе анализируются те параметры сигнала на фонемных переходах в непрерывной речи, которые могут служить индикаторами границ фонем.

1. Сегментация на основе энергии сигнала и коэффициента ZCR

В качестве одного из первых таких параметров использовался коэффициент перехода уровня сигнала через ноль [3]:

$$ZCR(f) = \frac{1}{L} \sum_{i=(f-1)\times L}^{f\times L-1} \frac{\left|\text{sign}(s_{i+1}) - \text{sign}(s_i)\right|}{2}, \quad (1)$$

где f – номер фрейма; L – количество отсчетов сигнала в данном фрейме; s_i – значение i-го отсчета сигнала.

Главное преимущество коэффициента ZCR состоит в том, что он не зависит от амплитуды сигнала в данном фрейме. В статье [3] вычисление коэффициента ZCR лежит в основе классификации типа «гласный/согласный/пауза» («vowel/consonant/pause», v/c/p). Классификация v/c/p необходима как базовый этап для выявления кандидатов пауз и границ предложений, а также извлечения просодических признаков речи. Для отделения пауз речи предложен адаптивный расчет уровня фонового шума. Использование конкретного порога энергии сигнала неприменимо, так как параметры окружающей обстановки меняются во времени, особенно в видео сценариях. Ниже подробно описан алгоритм сегментации v/c/p [3].

1. Входной сигнал квантуется во времени на неперекрывающиеся фреймы длиной 20мс, в которых вычисляется энергия сигнала, коэффициент ZCR и частота основного тона. Энергия рассчитывается путем усреднения значений модуля отсчетов сигнала в данном фрейме.

2. Огибающая частоты основного тона и энергии сигнала сглаживается.

3. Уровень шума (NoiseLevel) вычисляется на основе среднего (Mean_En) и дисперсии (Std_En) уровня энергии сигнала:

$$NoiseLevel = Mean\_En - 0.75 \cdot Std\_En. \quad (2)$$

Аналогичным образом, на основе математического ожидания (Mean_ZCR) и дисперсии (Std_ZCR) коэффициента перехода уровня сигнала через ноль, рассчитывается и ZCR-порог (ZCR_dyna):

$$ZCR\_Dyna = Mean\_ZCR + 0.5 \cdot Std\_ZCR. \quad (3)$$

4. Фреймы классифицируются в соответствии со следующими правилами (тип сегмента Consonant – согласный звук, сегмент Vowel – гласный звук, сегмент Pause – пауза):

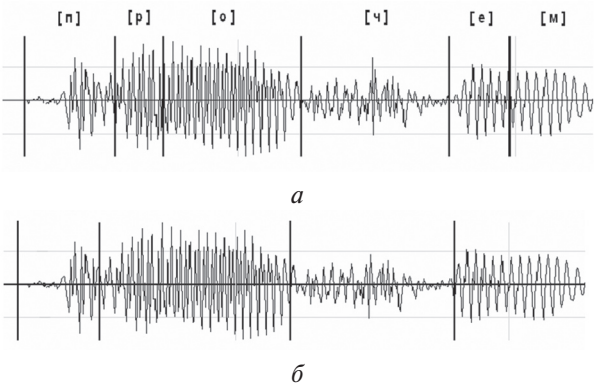
$$\begin{aligned} &IF\ ZCR > ZCR\_Dyna\ THEN \\ &FrameType = Consonant \\ &ELSE\ IF\ Energy < NoiseLevel\ THEN \\ &FrameType = Pause \\ &ELSE\ FrameType = Vowel \end{aligned} \quad (4)$$

5. Величина уровня шума пересчитывается как взвешенная средняя энергия фреймов на каждой границе гласного и фоновых сегментов.

6. Фреймы повторно классифицируются, согласно с правилами (4), с учетом обновленной величины уровня шума. Паузы объединяются путем удаления изолированных коротких согласных. Гласный разделяется надвое в позиции падения уровня энергии, если его длительность слишком велика.

На практике расчет уровня шума и классификация v/c/p выполняется каждые 4-5 секунд, чтобы учесть фоновые изменения.

Проведенные автором эксперименты показывают, что с помощью этого метода можно эффективно выделить шипящие и глухие согласные, однако точность сегментирования многих других звуков остается невысокой. На рис. 1 показаны результаты работы метода на примере фонемной последовательности «*прочем*». Видно, что хорошо отделены глухие согласные [п] и [ч], в то время как границы между остальными (звонкими) фонемами обнаружены не были. Тем не менее, идея вычисления уровня энергии сигнала для учета фонового шума на практике оказывается важной и полезной.



а

б

Рис. 1. Результаты сегментации фонемной последовательности «прочем»: ручная сегментация (*а*); сегментация методом v/c/p (*б*)

2. Сегментация на основе меры спектрального перехода

Более удовлетворительные результаты, по сравнению с методом v/c/p (ZCR), показывает метод автоматической сегментации, основанный на мере спектрального перехода (Spectral Transition Measure, STM), предложенной в работе [4] следующим образом:

$$STM(f) = \frac{\sum_{i=1}^{D} a_i^2(f)}{D}, \quad (5)$$

где D – размерность вектора признаков речевого сигнала, скорость изменения элементов которого рассчитывается в соответствии с формулой:

$$a_i(f) = \frac{\sum_{n=-I}^{I} c_i(f+n) * n}{\sum_{n=-I}^{I} n^2}, \quad (6)$$

где f – номер текущего фрейма; $c_i(f)$ – i-ый коэффициент MFCC-кепстра речевого сигнала, вычисленного для фрейма f; I – количество соседних фреймов, слева и справа от текущего, используемых для расчета скорости изменения кепстральных коэффициентов.

Параметр STM может быть интерпретирован как модуль скорости изменения спектра (в данном случае – MFCC-кепстра) сигнала. Согласно [4] положения максимумов величины спектрального перехода связаны с критическими точками, содержащими наиболее важную информацию для восприятия согласных звуков и слогов.

Алгоритм сегментации, основывающейся на параметре STM, включает следующие шаги [4]:

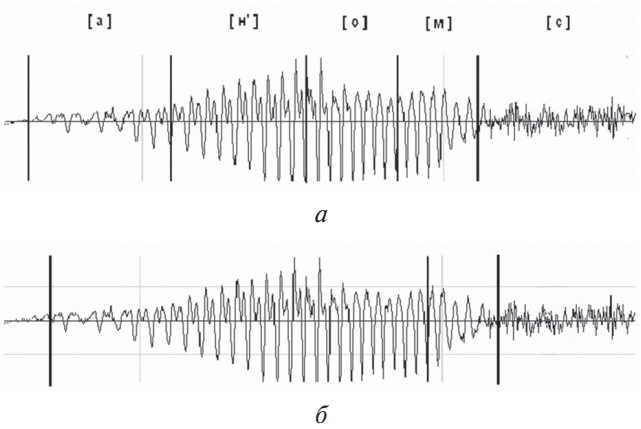
1. Речевой сигнал разбивается на фреймы длительностью 10 мс.

2. В каждом фрейме рассчитывается величина спектрального перехода (5). Параметр I в формуле (6) устанавливается равным 2.

3. Выделяются локальные пики кривой STM, построенной на шаге 2. Эти пики соответствуют фонемным границам.

Недостатком данного метода является то, что он редко определяет границы между гласными и согласными, в случаях выраженной коартикуляции или «плавного» звучания речи [4]. Пример на рис. 2 иллюстрирует данную проблему.

На рис. 2 видно, что коартикуляция фонем [а], [н’] и [о] в начале фразы «О нем существовали самые разные легенды» настолько сильна, что на границах данных фонем не наблюдается существенного изменения меры спектрального перехода.



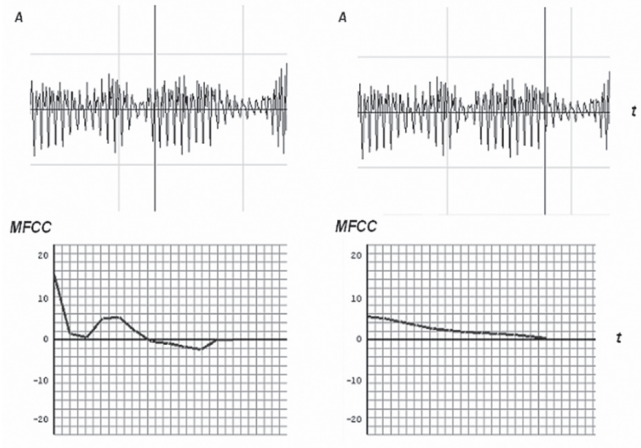
а

б

Рис. 2. Результаты сегментации фонемной последовательности «*о нём с...*»: ручная сегментация (*а*); сегментация на основе характеристики STM (*б*)

3. Сегментация с учетом меры гладкости кепстра

С другой стороны, автором экспериментально было установлено, что в подобных ситуациях кепстральная картина может приобретать специфическую форму – кепстр становится «гладким» (рис. 3).



а *б*

Рис. 3. Пример MFCC-кепстра: на границе между согласным и согласным звуком (*а*); на границе между согласным и гласным звуком (*б*)

Для отслеживания указанных ситуаций в работе вводится мера кепстральной гладкости (Cepstral Smoothness Measure, CSM):

$$CSM(f)=\begin{cases}\dfrac{c_{\max}(f)-c_{\min}(f)}{\sum_{i=1}^{D-1}\left|c_{i+1}(f)-c_i(f)\right|}, & c_{\max}(f)\neq c_{\min}(f)\\ 1, & c_{\max}(f)=c_{\min}(f),\end{cases} \quad (7)$$

где f – номер фрейма; $c_i(f)$ – i-ый коэффициент MFCC-кепстра речевого сигнала, вычисленного для фрейма f; $c_{max}(f)$ – значение максимального MFCC-коэффициента кепстра; $c_{min}(f)$ - значение минимального MFCC-коэффициента кепстра; D – размерность кепстра.

Величина CSM стремится к 1 при увеличении степени гладкости кепстра (CSM равняется 1, в случае, если кепстр представляет собой прямую линию).

Предложен новый метод сегментации речевых сигналов, учитывающий меру гладкости кепстра. Метод «STM+CSM» является дополнением к методу STM, описанному в работе [4], а также учитывает уровень фонового шума, по аналогии с методом классификации v/c/p [3]. Алгоритм «STM+CSM» включает следующие действия:

1. Речевой сигнал разбивается на фреймы длительностью 10мс.

2. Вычисляется уровень шума (NoiseLevel) на основе среднего (Mean_En) и дисперсии (Std_En) уровня энергии сигнала по формуле (2).

3. Энергия сигнала в каждом фрейме сравнивается с уровнем шума. Если оно не превосходит уровень шума, то переход к следующему фрейму.

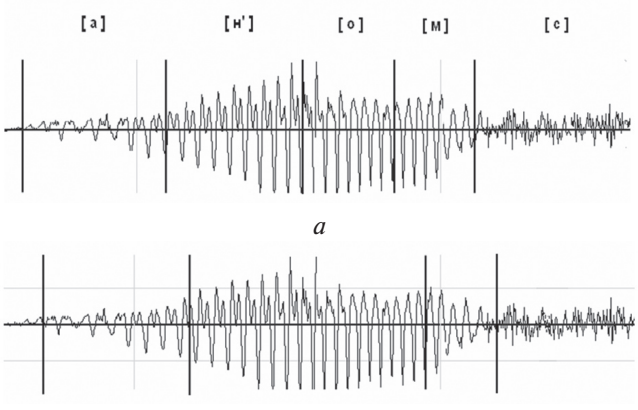
4. В текущем фрейме рассчитывается величина спектрального перехода (5).

5. В текущем фрейме рассчитывается величина кепстральной гладкости (7).

6. Выделяются локальные пики кривой STM, построенной на шаге 2. Эти пики соответствуют фонемным границам.

7. Находятся точки, в которых выражение *(CSM(f) – Delta)* меняет знак. Иначе говоря, находятся точки, в которых кепстр либо становится, либо перестает быть гладким. Найденные точки также считаются соответствующими фонемным границам. Значение порога *Delta* выбирается равным около 0.7-0.8.

На рис. 4 показаны результаты работы предложенного метода для предыдущего примера фонемной последовательности «*о нём с...*». Видно, что была обнаружена граница между фонемами [а] и [н‘], которая не фиксировалась методом STM. Тем не менее, остается некоторый процент случаев сложно разделяемых (даже вручную) фонем, при которых не происходит изменений ни меры спектрального перехода, ни меры кепстральной гладкости. В данном примере это граница между фонемами [н’] и [о].



а

б

Рис. 4. Результаты сегментации фонемной последовательности «*о нём с...*»: ручная сегментация (*а*); сегментация на основе характеристик STM и CSM (*б*)

4. Сравнение методов автоматической сегментации речевых сигналов

Тестирование методов сегментации проводилось на выборке из 50 файлов 3 дикторов речевой базы VoxForge [5]. Для экспериментов использовалась инструментальная среда редактирования речевых файлов [6]. На первом этапе файлы были размечены вручную. На втором этапе тестировались методы V/C/P, STM и STM+CSM (табл. 1).

Таблица 1

Сравнение показателей точности трех методов автоматической сегментации речи

Метод	Обнаруженные границы (%)	Пропущенные границы (%)	Ложно обнаруженные границы (%)
V/C/P	70,41	29,59	16,76
STM	82,92	17,08	20,19
STM+CSM	90,87	9,13	22,25

Для оценки качества сегментации использовались традиционные показатели – процент верно обнаруженных маркеров границ сегментов, процент пропущенных (не обнаруженных) маркеров и процент ложно обнаруженных маркеров сегментации.

Следует отметить, что в работе маркеры сегментации, расставленные вручную и рассчитанные одним из методов, считаются совпадающими, если разница в их позициях не превышает размера фрейма. Это более мягкое условие по сравнению с условием полного совпадения позиций маркеров. Таким образом, во-первых, учитываются особенности алгоритмов автоматической сегментации речевых сигналов, «разрешающая способность» которых ограничена размером фрейма. Во-вторых, снижается влияние погрешности, с которой эксперт сегментирует сигнал вручную.

Как видно из табл. 1, метод STM+CSM позволяет добиться лучшей точности сегментации по сравнению с методами v/c/p и STM: процент верно обнаруженных границ между фонемами на тестовой выборке составил 90.87%. При этом наблюдается незначительный рост числа ложно обнаруженных маркеров, который можно считать некритичным по отношению к числу верно найденных маркеров. Появление «новых» меток сегментации можно объяснить тем фактом, что при определенных условиях звуки речи, представляемые в фонетической транскрипции одной фонемой, характеризуются несколькими стадиями звучания, в каждой из которых кепстральная картина может изменяться и иметь место пик кривой меры спектрального перехода.

5. Применение метода STM+CSM

Стоит подчеркнуть тот факт, что в табл. 1 приведены показатели эффективности сегментации с точки зрения именно фонемной сегментации, то есть разбиения речевого сигнала на абстрактные единицы речи (фонемы), используемые при транскрипции фразы. Вопрос о минимальных единицах описания речевых сигналов до сих пор остается открытым и спорным среди ученых, но можно утверждать однозначно, что такие единицы, как фонемы, не описываются целиком своими акустическими (спектральными и просодическими) параметрами. Поэтому достичь хороших результатов (98% обнаруживаемых границ и выше) методами «слепой» сегментации невозможно в принципе. С другой стороны, метод STM+CSM, предложенный в работе, полезен с точки зрения определения участков сигнала с медленно изменяющимся кепстром. Результаты экспериментов можно интерпретировать таким образом, что в 90,87% случаев указанные участки совпадают с «реальными» фонемами. В остальных ситуациях эти участки являются либо частью (одной из стадий) звучания фонемы, либо последовательностью двух-трех очень плавно сменяющихся фонем. В любом случае, сегментация представляет собой лишь предварительный этап обработки речевого сигнала, и интерпретация ее результатов во многом зависит от модуля постобработки сигнала (back-end). В контексте сказанного, метод STM+CSM позволяет повысить точность предварительной сегментации, что способствует также росту эффективности распознавания речи в целом.

Важной особенностью является также то, что разработанный подход теоретически можно совместить с методами контролируемой сегментации, например с методом максимального правдоподобия (Maximum Likelihood Segmentation, ML) [1,2]. Данный метод основывается на принципах динамического программирования и для его использования необходимо задать число фонем, на которые необходимо сегментировать определенный участок речевого сигнала. С помощью метода STM+CSM можно точнее оценить количество фонем на данном участке речи и далее с помощью метода ML более качественно определить непосредственно границы между фонемами.

Заключение

Автором исследованы параметры речевого сигнала на границах фонем в непрерывной речи, которые важны для автоматической фонемной сегментации речевых сигналов. Проанализированы такие методы «слепой» сегментации, как v/c/p классификация и сегментация на основе меры спектрального перехода (STM). Введена мера гладкости кепстра для нахождения фонемных переходов и предложен новый метод STM+CSM, учитывающий эту характеристику при сегментации наряду с мерой спек-

трального перехода. Эксперименты подтвердили повышение точности выявления границ фонем. Дальнейшая работа связана с усовершенствованием алгоритма с целью увеличения процента верно определяемых меток, а также уменьшения числа неверно обнаруживаемых лишних меток фонемных сегментов.

Список литературы: 1. Parse Structure and Segmentation for Improving Speech Recognition [text] / W.P. McNeill, J.G. Kahn, D.L. Hillard, M. Ostendorf // IEEE Spoken Language Technology Workshop. – 2006. – P.90–93. 2. On the Robust Automatic Segmentation of Spontaneous Speech [text] / B. Petek, O. Andersen, P. Dalsgaard // Proceedings of ICSLP'96. – 1996. – P.913–916. 3. Speech segmentation without speech recognition [text] / D.Wang, L.Lu, H.J.Zhang // International Conference on Multimedia and Expo (ICME '03). – 2003. – Vol.1. – P.405–408. 4. On the Relation Between Maximum Spectral Transition Positions and Phone Boundaries [text] / S. Dusan, L.R. Rabiner // Proceedings of ICSLP'06. – 2006. – P.17–21. 5. VoxForge home page [electronic resource] / URL: http://www.voxforge.org/home / 30.10.2009. 6. Разработка инструментальной среды интеллектуального анализа аудиальных данных [текст] / А.А.Каргин, Т.В.Шарий // Труды VIII международной конференции «Интеллектуальный анализ информации ИАИ-2008», г.Киев. – 2008. – С. 558–564.

Поступила в редколлегию 26.10.2009

УДК 004.934.2

Про один метод автоматичної сегментації мовних сигналів / Т. В. Шарій // Біоніка інтелекту: наук.-техн. журнал. – 2009. – № 2 (71). – С. 61-65.

Стаття присвячена задачі автоматичної сегментації мовних сигналів. Досліджені акустичні параметри сигналу на межах фонем. Запропоновано новий метод «сліпої» сегментації STM+CSM, що грунтується на мірі спектрального переходу й мірі кепстральної гладкості. Експериментально підтверджено підвищення точності сегментації мовних сигналів з використанням нового методу.

Табл. 1. Іл. 4. Бібліогр.: 6 найм.

UDK 004.934.2

On one approach to automatic segmentation of speech signals / T.V. Shariy // Bionics of Intelligence: Sci. Mag. – 2009. – № 2 (71). – P. 61-65.

The paper is devoted to automatic speech segmentation task. The acoustic parameters at the phoneme boundaries are studied. The novel "blind" segmentation technique STM+CSM based on the Spectral Transition Measure and the Cepstral Smoothness Measure is developed. Experiments showed the increase of the segmentation accuracy of STM+CSM technique.

Tab. 1. Fig. 4. Ref.: 6 items.