электронный журнал
МОЛОДЕЖНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
Издатель ФГБОУ ВПО "МГТУ им. Н.Э. Баумана". Эл No. ФС77-51038.

УДК 004.52

Отложенная слоговая сегментация при распознавании речи

***Ду Цзяньмин**, студент*
кафедра «Информационные системы и телекоммуникации»,
МГТУ им. Н.Э. Баумана, Москва, Россия

Научный руководитель: Выхованец В.С., д.т.н.,
профессор кафедры «Информационные системы и телекоммуникации»
МГТУ им. Н.Э. Баумана, Москва, Россия
bauman@bmstu.ru

1. Введение

Распознавание речи – область речевых технологий, включающая целый ряд достаточно обособленных направлений, каждое из которых ориентировано на решение конкретных прикладных проблем, требующих отдельной теоретической и практической проработки.

Одно из таких направлений – автоматическое распознавание речи, или преобразование речи в текст. Выделяют несколько задач автоматического распознавания речи [**Ошибка! Источник ссылки не найден.**]. Это распознавание отдельных команд для управления устройствами и приложениями, распознавание отдельных фраз в системах массового обслуживания, поиск ключевых слов в слитной речи в системах мониторинга и распознавание слитной речи на больном словаре.

Несмотря на то, что распознавание отдельных команд, фраз и ключевых слов имеет эффективную практическую реализацию, задача достоверного распознавания слитной речи пока не решена полностью. Трудности возникают на всех стадиях обработки [**Ошибка! Источник ссылки не найден.**]: на стадии предварительной обработки речевых сигналов, при принятии решения «речь-пауза» и «вокализованный-невокализованный фрагмент», на стадии сегментации сигнала на значимые части, при выделении и распознавании фонем, при сопоставлении последовательности фонем словам из словаря, при выявлении и расстановке ударений, на стадии учета грамматических правил словообразования и построения предложений, и т.п.

Причины возникающих проблем кроются в том, что речь, зафиксированная в форме звука, многоаспектна, а ее транскрипция в текстовую форму настолько обедняет звуковой поток, что становится невозможным надежное решение даже простых прикладных задач. Другая группа причин низкой надежности распознавания связана с

тем, что ненадежность является объективным фактором речевого способа коммуникации: при восприятии речи человек тоже не все распознает надежно. Однако воспринимая речь, человек соотносит сказанное с действительностью, со своими знаниями о ней, со своим опытом, благодаря чему происходит восстановление нераспознанных фрагментов. В процессе восприятия речи человек активен, выдвигает гипотезы относительно содержания фрагментов речи и осуществляет смысловые замены.

Настоящая статья посвящена отложенной слоговой сегментации речи, которая необходима при использовании фонетических грамматик естественных языков, описывающих некоторые предметные области устного дискурса [**Ошибка! Источник ссылки не найден.**]. Фонетическая грамматика выражает правила построения правильных речевых отрезков с учетом как синтаксических, так и фонетических правил языка. Эти правила определяют некоторый язык-речь, рассматриваемый как множество тех речевых отрезков, которые могут быть распознаны с помощью правил грамматики. При этом анализу подвергаются не все имеющиеся фонетические единицы языка, а только те, которые предусмотрены текущим правилом фонетической грамматики и могут быть выявлены во входном речевом потоке. Отложенная слоговая сегментация речи заключается определение фонетического и семантического значения слога по их месту в слитной речи с учетом предыстории распознавания. В этом случае один и то же слог может быть сопоставлен различным словам в зависимости от текущего контекста.

2. Отложенная сегментация речи

Процесс автоматического распознавания речи разделяется на несколько стадий. На первой стадии выполняется предварительная обработка речевого сигнала. На второй стадии реализуется процесс определения граничных точек речи, или процесс принятия решения «сигнал-шум-пауза». Следующая стадия – сегментация речевого сигнала на значимые части (слоги, фонемы, аллофоны), является наиболее важным этапом в процессе распознавания. Сегментация речи – это процесс поиска границ между фразами, словами, слогами или артикуляторно-акустическими сегментами речевого сигнала.

Существует два основных метода сегментации речи. В первом методе производится сегментация речи при условии, что известна последовательность фонем распознаваемой фразы.

В другом методе априорные данные о фразе не использует, а границы сегментов определяются по степени изменения акустических характеристик сигнала. В

комбинированном методе решение о сегментации речи принимается как на основе априорных данных, так и на основе изменения акустических характеристик сигнала.

Сегментация потока устной речи на значимые единицы (паузы, слова, фонемы, аллофоны) оказалась довольно затруднительной. Выделяемые из этого потока фрагменты далеко не всегда обладают признаками синтаксических единиц, свойственных письменной речи. Установлено, что речевая коммуникация предполагает не только декодирование языковых единиц, но одновременное раскрытие их смыслов, т.е. выполнения сегментации речи через ее идентификацию [**Ошибка! Источник ссылки не найден.**]. Необходим также учет фоновых знаний, ситуации общения и структуры дискурса в самом процессе такого декодирования [**Ошибка! Источник ссылки не найден.**]. По этой причине прямая транскрипция речевого потока в текстовую форму видится малоперспективной.

Отложенная слоговая сегментация относится к комбинированным методам сегментации речи и основана на выделении гласных звуков. В русской речи гласные звуки обычно имеют большую энергию, чем согласные, а согласные звуки несут основную смысловую нагрузку [**Ошибка! Источник ссылки не найден.**]. Хотя в формантной картине гласных в различных видах речи наблюдается определенная вариативность [**Ошибка! Источник ссылки не найден.**], сегментация слитной речи на основе выделения гласных звуков и построением на их основе фонетических слогов путем присоединения близлежащих согласных звуков видится наиболее надежной.

Фонетический слог – гласный или сочетание гласного с одним или несколькими согласными, произносимый одним выдыхательным толчком речевого аппарата. Выделение фрагментов слитной речи, которые соответствуют фонетическим слогам, при отложенной слоговой сегментации производиться с учетом контекста распознавания, призванного учесть фоновые знания о предметной области дискурса и конкретную ситуацию речевого общения.

Контекст распознавания фонетического слога – это состояние фонетического анализатора, при котором гласный звук по-разному может быть соотнесен с образующим его слогом. При этом возможно присоединение к гласному звуку, как предшествующих фрагментов речи, так и последующих, вплоть до граничащих с ним другими гласными звуками. В этом случае принятие решения о слоговой сегментации осуществляется на поздних этапах распознавания, когда к этому процессу присоединяется фонетическая грамматика языка.

Процесс отложенной слоговой сегментации можно пояснить интерфейсом программного модуля, выполняющего выделение из речевого потока гласных и согласных

звуков (Листинг 1), где определены структуры вероятностей гласных и согласных звуков Vowel (строки 1-10) и Consonant (строки 11-32) соответственно, типы данных для дескрипторов структур (строка 34), кодов ошибок (строка 36), а также интерфейсные функции модуля (строки 37-50).

Листинг 1. Интерфейс модуля отложенной слоговой сегментации

```
1   /* Структура описания гласного звука */
2   typedef struct tagVowel {
3        int v1;       /* Вероятность звука А */
4        int v2;       /* Вероятность звука Э */
5        int v3;       /* Вероятность звука И */
6        int v4;       /* Вероятность звука О */
7        int v5;       /* Вероятность звука У */
8        int v6;       /* Вероятность звука Ы */
9        int v7;       /* Вероятность паузы */
10  } Vowel;
11  /* Структура описания согласного звука */
12  typedef struct tagConsonant {
13       int c01;      /* Вероятность звука П */
14       int c02;      /* Вероятность звука Б */
15       int c03;      /* Вероятность звука М */
16       int c04;      /* Вероятность звука Ф */
17       int c05;      /* Вероятность звука В */
18       int c06;      /* Вероятность звука Т */
19       int c07;      /* Вероятность звука Д */
20       int c08;      /* Вероятность звука Н */
21       int c09;      /* Вероятность звука С */
22       int c10;      /* Вероятность звука З */
23       int c11;      /* Вероятность звука Р */
24       int c12;      /* Вероятность звука Л */
25       int c13;      /* Вероятность звука К */
26       int c14;      /* Вероятность звука Г */
27       int c15;      /* Вероятность звука Х */
28       int c16;      /* Вероятность звука Ж */
29       int c17;      /* Вероятность звука Щ */
30       int c18;      /* Вероятность звука Ц */
31       int c19;      /* Вероятность звука Ч */
32  } Consonant;
33  /* Тип данных для дескрипторов */
34  typedef int Handle;
35  /* Тип данных для ошибок */
36  typedef int Error;
37  /* Открытие звукового файла */
38  Error    Open_Speech( char* file_path, Handle* file_handle );
39  /* Получение дескриптора первого гласного звука */
40  Error    First_Vowel( Handle file_handle, Handle* vowel_handle );
41  /* Получение дескриптора следующего гласного звука */
42  Error    Next_Vowel( Handle file_handle, Handle* vowel_handle );
43  /* Получение дескриптора предыдущего гласного звука */
44  Error    Previous_Vowel( Handle file_handle, Handle* vowel_handle );
45  /* Получение структуры гласного звука */
46  Error    Get_Vowel( Handle vowel_handle, Vowel* );
47  /* Получение структуры согласного звука */
48  Error    Get_Consonant( Handle vowel_handle, int number, Consonant* );
49  /* Закрытие звукового файла */
50  Error    Close_Speech( Handle file_handle );
```

Функции модуля позволяют открывать и закрывать звуковые файлы (строки 38 и 50), получать дескрипторы первого, следующего и предыдущего гласного звука (строки 40, 42, 44), получать структуру вероятностей гласных звуков Vowel по ее дескриптору (строка 46), а также структуру вероятностей согласных звуков Consonant по дескриптору слогообразующего гласного звука (строка 48), где number – номер по порядку предшествующего (number меньше нуля) или последующего (number больше нуля) согласного звука.

3. Первая сегментация

В процессе первой сегментации речи используются такие характеристики звукового сигнала как мгновенная энергия и число переходов через нуль. Знание этих численных характеристик позволяет надежно отделить гласные и звонкие согласные звуки от глухих согласных и пауз.

На рис. 1 показано зависимость мгновенной энергии сигнала eng и число переходов сигнала через ноль zrc для слога «ка». Из рисунка видно, что глухой согласный звук «к» имеют высокую частоту перехода через ноль и малую энергию, в то время как гласный звук «а» – высокую энергию и малое число переходов через ноль.

Однако для слога «за» со звонкой согласной картина совсем другая: энергия и число переходов через ноль звонкой согласной «з» мало отличаются от аналогичных характеристик гласного звука «а» (рис. 2).

Таким образом, после первой сегментации, отделяющей гласные и звонкие согласные от глухих согласных и пауз, необходимо выполнить вторую сегментацию, отделяющую звонкие согласные от гласных.

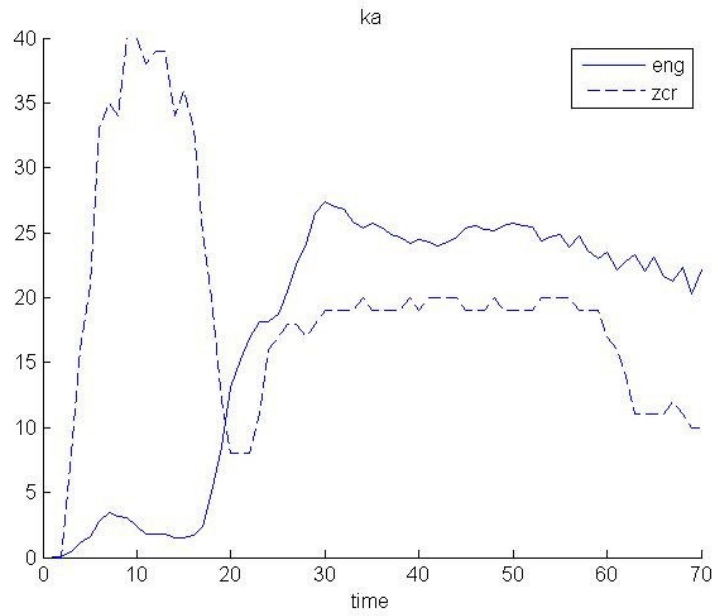

Рис. 1. Мгновенная энергия и число переходов через ноль в слоге «ка»

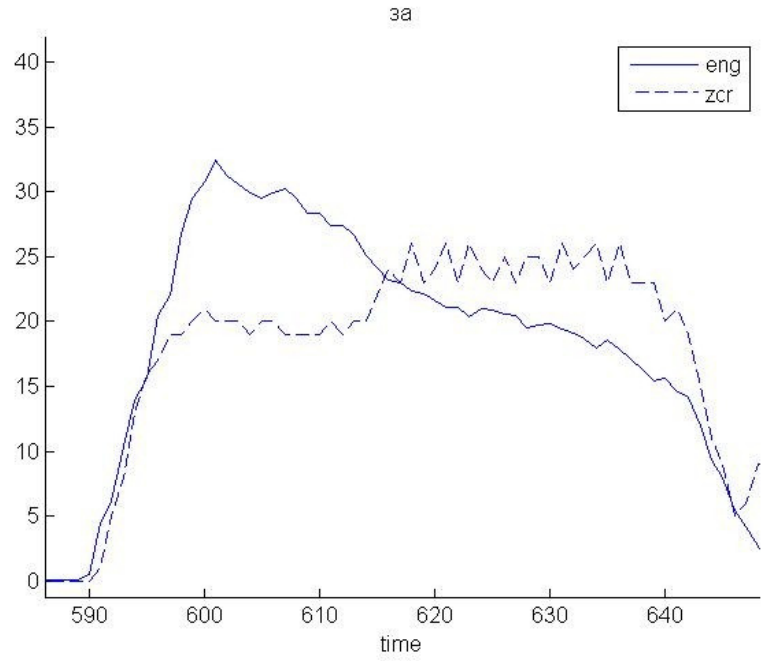

Рис. 2. Мгновенная энергия и число переходов через ноль в слоге «за»

4. Вторая сегментация

Как показано ранее, звонкие согласные трудно различимы по мгновенной энергии сигнала и числу переходов через ноль. Однако известно, что гласный звук имеет две или три ярко выраженные резонансные частоты в нижней части спектра [**Ошибка! Источник ссылки не найден.**]. Это позволяет с помощью спектрограммы сигнала отделять гласные звуки от звонких согласных (рис. 3).

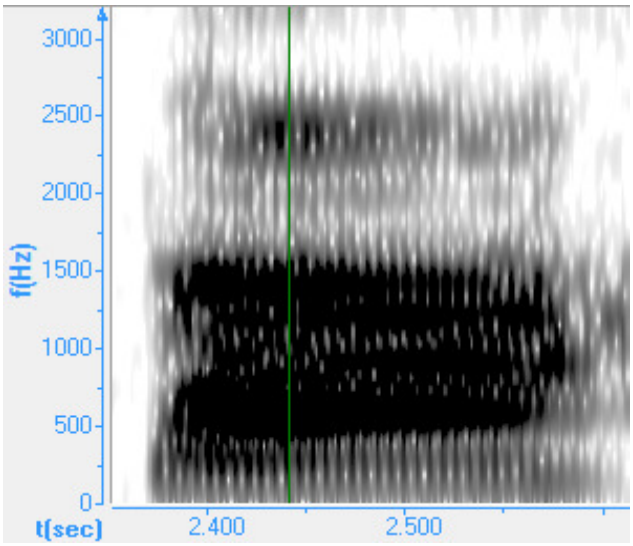

Рис. 3. Спектрограмма

На рис. 4 приведен временные диаграммы изменения мгновенной энергии сигнала для слога «за», полученные в различных диапазонах частот. Из рисунка видно, что в интервале от 590 до 620 мс в области высоких частот наблюдается повышенная энергия сигнала (произносится звонкий согласной «з»), а в интервале 620-660 мс – эта энергия мала (произносится гласный звук «а»). В обоих случаях в низкочастотной части спектра энергия сигнала достаточно высока.

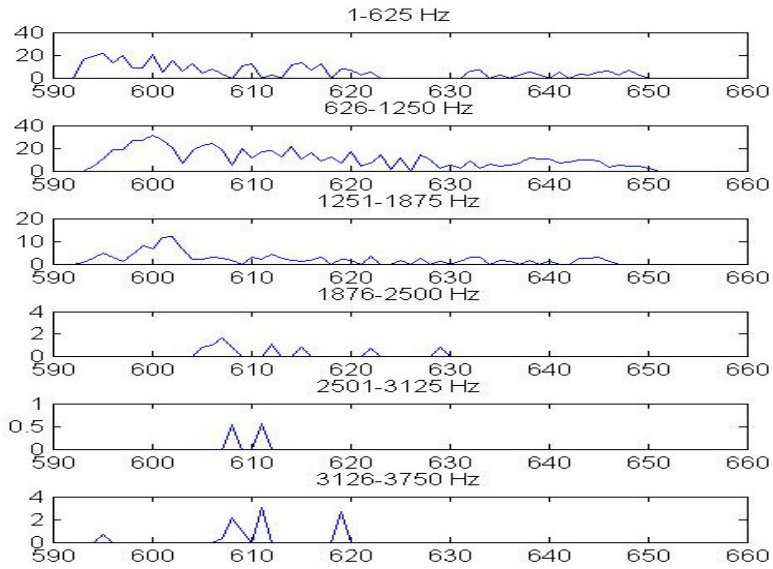

Рис. 4. Спектрограмма слога «за»

Для получения спектрограмм используются 24 треугольных фильтра, каждый их которых настаивается на свою полосу частот, определяемую характеристиками спектральной чувствительности человеческого уха (рис. 5).

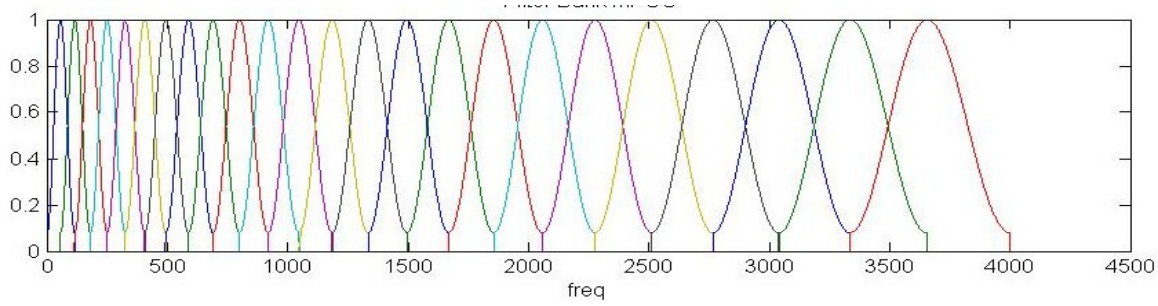

Рис. 5. Банк треугольных фильтров

Так как человеческое ухо не чувствительно к фазе сигнала и имеет логарифмическую чувствительность к его уровню, то для получения спектрограмм применено косинусное преобразование логарифма квадрата амплитуды звукового сигнала (кепстральное преобразование).

5. Распознавание

После сегментации звукового сигнала наступает этап распознавания звуков. Для распознавания звуков использован алгоритм динамической трансформации шкалы времени DTW (англ. Dynamic Time Warping), который позволяет измерять сходство двух временных последовательностей, различающихся скоростью своего воспроизведения [**Ошибка! Источник ссылки не найден.**].

Для работы алгоритма необходимо иметь стандартные образцы для всех звуков, которые получаются в процессе обучения алгоритма. Для обучения использованы образцы речи различных дикторов, которые размечены на фонетико-артикуляторные сегменты вручную.

Для исключения влияния на распознавание звуков индивидуальных характеристик дикторов применено вычисление основного тона каждого фрагмента речи, и пересчет звукового сигнала к одному и тому же тону. Такому же преобразованию должны подвергаться и фрагменты речи, подлежащие отложенной слоговой сегментации.

6. Заключение

В настоящей статье рассмотрены накопившиеся проблемы в области распознавания слитной речи и предложен подход к их решению на основе использования фонетических грамматик и отложенной слоговой сегментации. В отличие от других подходов при отложенной слоговой сегментации решение о границах слогов принимаются на заключительных этапах распознавания, что значительно повышает вероятность распознавания речи в целом. При выполнении слоговой сегментации не требуется высокая

надежность распознавания звуков, а также точное деление речи на слоги. Возникающие при распознавании ошибки исправляются контекстом распознавания, задаваемым текущим правилом фонетической грамматики.

Список литературы

1. Потапова Р.К. Речевое управление роботом. М., Комкнига, 2005. 328 с.
2. Шелепов В.Ю. К проблеме пофонемного распознавания // Искусственный интеллект. 2005. № 4. С. 662-668.
3. Выхованец В.С., Ду Цзяньмин, Назарова С.И. Контекстное распознавание слитной речи // VII-я Международная конференция «Управление развитием крупномасштабных систем» (MLSD'2013). Т. II. М.: ИПУ РАН, 2013. С. 312-315.
4. Венцов А.В., Касевич В.Б. Проблемы восприятия речи. М.: Едиториал УРСС, 2003. 240 с.
5. Андреева С.В. Речевые единицы русской речи: Система, зоны употребления, функции. М.: УРСС, 2006. 192 с.
6. Демачева Ю.С., Заранко К.М. Стенография: Практическое пособие. М.: Высшая школа, 1991. 368 с.
7. Сорокин В.Н., Цыплихин А.И. Сегментация и распознавание гласных // Информационные процессы. 2004, Т. 4, № 2. С. 202-220.
8. Keogh E., Ratanamahatana C.A. Exact indexing of dynamic time warping // Knowledge and Information Systems. 2005, No 7(3). P. 358–386.