Отложенная слоговая сегментация при распознавании слитной речи

Выхованец В.С., доктор технических наук

Ду Цзяньмин, аспирант

Рассмотрена проблема сегментации речи. Показано, что принятие решения о границах слога может быть осуществлено только на заключительных этапах распознавания. Предложено использование грамматического распознавания речи совместно с отложенной слоговой сегментацией. При отложенной слоговой сегментации первоначально выделяются гласные звуки, относительно которых на заключительных этапах распознавания принимается решение о группировке соседних с ними согласных звуков в слоги. При использовании предложенного подхода не требуется достижение высокой вероятности распознавания звуков слитной речи.

• автоматическое распознавание речи • сегментация речи • фонетические грамматики • контекстное распознавание

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 [1]. 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 большом 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 [2]: 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 объ-

ективным 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 [3]. 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 в определении фонетического и семантического значения слога по его 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.

1. ОТЛОЖЕННАЯ СЕГМЕНТАЦИЯ РЕЧИ

Процесс автоматического распознавания речи делится 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.

Существуют два основных метода сегментации речи. В первом случае производится сегментация речи при условии, что известна последовательность фонем распознаваемой фразы. В другом – априорные данные о фразе не используются, 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) оказывается довольно затруднительной процедурой. 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 [4]. Необходим также учет фоновых знаний, ситуации общения и структуры дискурса в самом процессе такого декодирования [5]. 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 [6]. Хотя в формантной

картине гласных в различных видах речи наблюдается определенная вариативность [7], сегментация слитной речи на основе выделения гласных звуков и построение на их основе фонетических слогов путем присоединения близлежащих согласных звуков видится наиболее надежной.

Фонетический слог – это гласный или сочетание гласного с одним или несколькими согласными, произносимый одним выдыхательным толчком речевого аппарата. В русском языке слоговыми (слогообразующими) являются только гласные звуки. Сколько в словоформе гласных, столько и слогов. При отложенной слоговой сегментации выделение фрагментов слитной речи, которые соответствуют фонетическим слогам, производиться с учетом контекста распознавания, призванного учесть текущее состояние фонетического анализатора.

1.1. Контекстное распознавание

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 других гласных звуков. 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), 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 фонемы У (звуки у, ю) */
4               int v2;    /* Вероятность фонемы А (звуки а, я) */
5               int v3;    /* Вероятность фонемы Е (звуки э, е) */
6               int v4;    /* Вероятность фонемы И (звук и) */
7               int v4;    /* Вероятность фонемы Ы (звук ы) */
8               int v5;    /* Вероятность фонемы О (звуки о, е, ё) */
9               int v6;    /* Вероятность паузы */
10  } Vowel;
11  /* Структура описания согласного звука */
12  typedef struct tagConsonant {
13              int c01;   /* Вероятность фонемы Л (твердый звук л) */
13              int c02;   /* Вероятность фонемы Л' (мягкий звук л) */
14              int c03;   /* Вероятность фонемы М (твердый звук м) */
14              int c04;   /* Вероятность фонемы М' (мягкий звук м) */
15              int c05;   /* Вероятность фонемы Н (твердый звук н) */
15              int c06;   /* Вероятность фонемы Н' (мягкий звук н) */
16              int c07;   /* Вероятность фонемы Р (твердый звук р) */
16              int c08;   /* Вероятность фонемы Р' (мягкий звук р) */
17              int c09;   /* Вероятность фонемы Б (твердый звук б) */
17              int c10;   /* Вероятность фонемы Б' (мягкий звук б) */
```

```
18          int c11;  /* Вероятность фонемы П (твердый звук п) */
18          int c12;  /* Вероятность фонемы П' (мягкий звук п) */
19          int c13;  /* Вероятность фонемы З (твердый звук з) */
19          int c14;  /* Вероятность фонемы З' (мягкий звук з) */
20          int c15;  /* Вероятность фонемы С (твердый звук с) */
20          int c16;  /* Вероятность фонемы С' (мягкий звук с) */
21          int c17;  /* Вероятность фонемы В (твердый звук в) */
21          int c18;  /* Вероятность фонемы В' (мягкий звук в) */
22          int c19;  /* Вероятность фонемы Ф (твердый звук ф) */
22          int c20;  /* Вероятность фонемы Ф' (мягкий звук ф) */
23          int c21;  /* Вероятность фонемы Д (твердый звук д) */
23          int c22;  /* Вероятность фонемы Д' (мягкий звук д) */
24          int c23;  /* Вероятность фонемы Т (твердый звук т) */
24          int c24;  /* Вероятность фонемы Т' (мягкий звук т) */
25          int c25;  /* Вероятность фонемы Г (твердый звук г) */
25          int c26;  /* Вероятность фонемы Г' (мягкий звук г) */
26          int c27;  /* Вероятность фонемы К (твердый звук к) */
26          int c28;  /* Вероятность фонемы К' (мягкий звук к) */
27          int c29;  /* Вероятность фонемы Х (твердый звук х) */
27          int c30;  /* Вероятность фонемы Х' (мягкий звук х) */
28          int c31;  /* Вероятность фонемы Ж (звук ж) */
29          int c32;  /* Вероятность фонемы Ш (звук ш) */
30          int c33;  /* Вероятность фонемы Щ (звук щ) */
31          int c34;  /* Вероятность фонемы Ч (звук ч) */
31          int c35;  /* Вероятность фонемы Ц (звук ц) */
31          int c36;  /* Вероятность фонемы Й (звук й) */
32 } Co      nsonant;
33 /* Тип данных для дескрипторов */
34 typedef int Handle;
35 /* Тип данных для ошибок */
36 typedef int Error;
37 /* Открытие звукового потока */
38 Error      Open_Speech( char* stream, Handle* stream_handle );
39 /* Получение дескриптора первого гласного звука */
40 Error      First_Vowel( Handle stream_handle, Handle* vowel_handle );
41 /* Получение дескриптора следующего гласного звука */
42 Error      Next_Vowel( Handle stream_handle, Handle* vowel_handle );
43 /* Получение дескриптора предыдущего гласного звука */
44 Error      Previous_Vowel( Handle stream_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 потока */
50 Error      Close_Speech( Handle stream_handle );
```

Функции модуля позволяют открывать и закрывать звуковые потоки (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.

1.2. Первая сегментация

Голосовые сигналы являются нестационарными, но в достаточно короткие периоды времени их можно считать стационарными. Это явление используется при выделении кадров конечной длительности, в которых нестационарностью сигнала можно пренебречь.

Для каждого выделенного кадра могут быть найдены характеристики, которые называются мгновенными характеристиками звукового сигнала. При частоте дискретизации звукового сигнала 8000 Гц размер кадра примерно равен 200 отсчетам или 20–30 мс. Для устранения резких изменений мгновенных характеристик сигнала, которые могут быть вызваны различными причинами, используется метод медианного сглаживания.

В процессе первой сегментации речи используются такие характеристики звукового сигнала во временной области, как мгновенная амплитуда, мгновенная энергия и число переходов сигнала через ноль (мгновенная частота). Знание этих численных характеристик позволяет надежно отделить гласные и звонкие согласные звуки от глухих согласных и пауз.

Известно, что глухие согласные звуки (в отличие от звонких) и гласные звуки легко могут быть распознаны с использованием мгновенных характеристик речевого сигнала. Глухие согласные имеют большие мгновенные частоты при небольшой мгновенной энергии. В свою очередь пауза характеризуется низкой мгновенной энергией и низкой мгновенной частотой.

На *рис. 1* показана 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. Аналогично обстоит дело и с другими глухими согласными.

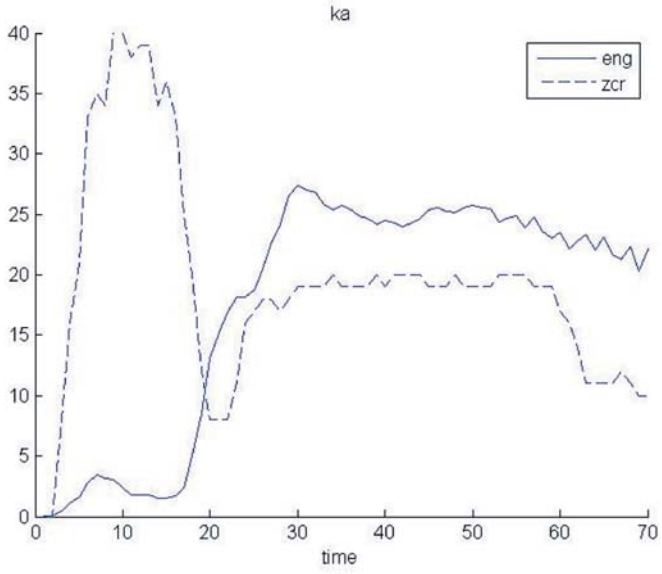

Рисунок 1 – Мгновенная энергия и мгновенная частота слога «ка»

Однако для слога «за» со звонкой согласной картина совсем другая: мгновенная энергия и мгновенная частота 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.

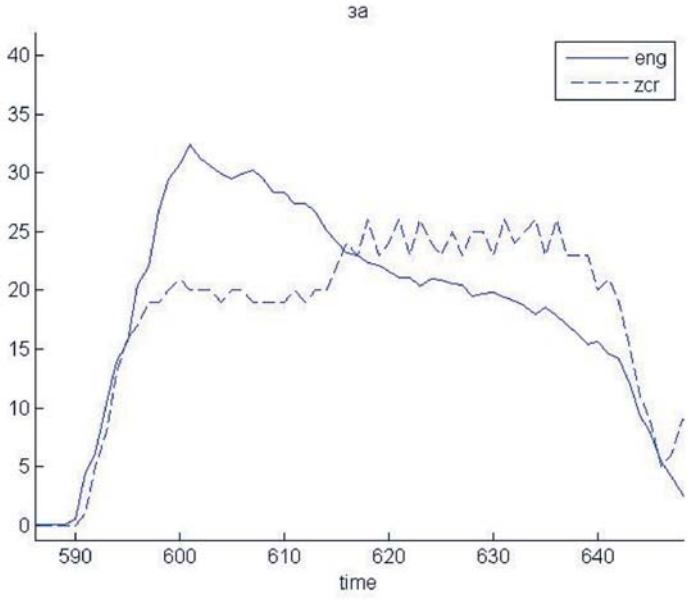

Рисунок 2 – Мгновенная энергия и мгновенная частота слога «за»

1.3. Вторая сегментация

Звонкие согласные трудно различимы по мгновенной энергии и мгновенной частоте сигнала. Однако известно, что гласный звук имеет две или три ярко выраженные резонансные частоты в нижней части спектра (форманты). Это позволяет с помощью спектрограммы сигнала отделять гласные звуки от звонких согласных (*рис. 3*).

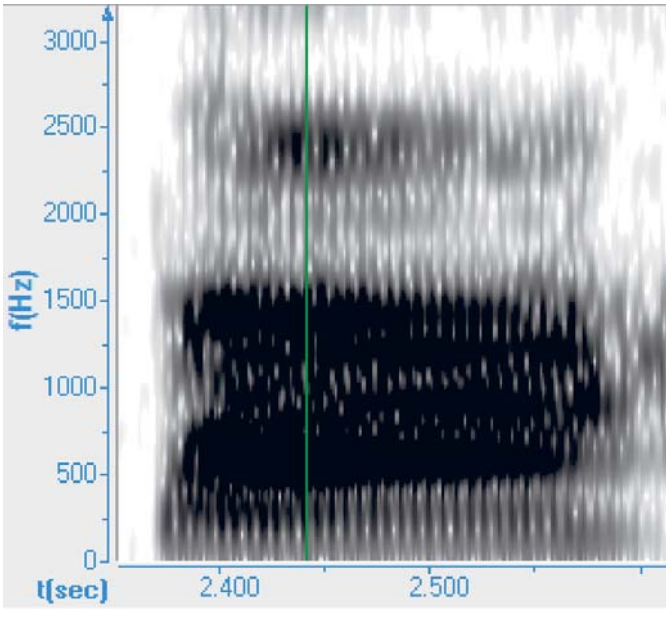

Рисунок 3 – Спектрограмма слога «за»

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.

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

Так как человеческое ухо не чувствительно к фазе сигнала и имеет логарифмическую чувствительность к его уровню [8], то для получения спектрограмм применено косинусное преобразование логарифма квадрата амплитуды звукового сигнала (кепстральное преобразование).

После определения момента перехода гласного или звонкого согласного звука в другие звуки используется значение первой производной мгновенной энергии сигнала, которая оценивается разностью энергий соседних кадров.

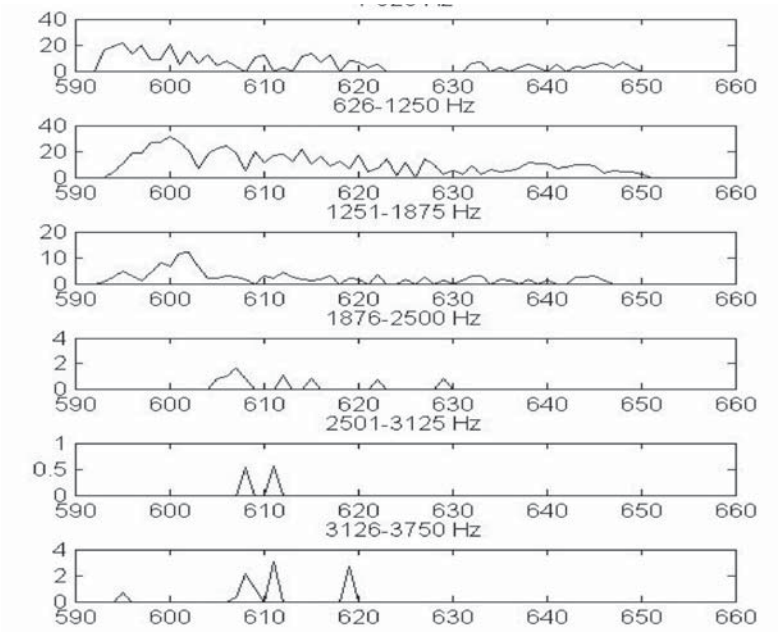

Рисунок 4 – Мгновенная энергия слога «за» в различных полосах частот

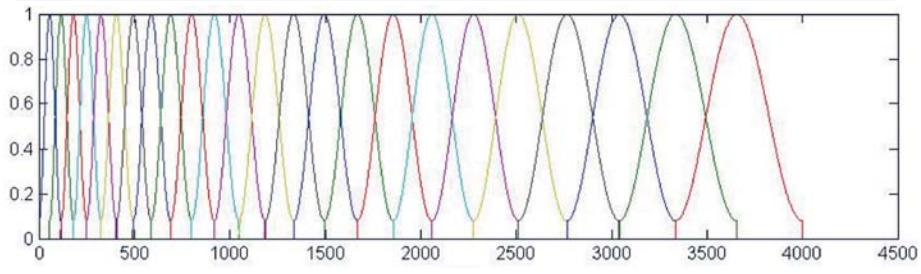

Рисунок 5 – Банк треугольных фильтров

2. РАСПОЗНАВАНИЕ

После сегментации звукового сигнала наступает этап распознавания звуков. Для исключения влияния на распознавание звуков индивидуальных характеристик дикторов применено вычисление основного тона каждого фрагмента речи, и пересчет звукового сигнала к одному тону. Распознавание гласных звуков осуществляется во временной области, а согласных – в частотной.

2.1. Распознавание гласных

Для распознавания гласных звуков использован алгоритм динамической трансформации шкалы времени DTW (англ. Dynamic Time Warping), который позволяет измерять сходство двух временных последовательностей, различающихся скоростью своего воспроизведения [9].

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 расчета вероятностей гласных звуков $P_v\ (v = \overline{1, N_v})$ используется формула

$$P_v = 1 - \frac{B_v}{\sum_{i=1}^{N_v} B},$$

где N_v – число гласных звуков, B_i – расстояния (стоимость пути) между выборочными значениями исследуемого и эталонных звуков.

Таблица 1

Энергия и частота сигнала для глухого согласного «П»

Полоса, Гц	«па»	«пе»	«пи»	«по»	«пу»	Среднее
F_1, 300-900	1,00	1,00	1,00	1,00	1,00	1,00
F_2, 900-1500	2,40	0,86	0,88	0,70	1,64	1,30
F_3, 1500-2500	0,77	0,92	1,62	0,13	0,80	0,85
F_4, 2500-4000	0,56	0,76	6,84	0,08	0,25	1,70
F_5, 4000-6000	0,30	0,30	3,01	0,06	0,15	0,76
F_6, 6000-8000	0,16	0,13	1,80	0,03	0,22	0,47
Z, частота	1120	1472	3456	800	416	1452,80

Таблица 2

Энергия и частота сигнала для шипящего согласного «С»

Полоса, Гц	«са»	«се»	«си»	«со»	«су»	Среднее
F_1, 300-900	1,00	1,00	1,00	1,00	1,00	1,00
F_2, 900-1500	1,23	0,78	0,63	0,92	1,11	0,93
F_3, 1500-2500	0,42	0,92	1,62	0,13	0,80	0,78
F_4, 2500-4000	0,39	0,76	6,84	0,08	0,25	1,66
F_5, 4000-6000	2,31	0,30	3,01	0,06	0,15	1,17
F_6, 6000-8000	2,69	0,13	1,80	0,03	0,22	0,97
Z, частота	5728	5184	6080	5472	3552	5203,20

2.2. Распознавание согласных

В отличие от гласных для согласных звуков алгоритм DTW дает плохие результаты. В результате исследований установлено, что разные согласные имеют различные распределения мгновенной энергии и мгновенной частоты в разных диапазонах частот звукового сигнала (*табл. 1-2*).Для расчета вероятностей согласных звуков P_c $(c = 1, N_c)$ используются формулы:

$$P_c = 1 - \frac{D_{c'}}{\sum_{i=1}^{N_c} D} \quad ; \quad D_i = \sqrt{1/2(Z_i - Z)^2 + \sum_{i=1}^{N_f}(F_j - F_i)^2},$$

где N_c – число согласных звуков, D_i – расстояния между выборочными значениями частоты и энергии в различных полосах частот исследуемого и эталонных звуков, Z_i и F_i – эталонные значения частоты и энергии в различных полосах частот эталонных звуков i, Z и F_i – значения частоты и энергии в различных полосах частот исследуемого звука, N_i – число анализируемых полос частот.

Таблица 3

Точность сегментации, %

Речь	Мужчина 1	Мужчина 1	Мужчина 1	Женщина 1	Женщина 2
Фрагмент 1	87	83	83	75	80
Фрагмент 2	85	85	90	71	75
Фрагмент 3	85	89	93	79	77

Таблица 4

Точность распознавания, %

Речь	Мужчина 1	Мужчина 1	Мужчина 1	Женщина 1	Женщина 2
Гласные	86	90	89	80	78
Согласные	61	55	62	59	53

2.3. Результаты экспериментов

Для определения точности сегментации и распознавания звуков использовались пять индивидуальных речевых сигналов, полученных от трех мужчин и двух женщин. Результаты точности сегментации и распознавания показаны в табл. 3–4.Средняя точность сегментации оказалась равна 82%, средняя точность распознавания гласных – 85%, а средняя точность распознавания согласных – 58%.

ЗАКЛЮЧЕНИЕ

В настоящей статье рассмотрены 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.

ЛИТЕРАТУРА

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.

СВЕДЕНИЯ ОБ АВТОРАХ

Выхованец Валерий Святославович,
доктор технических наук, доцент, ведущий научный сотрудник Института проблем управления им. В.А. Трапезникова РАН, профессор Московского государственного технического университета им. Н.Э. Баумана. Область научных интересов: многозначная логика, цифровая обработка сигналов, теория и технологии программирования, представление и обработка знаний, экспертные системы.

Ду Цзняьмин,
аспирант кафедры информационных систем и телекоммуникаций Московского государственного технического университета им. Н.Э. Баумана. Область научных интересов: обработка речевых сигналов, контекстное распознавание речи.

Vykhovanets V.S., Du Tzyan'ming

DELAYED SYLLABIC SEGMENTATION IN CONTINUOUS SPEECH RECOGNITION

The paper deals with the problem of segmentation of speech. It is shown that the decision of the boundaries of syllables can be carried out only in the final stages of recognition. The paper proposes the use of grammatical speech recognition together with deferred syllabic segmentation. When deferred syllabic segmentation originally distinguished vowel sounds, for which the final stages of the recognition decision about grouping neighboring consonants in syllables. When using the proposed approach is not required to achieve a high probability of recognition of speech sounds.

• automatic speech recognition • segmentation of speech • phonetic grammar • context recognition

REFERENCES

1. *Potapova R.K.* Rechevoe uprevlenie robotom. – M., Komkniga? 2005/ - 328 s.»

2. *Shelepov V.Yu.* K problem pofonemnogo raspoznavaniya // Iskusstvenny intellect. – 2005, № 4. – S. 662-668.»