УДК 004.934.2

А.К. Бурибаева, Г.В. Дорохина, А.В. Ниценко, В.Ю. Шелепов

СЕГМЕНТАЦИЯ И ДИФОННОЕ РАСПОЗНАВАНИЕ РЕЧЕВЫХ СИГНАЛОВ

Бурибаева А.К., Дорохина Г.В., Ниценко А.В., Шелепов В.Ю. **Сегментация и дифонное распознавание речевых сигналов.**

Аннотация. Статья посвящена описанию разработанной в Институте проблем искусственного интеллекта НАН и МОН Украины (Донецк) технологии распознавания речи, основанной на следующих основных этапах обработки: сегментация с использованием численного аналога полной вариации; создание дифонной базы; DTW-распознавание слов по эталонам, автоматически создаваемым из эталонов дифонов. Разработанная технология применима к распознаванию сверхбольших словарей, а также при разработке текстовых редакторов с голосовым вводом.

Ключевые слова: сегментация речевого сигнала, дифон, DTW-распознавание.

Buribayeva A.K., Dorokhina G.V., Nitsenko A.V., Shelepov V.Ju. **Segmentation and diphone recognition of speech signals.**

Abstract. The paper is devoted to speech recognition technology developed in Artificial intelligence Institute (Donetsk, Ukraine). It is based on the following main stages: segmentation with the help of full variation digital analogue; diphone-database creation; DTW-recognition of words based on diphone templates. The technology could be used for large vocabulary speech recognition as well as for development of text editors with voice input.

Keywords: segmentation of speech signal, diphone, DTW-recognition.

1. Введение. Цель настоящей статьи – дать сжатое, но по мере возможности цельное и законченное изложение технологии распознавания слов, которая разработана в Институте проблем искусственного интеллекта НАН и МОН Украины при участии Евразийского национального университета имени Л.Н. Гумилева (Казахстан). Она суммирует и обобщает отдельные результаты, опубликованные в работах с участием авторов, приведенных в списке литературы. Обсуждается распознавание русской речи. Аналогичная работа выполнена А. К. Бурибаевой для казахского языка. Публикации [1-16] представляют ряд последних работ по сегментации речи и примыкающим вопросам.

2. Упрощенная транскрипция русских слов (автоматическое построение). Говоря о так называемом пофонемном распознавании, в качестве объектов распознавания обычно имеют в виду не фонемы или аллофоны, а то, что уместно обозначать не очень строгим, но емким русским термином «звуки речи». Именно этим термином мы и будем пользоваться. При распознавании слов на основе использования звуков речи, очевидно, необходима предварительная транскрипция. Мы

используем свой упрощенный автоматический транскриптор, обеспечивающий потребности распознавания.

Транскриптор реализован как программа, заменяющая одни символы другими в соответствии с правилами, содержащимися в управляющем файле. Каждое из них записано в виде двух частей, соединенных знаком равенства. Слева стоят исходные символы буквенной записи слова, справа – символы которыми они заменяются в транскрипции. Значок \ означает ударение. Машина, транскрибируя слово, последовательно ищет вхождение левой части очередного правила, и если таковое обнаруживается, заменяет его правой частью. Вот примеры используемых правил:

ндс=нс – исключение в транскрипции непроизносимого [д], как в слове «ирландский»;

\о=1, о=а, 1=\о – комплекс правил для замены безударного [о] на [а] и сохранения [о] под ударением; ъю=jю, ью=jю – твердый или мягкий знак перед [ю] приводит к появлению перед ним звонкого согласного [j]; бк=пк – оглушение звонкого звука перед глухим; кб=гб – озвончение глухого звука перед звонким и так далее. На сегодняшний день управляющий файл содержит 747 правил замены. Наконец, прежде чем транскрибировать по указанным правилам, компьютер обращается к файлу исключений, в котором описываются процедуры транскрибирования некоторых целых слов, например, чт\о=што. Более подробное описание транскриптора для отдельных слов и слитной речи приведено в работе [17].

3. «В-Н» - обработка числового массива. Сглаживание сигнала. Мы используем 8-битную оцифровку звукового сигнала с частотой дискретизации 22050 Гц. В этом разделе описаны некоторые процедуры обработки сигналов и числовых массивов, применяемые в процессе сегментации.

Пусть имеется одномерный числовой массив и задан некоторый порог p. Построим символьную последовательность S, поставив в соответствие членам массива, которые больше либо равны p, символ «В» (выше порога), остальным символ «Н» (ниже порога). Для того чтобы устранить случайные единичные включения, для каждого промежуточного i-го элемента полученной символьной последовательности S выполняются две дополнительные обработки, обработка «тройками»:

$$\text{если } s[i-1]=s[i+1] \text{ и } s[i]\neq s[i-1], \text{ то полагается } s[i]=s[i-1]; \quad (1)$$

и обработка «четверками»:

$$\text{если } s[i]=s[i+3], \text{ но } s[i+1]\neq s[i] \text{ и (или) } s[i+2]\neq s[i], \text{ то}$$
$$\text{полагается } s[i+1]=s[i] \text{ и } s[i+2]=s[i]. \qquad (2)$$

Далее, имея дело с описанной процедурой будем называть ее «В-Н»-обработкой исходного числового массива с заданным порогом.

Назовем сглаживанием сигнала $y_1, y_2, \ldots$ обработку его 3-точечным скользящим фильтром

$$y_i = \frac{y_{i-1} + y_i + y_{i+1}}{3}. \qquad (3)$$

Рисунки 1 и 2 иллюстрируют влияние сглаживания на речевой сигнал и то, что оно уменьшает его полную вариацию

$$\sum_{i=0}^{N-1} \left| x_{i+1} - x_i \right|,$$

(здесь N – количество отсчетов в сигнале).

Рис. 1. Сигнал, отвечающий слову «досада».

Рис. 2. Сигнал, отвечающий слову «досада», после 10-кратного сглаживания.

Следующие три раздела описывают алгоритмы автоматической сегментации, то есть разбиения сигнала на участки, отвечающие отдельным звукам с одновременным отнесением последних к числу гласных (W), звонких согласных (C), фрикативных (F) и выделением паузообразных участков (P), отвечающих глухим взрывным или части аффрикаты.

4. Выделение в речевом сигнале глухих согласных. Распознавание в паре классов «шипящая-пауза». В данном разделе предлагается алгоритм выделения согласных [с],[ш],[ф],[х],[к],[п],[т] и их мягких аналогов, а также аффрикат [ц],[ч], произнесение которых происходит без участия голосовых связок. В основе его лежит обработка сигнала полосовым фильтром с полосой пропускания от 100 до 200 Гц. При этом можно использовать простейший фильтр, описанный, например, в [18, с. 147–160], нормируя получающийся сигнал делением на амплитуду и умножением на 256. Вот как выглядит запись слова «ОСА» до и после такой фильтрации (рис. 3–4).

Рис. 3. Визуализация сигнала «оса».

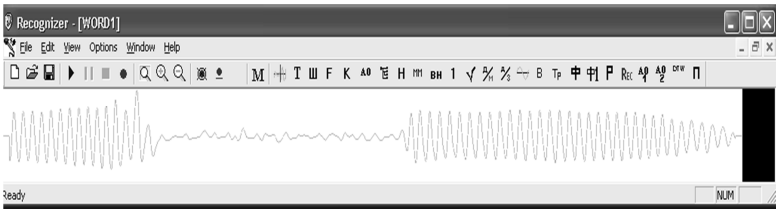

Рис. 4. Визуализация того же сигнала после фильтрации.

Назовем момент дискретного времени точкой постоянства, если в следующий момент сигнал принимает то же значение. Глухие звуки отличаются от всех остальных тем, что после упомянутой фильтрации их участки становятся подобными паузе и содержат большое число точек постоянства. Таким образом, на этих участках разность между числом точек непостоянства и числом точек постоянства будет отрицательной, что позволяет выделить их в массиве таких разностей, построенном для последовательности окон в 256 отсчетов. Отметим, что выделенный участок глухих звуков, может содержать шипящую,

как в слове «лошадь», паузу, как в слове «лапа», или и то и другое вместе.

Рассмотрим для произвольно выделенного участка речевого сигнала численный аналог полной вариации «с переменным верхним пределом»:

$$V(0)=0,\ V(n)=\sum_{i=0}^{n-1}\left|x_{i+1}-x_i\right|. \qquad (4)$$

Построим функцию $W(n) \overset{\bmod 256}{=} V(n)$. Пусть N_1 — максимальное число такое, что $V(N_1)\le 255$. Полагаем $W(n)=V(n)$ при $0\le n\le N_1$, $W(N_1+1)=0$, $W(n)=\sum_{i=N_1+1}^{n-1}\left|x_{i+1}-x_i\right|$ при $N_1+1<n\le N_2$, где N_2 — максимальное число такое, что $W(N_2)\le 255$ и так далее (рис. 5-а, 5-б). В результате возникает массив чисел

$$N_1, N_2-N_1, N_3-N_2\ldots \qquad (5)$$

Каждое из них — это длина участка, на котором величина $W(n)$ возрастает от 0 до 255.

На сегменте шипящей величина (4) быстро растет, поэтому участки возрастания величины $W(n)$ от 0 до 255 коротки, то есть числа (5) относительно малы. На сегменте паузы ([к],[п],[т],[ф],[х] и часть аффрикаты) величина (4) растет медленно и поэтому числа (5) относительно велики. Для различения шипящей и паузы введем порог p (в системе авторов он взят равным 200). Возьмем выделенный методами предыдущего раздела сегмент глухих согласных и построим для него последовательность чисел (5). Те участки, для которых числа (5) превосходят p, относим к паузе (их объединение маркируем символом P), остальные — к шипящей (маркируем ее символом F). В результате компьютер расставит маркированные границы шипящих и пауз, как на первом рис. 5-а.

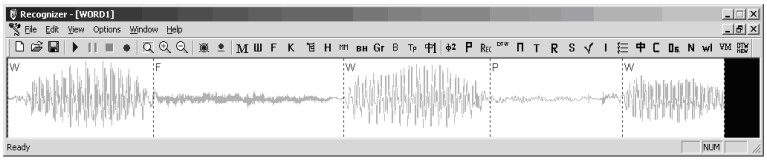

Рис. 5-а. Сигнал, отвечающий слову «осока».

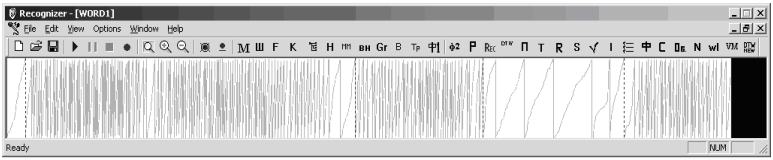

Рис. 5-б. График функции *W(n)*, отвечающей сигналу на рисунке 5-а.

5. Сегментация речевого сигнала без [ж], [з]. Рассмотрим сигнал, отвечающий слову, состоящему только из голосовых звуков и не содержащему **[ж]** и **[з]**. Разобьем его на окна по 256 отсчетов, и на каждом из них вычислим численный аналог полной вариации

$$V = \sum_{i=0}^{254} \left| x_{i+1} - x_i \right| . \qquad (6)$$

Создадим числовой массив для набора первых 20-ти окон и в качестве порога возьмем среднее. Сделаем «В-Н»-обработку с этим порогом. Затем упомянутый набор сдвинем на одно окно вправо и повторим процедуру. Продолжая, получим таблицу (рис. 6).

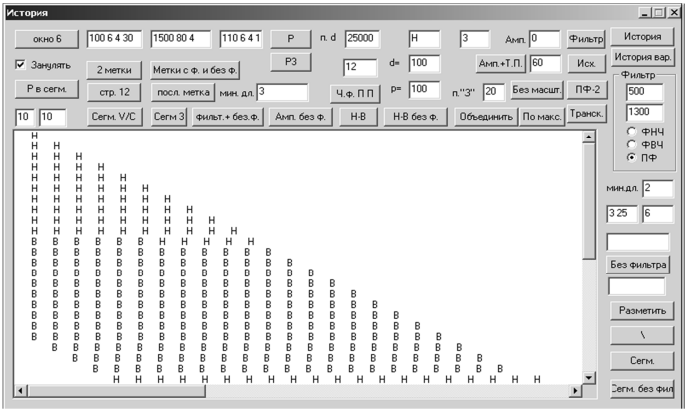

Рис. 6. Таблица, используемая при сегментации.

Далее просматриваются все строки полученной таблицы, и создается новая символьная последовательность *S*. Если текущая *i*-я строка таблицы начинается и заканчивается одним и тем же символом («Н» или «В»), то на *i*-ю позицию в *S* записывается соответствующий символ. Иначе считается количество вхождений каждого из символов в данной строке. Если количество «В» превышает количество «Н» или равно ему, то в *S* на соответствующую позицию записывается «В», иначе «Н». К полученной последовательности применяется обработка (1) и (2). Метки сегментации ставятся там, где происходит смена символов «Н» на «В», или «В» на «Н». В-участок считается соответствующим гласному (возле левой метки проставляется символ W). Н-участок считается соответствующим звонкому согласному (возле левой метки проставляется символ C).

На рисунке 7 показан результат для слова «мимо», отсегментированного в соответствии с только что описанным алгоритмом.

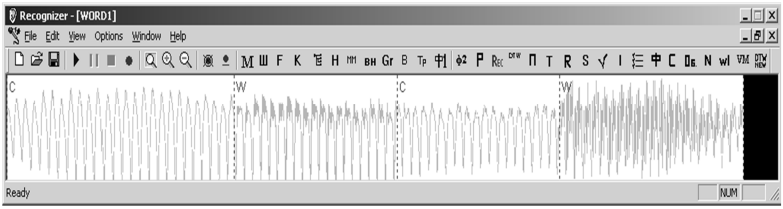

Рис. 7. Сегментация слова “мимо”.

Если слово содержит шипящие или паузы, то мы выделяем их, как описано выше, после чего значения величины (6) для соответствующих им окон полагаем равными нулю и сегментируем сигнал подряд только что описанным способом (шипящие и паузы автоматически попадают в число Н-участков). Для надежного выделения звонкого согласного после шипящего или паузы порядок формирования S непосредственно после шипящего или паузы меняется: если в строке появляется «В», но она заканчивается на «Н», то ей сопоставляется «Н». Дальше все, как изложено выше. Аналогичная ситуация с голосовым согласным непосредственно перед шипящей или паузой.

6. Сегментация при наличии согласных «Ж, З». Определенная трудность возникает при выделении в ходе сегментации участков согласных [ж], [з]. Они содержат существенную шумную компоненту (произносятся с той же артикуляцией, что и [*ш*], [с] и отличаются от них добавлением голоса). Поэтому для них значения величины (6) являются относительно большими (существенно превосходящими ее значения для других звонких согласных) и они могут не попасть при вышеописанной сегментации в число «Н»-участков. Участки, соответствующие указанным звукам, целесообразно определять заранее, используя модифицированный алгоритм сегментации.

Для того чтобы перевести искомые участки в число согласных, для которых величина (6) мала, сигнал подвергается 5-кратному сглаживанию (см. (3)). После этого он сегментируется описанным выше способом. Для участков [ж], [з] характерна относительно большая вариация и значительное ее уменьшение при 5-кратном сглаживании Поэтому выделенные с применением сглаживания «Н»-участки целесообразно промаркировать, вычисляя для каждого полученного «Н»-участка среднее значение величин (6) для поточечной разности исходного и пятикратно сглаженного сигнала. Если оно превышает некоторый порог p, участок считается

маркированным, в противном случае – нет. Для одного из авторов и его оборудования эффективным оказывается порог $p = 200$.

Описанное сглаживание сигнала перед сегментацией помогает таким образом выделить участки [ж] и [з] как «Н»-участки, но приводит к значительному ухудшению результата при разделении *М* и *И*. Поэтому целесообразно проводить окончательную сегментацию для исходного сигнала, сохранив полученную информацию относительно границ маркированных участков. Именно, после того, как маркированные участки выделены, величины (6), отвечающие соответствующим окнам, полагаются равными нулю. В результате маркированные участки оказываются при работе с таблицей рисунка 6 автоматически включенными в число «Н» - участков и можно провести сегментацию исходного сигнала в соответствие с описанным выше общим алгоритмом. Если при этом возникают метки, отстоящие от уже существующих меток для маркированных участков на расстояние не более чем в три окна по 256 отсчетов, они рассматриваются как лишние и удаляются. В работе [19] приведен алгоритм распознавания звука [р].

7. Ограничения на длину фонемы. Случай двух идущих подряд согласных. Содержание предыдущих разделов требует некоторых дополнений. При произнесении двух согласных подряд возможна промежуточная перестройка артикуляционного аппарата, при которой голосовая щель на какое-то время оказывается открытой. В результате возникает участок сигнала, который уместно назвать ***гласной вставкой.*** На рисунке 8 приведен пример возникновения гласной вставки между согласными [б] и [л] в слове «благо»:

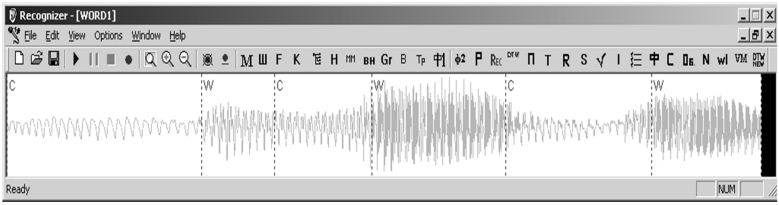

Рис. 8. Сегментация слова “благо” с гласной вставкой между [б] и [л].

Поскольку мы стремимся к тому, чтобы сегментация всегда соответствовала транскрипции слова, гласная вставка должна убираться. Это достигается за счет априорного ограничения на длину гласного: гласный, длина которого (количество 256-окон) меньше

задаваемой величины, должен убираться. Рекомендуемое значение этой величины 5.

Аналогичные ограничения на минимальную длину целесообразно ввести и для звуков всех остальных классов – звонких согласных, шипящих и пауз, задавая минимальные длины в начале, в середине и в конце слова.

Далее, описанный метод сегментации ориентирован на слова со строгим чередованием гласных и согласных. В русском языке немного слов, где налицо два гласных или более двух звонких согласных подряд. Эти случаи мы пока не рассматриваем. Зато имеется очень много слов, где идут подряд два звонких согласных. Проставить между ними метку несложно, ее нужно ставить просто посредине соответствующего Н-участка. Действительно, как показывает опыт, два идущих подряд звонких согласных при естественном произнесении всегда имеют практически одинаковую длину. Сделать их существенно отличающимися по длине можно только с помощью специальных усилий. Значительно сложнее решить на уровне сегментации, сколько же в рассматриваемом Н-участке согласных: один или два? Предлагается делать это, учитывая длину Н-участка. Для участка в середине слова задаются два порога (на сегодня мы работаем с числами 18 и 20). Если длина участка меньше первого порога, считается, что согласный один. Если длина участка больше второго порога, считается, что согласных два, и участок разбивается на две равные части, каждая из которых маркируется как «С». Если длина участка заключена между порогами, допускаются оба варианта, участок делится пополам, но символ «С» в начале второй половины не проставляется. Аналогичные процедуры применяются к Н-участку в начале и в конце слова. Согласный в начале слова обычно длиннее, чем в середине, поэтому здесь предлагается брать другие пороги (у нас это 22 и 24).

Сказанное в равной мере относится к шипящим и паузам. Напомним, например, что, согласно правилам русской фонетики буквосочетание «сш» в произношении реализуется долгим (удвоенным) звуком [ш]. Особо подчеркнем, что в русском языке звук [щ] (мягкое [ш]) - всегда долгий. Пороги для шипящих и пауз свои.

8. Построение системы признаков. Представление слова. Пусть записанный сигнал, соответствующий какому-либо русскому слову, состоит из N чисел $y_1, y_2, \ldots, y_N$. Моменты дискретного времени, как и сами числа, будем называть отсчетами. Пусть l – число отсчетов между двумя соседними локальными максимумами функции

$y(i) = y_i$, $(i = 1,2\ldots N)$. Назовем сужение функции на соответствующий интервал полным колебанием. Если максимумы - не строгие, то под l будем понимать число отсчетов от начала первого максимума до начала второго. Определим величину z:

$$z = l,\ 1 \le l < 20;$$

$$z = 20 + \frac{l-20}{6},\ 20 \le l < 50;$$

$$z = 25 + \frac{l-50}{10},\ 50 \le l < 90;$$

$$z = 29,\ l \ge 90.$$

Ближайшее целое число, не превосходящее z, назовем длиной соответствующего полного колебания. Таким образом, длина полного колебания учитывается тем точнее, чем оно короче и тем менее точно, чем оно длиннее. Выделим участок сигнала и обозначим через n общее число полных колебаний на этом участке, через n_1 — число полных колебаний длины 2,...,через n_{28} — число полных колебаний длины 29.

Поставим в соответствие выделенному участку вектор

$$(x_1, \ldots, x_{28}, \varepsilon), \qquad (7)$$

где $x_i = \frac{n_i}{n}$, $i = 1,2,...28$, ε – отношение амплитуды рассматриваемого участка сигнала к амплитуде всего сигнала. Величина ε вводится для того, чтобы надежно выделять паузу, а нормировка ее делается, чтобы отвлечься от громкости произносимого.

Разобьем записанный сигнал на отрезки по 368 отсчетов в каждом (удвоенный квазипериод основного тона для мужского голоса средней высоты). Пусть в нашем сигнале содержится k таких полных отрезков. Для каждого из них вычислим вектор (7). В результате мы представляем сигнал в виде траектории, то есть последовательности k точек в 29-мерном пространстве:

$$A = (a_1, a_2, \ldots, a_k).$$

9. Распознавание слов по эталонам. Алгоритм DTW. Пусть представление некоторой реализации слова принимается за эталон. Как изложено в предыдущем разделе, мы представляем его в виде набора 29-мерных векторов:

$$E = \left(e_1, e_2, \ldots, e_m\right). \tag{8}$$

Такой эталон записывается для каждого из слов распознаваемого словаря.

Пусть теперь

$$A = \left(a_1, a_2, \ldots, a_k\right) — \tag{9}$$

представление слова, которое подлежит распознаванию. Имея в виду потребность выравнивания темпа на протяжении произнесения слова, расстояние между (8) и (9) будем, следуя Т.К. Винцюку (см. [20]), определять методом динамического программирования (алгоритм DTW).

Выберем для определенности за расстояние между векторами сумму модулей разностей соответствующих координат (l_1 – метрика). Обозначим расстояние между векторами e_j и a_j наборов (8), (9) через D_{ij} и для всех $1 \le i \le m$, $1 \le j \le k$ определим величину C_{ij}:

$$\begin{gathered} C_{11} = D_{11}, \;\; C_{i1} = D_{i1} + C_{i-1,1}, \\ C_{1j} = D_{1j} + C_{1,j-1}, \;\; C_{ij} = D_{ij} + \min\left(C_{i-1,j}, C_{i,j-1}, C_{i-1,j-1}\right), \\ 2 \le i \le m, \;\; 2 \le j \le k. \end{gathered} \tag{10}$$

C_{ij} используется как величина, пропорциональная расстоянию между частью сигнала, соответствующего (8), от начала до *i*-го отрезка включительно и частью сигнала, соответствующего (9), от начала до *j*-го отрезка включительно. Расстояние между полными сигналами определим как $\dfrac{C_{mk}}{\sqrt{m^2 + k^2}}$. Деление на корень введено для баланса расстояний распознаваемого слова до длинных и коротких слов словаря (см.[21]).

Числа C_{ij} образуют DTW-матрицу, которая заполняется при движении из нижнего левого угла вправо и вверх. Двигаясь из правого верхнего угла влево и вниз, устанавливаем понятие соответствия

между векторами: $e_m \sim a_k$ по определению; если $e_i \sim a_j$, то в соответствии с тем, какой член в скобках (10) реализует минимум, получаем $e_{i-1} \sim a_j$ или $e_i \sim a_{j-1}$ или $e_{i-1} \sim a_{j-1}$. Тогда ясно, что при вычислении расстояния между (8) и (9) суммируются только расстояния между соответствующими векторами и происходит желаемое выравнивание по времени.

Эталоны одинаковой длины T можно усреднять. Если $E = (e_1, e_2, \ldots, e_T)$ результат усреднения n эталонов и $A = (a_1, a_2, \ldots, a_T)$ – n+1-ый эталон, то полагаем $e_i^{'} = \frac{n}{n+1} e_i + \frac{1}{n+1} \frac{a_j + \ldots + a_{j+k}}{k+1}$, где $a_j, \ldots, a_{j+k}$ – все вектора, соответствующие вектору e_j.

10. DTW- распознаватель с эталонами слов, созданными из эталонов дифонов. Создание эталонов дифонов. До сих пор мы говорили о распознавании целых слов по эталонам, создаваемым с помощью голоса. Обучение такой системы предполагает предварительное (и зачастую неоднократное) произнесение каждого слова распознаваемого словаря. Такая процедура пригодна для малых словарей, при большом словаре она нереальна. Подход к решению проблемы состоит в переходе к использованию более мелких речевых единиц.

Учитывая, что в речи соседние звуки влияют друг на друга, перспективным представляется использование для распознавания пар соседних звуков. Взаимное влияние соседних звуков называют коартикуляцией, а участок перехода от одного звука к другому – межфонемным переходом. Дифоном называют отрезок речевого сигнала между серединами двух соседних звуков. Таким образом, дифон содержит соответствующий межфонемный переход. Кратко авторы решились бы сформулировать свою сегодняшнюю точку зрения в виде следующего тезиса:

Один из возможных ключей к распознаванию речи лежит в межфонемных переходах.

Создание автоматической системы распознавания требует формализации используемых понятий. В связи с этим введем понятие «укороченного дифона».

В предыдущих разделах мы научились автоматически разбивать речевой сигнал на участки, отвечающие отдельным звукам. Мы будем понимать под укороченным дифоном, соответствующим межфонемному переходу внутри слова, участок стандартной длины: 3 окна в 368 отсчетов слева от метки между звуками и 3 таких же окна справа от той же метки. В дальнейшем мы позволим себе опускать эпитет «укороченный» и обозначать описанный объект прежним термином «дифон». Эталон дифона – набор 6-ти соответствующих векторов, его имя – пара символов для соответствующих звуков. Кроме того, мы используем участок в 3 окна в начале слова (имя – символ начального звука с добавлением символа 0) и участок в 3 окна в конце слова (имя – символ конечного звука с добавлением символа 2), условно называя их соответственно начальным и конечным полудифонами (переход от молчания к речи и наоборот). Поскольку глухие взрывные [к], [п], [т] и их мягкие аналоги в начале слова при сегментации не выделяются, мы используем также начальные полудифоны вида ка0, пи0, тл0 и так далее.

Анализ ситуации можно начать со следующего простого эксперимента. Используя какую-либо известную программу работы со звуком, например «Sound Forge», запишем два произвольных слова, а затем вырежем стационарные (серединные) части составляющих их звуков. Воспроизведя получившиеся звуковые сигналы, мы можем на слух определить, какие слова звучат. Напротив, вырезав межфонемные переходы и оставив стационарные части звуков, мы затруднимся на слух различить, например, слова «мама» и «лама».

Следующий аргумент относительно роли межфонемных переходов – использование при DTW-распознавании эталонов, полученных при удалении стационарных частей звуков, из которых состоят слова. Эксперименты показывают, что такое распознавание не менее успешно, чем распознавание по «полным эталонам». На рисунке 9 представлено окно программы.

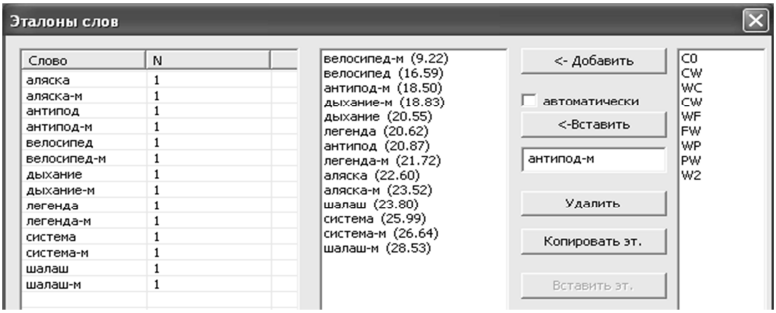

Рис. 9. Окно DTW-распознавателя с полными и урезанными сигналам.

В левом списке расположены слова, для которых построены эталоны по полному сигналу, и те же слова (они снабжены в конце символом –«м»), сигналы которых получены оставлением лишь межфонемных переходов (последнее достигается нажатием специальной кнопки). После этого эталоны строятся для таких урезанных сигналов. Если распознавать записанные сигналы без дополнительной обработки, то слова без «-м» распознаются стопроцентно. Если сигнал произнесенного слова урезать таким же образом, то стопроцентно распознаются слова с меткой «-м». Если оставить лишь урезанные эталоны, то слова с «-м» будут стопроцентно распознаваться и без урезания распознаваемого сигнала.

Можно пойти далее: создав эталоны всех дифонов, склеивать из них эталоны слов и распознавать слова по этим эталонам. Получится то, что естественно обозначить термином «распознавание через синтез». В результате процедура обучения системы распознавания произвольного словаря для конкретного диктора сведется к созданию базы дифонов. Количество дифонов равно квадрату от количества звуков. Для русского языка оно составляет около 1600. В п.12 показано, как организовать систему обучения, так чтобы его можно было выполнить за один сеанс работы. Подчеркнем еще раз, что создание дифонной базы в дальнейшем избавляет пользователя от необходимости создавать какие-либо эталоны голосом.

Спрашивается, зачем синтезировать из эталонов дифонов эталон слова? Не проще ли, распознавая дифоны между собой, получить их список, соответствующий слову, и по нему распознать слово? Нет, не проще. Дело в том, что дифонов достаточно мало для обучения, но слишком много для распознавания словаря этих дифонов, в особенности учитывая, что это короткие звуковые единицы, а DTW–

распознавание тем надежнее, чем длиннее и разнообразнее по составу распознаваемые объекты. Создание синтезированных эталонов слов – как раз шаг в этом направлении и он позволяет использовать DTW–распознавание целых слов со всеми его преимуществами.

Программа распознавания должна содержать модуль, который для каждого слова распознаваемого словаря создает его транскрипцию и по ней цепочку имен соответствующих дифонов, например,

остановка → астанофка → а0-ас-ст-та-ан-но-оф-фк-ка-а2.

Из эталонов дифонов склеивается эталон слова. Далее ведется DTW-распознавание слов по этим эталонам. Отметим, что в соответствии со сказанным выше относительно слов, начинающихся глухими взрывными звуками, для слова «палка», например, получится цепочка дифонов: па0-ал-лк-ка-а2. Все вектора, входящие в эталоны дифонов, играют роль кодовых векторов и образуют кодовую книгу *B*. Все эталоны дифонов нумеруются, нумеруются также все кодовые вектора.

При распознавании мы используем перечень эталонов слов словаря не в виде списка. Он реализуется в виде дерева дифонов, использование которого существенно ускоряет процесс распознавания. Дифоны представлены в дереве своими номерами. Эталон каждого слова представляется в виде отрезка ветви этого дерева. Если несколько ветвей имеют общую часть, то вычисления, заполняющие соответствующую часть DTW-матрицы, выполняются только один раз. Уровни дерева соответствуют позициям дифонов в слове. Каждый узел в рамках каждого уровня представляет собой номер дифона, находящегося в слове на соответствующей позиции. Вершины, соответствующие конечным дифонам слов, помечаются как концы соответствующих слов (в узле записывается порядковый номер соответствующего слова в словаре). Если узел не конечный, то записывается значение -1. Максимальная глубина дерева соответствует максимальной длине слова в словаре, выраженной в количестве дифонов.

Процесс распознавания строится следующим образом. Распознаваемое слово автоматически сегментируется и затем подвергается межфонемной обработке: удаляются стационарные части составляющих звуков и оставляются лишь дифоны в окрестностях межзвуковых меток. Затем создается представление слова в виде набора *N* векторов признаков и строится таблица *D* расстояний этих векторов до всех векторов кодовой книги *B*. Далее вычисляются DTW-

расстояния рассматриваемого слова до всех эталонов слов путем рекурсивного обхода дерева эталонов «в глубину». Вначале просматриваем начальный уровень, а затем спускаемся по ветви, пока не достигнем вершины, помеченной как конец слова. После того, как достигнут конец слова, возвращаемся назад вдоль пройденного пути пока не найдем вершину, у которой есть еще не посещенный сосед, а затем двигаемся в новом обнаруженном направлении. Процесс оказывается завершенным, когда мы вернулись в корень дерева, а все примыкающие к нему вершины уже оказались посещенными.

При прохождении ветвей дерева, по номерам дифонов строится цепочка соответствующих им номеров векторов, образующих эталон слова. Двигаясь в глубину, добавляем в цепочку номера, соответствующие пройденным узлам, а при движении назад они удаляются из нее. Достигнув узла, являющегося концом очередного слова, вычисляем DTW-расстояние от построенной цепочки векторов (эталона данного слова) до цепочки векторов распознаваемого сигнала. При этом расстояния между векторами берутся из таблицы *D*. В процессе вычисления расстояний матрица DTW не переписывается полностью, а обновляются только столбцы, соответствующие новым кодовым векторам, номера которых добавлены в цепочку после возврата назад по окончании предыдущего этапа. Дерево эталонов строится, когда словарь для распознавания загружается в программу в виде текстового списка.

Создание базы эталонов дифонов для конкретного диктора осуществляется с использованием программы, которая последовательно предлагает пользователю звукосочетания типа абавагада, ажазакала…, содержащие все дифоны. Записанное звукосочетание сегментируется, автоматически выделяются дифоны, создаются их эталоны и заносятся в базу. Создание полной базы для конкретного диктора требует около часа работы.

11. Заключение. При использовании дифонов появляется возможность различения слов, отличающихся звуками [б], [г], [д]. Появляется надежный способ распознавания слов, отличающихся звуками [к], [п], [т]. Пример – слова «папа», и «пата» (родительный падеж от шахматного термина «пат»). Становятся надежно различимыми парные сочетания с твердыми и мягкими согласными типа «мы-ми», «са-ся» и так далее. Хорошо различаются между собой также пары слов типа «кон-конь», «мол-моль» (отличие в твердости – мягкости в конце слова).

Предложенная методика DTW-распознавания по эталонам, построенным из эталонов дифонов, обеспечивает возможность распознавания сверхбольших словарей. Она тестировалась нами следующим образом. Из словаря А.А. Зализняка [22] случайным образом отбирались 30 тысяч слов (словарь для распознавания). Из них произносилось 100 произвольных слов. Число ошибочных распознаваний в самом неудачном случае равнялось 9.

Отметим, что при дифонном DTW-распознавании ошибки в сегментации в подавляющем большинстве случаев не приводят к ошибкам в распознавании. В наших распознавателях реализована также процедура дообучения: в случае ошибки пользователь указывает мышкой в списке или вводит с клавиатуры правильное слово; программа, сегментируя сигнал, создает эталоны прозвучавших дифонов и с их помощью модифицирует эталоны базы путем усреднения, которое было описано в конце раздела 9. Здесь ошибки в сегментации становятся важными, и программа, зная слово, в большинстве случаев исправляет их.

Отметим, что в этой статье мы не затрагивали важных вопросов о точном автоматическом определении начала и конца слитного речевого отрезка и проверке записанного на наличие речи (см. [17], [23–25]). Мы не касались также развиваемых нами подходов к распознаванию слитной речи (см. [17]).

Возможность автоматически создавать эталоны слов позволяет двигаться в направлении создания редактора с голосовым вводом, и постепенно пополняемым словарем, который позволяет с самого начала набирать произвольные тексты. При отсутствии в словаре нужного слова оно вводится с клавиатуры, и словарь автоматически пополняется его полной парадигмой, которая обеспечивает в дальнейшем возможность голосового ввода всех словоформ этого слова.

Литература

1. *Вишнякова О.А., Лавров Д.Н.* Автоматическая сегментация речевого сигнала на базе дискретного вейвлет-преобразования // Математические структуры и моделирование. 2011, вып. 23. С. 43–48.
2. *Жиляков Е.Г., Белов С.П., Белов А.С., Фирсова А.А., Глушак А.В.* Об эффективности различных подходов к сегментации речевых сигналов на основе обнаружения пауз // Научные ведомости Белгородского гос. ун-та. Серия История,…,Информатика, №7 (78), Выпуск14/1, 2010. С. 187-193.
3. *Кипяткова И.С., Карпов А.А.* Эксперименты по распознаванию слитной русской речи с использованием сверхбольшого словаря // Труды СПИИРАН, Вып. 12, 2010. С. 63–74.

4. *Конев А.А.* Модель и алгоритмы анализа и сегментации речевого сигнала: Кандидатская диссертация. Самара, 2007 г. 150 с.
5. *Мещеряков Р.В., Понизов А.Г.* Оценка качества слуха на основе мобильных вычислительных устройств // Труды СПИИРАН, Вып. 18, 2011. С. 93–107.
6. *Ручай А.Н.* Модифицированный метод сегментации речевого сигнала на основе непрерывного вейвлет-преобразования // Доклады ТУСУРа,№ 2 (26), часть 1, декабрь 2012. С. 189-192.
7. *Утробин В.А., Гай В.Е.* Алгоритм выделения вокализованных участков речевого сигнала // Вестник Нижегородского университета им. Н.И. Лобачевского, 2012, № 6 (1). С. 175–179.
8. *Цыплихин А.И.* Анализ и автоматическая сегментация речевого сигнала: Кандидатская диссертация. Москва, 2006 г. 149 с.
9. *Cherniz A., Torres M., Rufiner H., Esposito A.* Multiresolution Analysis applied to Text-Independent Phone Segmentation // Journal of Physics: Conference Series. 2007. Vol. 90, 012083.
10. *Greibus M., Telksnys L.* Rule Based Speech Signal Segmentation // Journal of telecommunications and information technology, 4/2010. P. 37-43.
11. *Heck M.* Segmentation of telephone speech based on speech and non-speech models // Specom, 2013.
12. *Hosom, J. P.* Automatic Phoneme Alignment Based on Acoustic-Phonetic Modeling // 2002 International Conference on Spoken Language Processing (ICSLP 2002), Boulder, Co., vol. I, Sep. 2002. P. 357-360.
13. *Petrushin V.A.* Adaptive Algorithms for Pitch-synchronous Speech Signal Segmentation // SPECOM'2004: 9th Conference, 2004.
14. *Rasanen O.J.* Speech Segmentation and Clustering Methods for a New Speech Recognition Architecture // Master's thesis, Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing, http://lib.tkk.fi/Dipl/2007/urn010123.pdf, 2007.
15. *Regine A.O.* A new statistical approach for the automatic segmentation of continuous speech signals // Acoustics, Speech and Signal Processing, IEEE Transactions. (Volume:36 , Issue: 1), 1988. P. 29-40
16. *Sarkar A., Sreenivas T.V.* Automatic speech segmentation using average level crossing rate information // Proc. ICASSP'05. 2005. Vol. 1. P. 397–400.
17. *Шелепов В.Ю., Ниценко А.В.* К проблеме распознавания слитной речи // Искусственный интеллект. № 4, 2012. С. 272-281.
18. *Шрюфер Е.* Обробка сигналів. Київ: Либідь,1992. 295 с.
19. *Шелепов В.Ю., Карабалаева М.Х., Ниценко А.В.* Обнаружение и выделение звука [р] в речевом сигнале. // Искусственный интеллект. № 1, 2011. С. 168-174.
20. *Винцюк Т.К.* Анализ, распознавание и интерпретация речевых сигналов. Киев: Наук. думка, 1987. 262 с.
21. *Дорохина Г.В.* Анализ методов распознавания речевых команд на основе алгоритма DTW // Труды шестого междисциплинарного семинара «Анализ разговорной русской речи», АРЗ-2012, 27-28 августа 2012, Санкт-Петербург. С. 29-34.
22. *Зализняк А.А.* Грамматический словарь русского языка. М.: Русский язык, 1977. 879 с.
23. *Федоров Е.Е., Шелепов В.Ю.* Защита речевых распознавателей от шума и посторонней речи // Искусственный интеллект. №3, 2001. С. 584-587.
24. *Федоров Е.Е., Шелепов В.Ю.* Автоматическое определение начала и конца записи речи // Искусственный интеллект. №4, 2002. С. 295-298.

25. *Шелепов В.Ю., Ниценко А.В.* Новый подход к определению границ речевого сигнала. Проблема конца сигнала. Москва. // Речевые технологии. №1, 2012. С. 74 – 78.

Бурибаева Айгерим Кеулимжаевна — Ph.D.; докторант Евразийского национального университета имени Л.Н. Гумилева (Астана, Казахстан). Область научных интересов: цифровая обработка сигналов автоматическое распознавание казахской речи. Число научных публикаций — 8. buribayeva@mail.ru. Научные руководители — А.А. Шарипбаев, В.Ю. Шелепов.

Buribajeva Aigerim Keulimzhajevna — Ph.D.; postdoc, Eurasian National Gumiljov University (Astana, Kazakhstan). Research interests: digital signal processing, kazakh speech recognition. The number of publications — 8. buribayeva@mail.ru. Scientific advisers — A. Sharipbajev, V.Ju. Shelepov.

Дорохина Галина Владимировна — исполняющая обязанности начальника отдела распознавания речевых образов, Институт проблем искусственного интеллекта НАН и МОН Украины. Область научных интересов: автоматическое распознавание речи, компьютерная лингвистика. Число научных публикаций — 30. sgv@iai.donetsk.ua. ул. Артема 118б, г. Донецк, 83048, Украина; тел.: 066-368-29-78. Научный руководитель — В.Ю. Шелепов.

Dorokhina Galina Vladimirovna — Acting Head, Speech recognition department, Institute of Artificial Intelligence. Research interests: automatic speech recognition, computational linguistics. The number of publications — 30. sgv@iai.donetsk.ua; Artyoma str. 118 b, Donetsk, 83048, Ukraine; office phone +38 (066)-368-29-78. Scientific advisor — V.Ju. Shelepov.

Ниценко Артем Владимирович — младший научный сотрудник отдела распознавания речевых образов, Институт проблем искусственного интеллекта НАН Украины и МОН Украины. Область научных интересов: цифровая обработка сигналов, автоматическое распознавание речи. Число научных публикаций — 20. nav_box@mail.ru; ул. Артема 118б, г. Донецк, 83048, Украина; тел. +38 (062) 311-34-24, факс +38 (062) 311-34-24. Научный руководитель — В.Ю. Шелепов.

Nitsenko Artem Vladimirovich — junior researcher, Speech recognition department, Institute of Artificial Intelligence. Research interests: digital signal processing, speech recognition. The number of publications — 20. nav_box@mail.ru; Artyoma str. 118 b, Donetsk, 83048, Ukraine; office phone +38 (062) 311-34-24, fax +38 (062) 311-34-24. Scientific advisor — V.Ju. Shelepov

Шелепов Владислав Юрьевич — д.ф.-м.н., проф.; зав. кафедрой систем искусственного интеллекта ДонНТУ, ведущий научный сотрудник Института проблем искусственного интеллекта НАН и МОН Украины. Область научных интересов: цифровая обработка сигналов, автоматическое распознавание речи. Число научных публикаций — более 100. vladislav.shelepov2012@yandex.ua; ул. Артема 118б, г. Донецк, 83048, Украина; тел.+3099-79-34-918, факс +38 (062) 311-34-24.

Shelepov Vladislav Jurievich — Ph.D., Dc.Sci., Prof.; head of the artificial intelligence systems department DonNTU; leading researcher, Institute of Artificial Intelligence. Research interests: digital signal processing, speech recognition. The number of publications – more than

100. vladislav.shelepov2012@yandex.ua; Artyoma str. 118 b, Donetsk, 83048, Ukraine; phone +3099-79-34-918, fax +38 (062) 311-34-24.

Рекомендовано лабораторией речевых и многомодальных интерфейсов, заведующий лабораторией Ронжин А.Л, д.т.н., доц.
Статья поступила в редакцию 10.10.2013.

РЕФЕРАТ

Бурибаева А.К., Дорохина Г.В., Ниценко А.В., Шелепов В.Ю.
Сегментация и дифонное распознавание речевых сигналов.

В статье описывается технология распознавания речи, основанная на использовании разработанного авторами механизма сегментации речевого сигнала, методике построения эталонов слов из дифонов и некоторой модификации метода DTW, дающей заметный выигрыш в скорости и объеме необходимой памяти.

Используются вектора признаков, связанные с относительными частотами длин полных колебаний. Эталоны слов распознаваемого словаря автоматически формируются из эталонов дифонов по автоматически же создаваемым транскрипциям слов. Полная база эталонов дифонов в объеме приблизительно 1600 создается для каждого диктора заранее. Это в дальнейшем избавляет пользователя от необходимости создавать эталоны каких-либо слов голосом.

Словарь эталонов слов реализуется в виде дерева дифонов. Дифоны представлены в дереве своими номерами. Эталон каждого слова представляется в виде ветви этого дерева. Если несколько ветвей имеют общую часть, то вычисления, заполняющие соответствующую часть DTW-матрицы, выполняются только один раз.

Предложенная методика DTW-распознавания по эталонам, построенным из эталонов дифонов, опробована на словарях, включающих до 30 тысяч слов, и показала надежность не менее 90%. Исследования также показали, что при дифонном DTW-распознавании ошибки в сегментации в подавляющем большинстве случаев не приводят к ошибкам в распознавании.

SUMMARY

Buribayeva A.K., Dorokhina G.V., Nitsenko A.V., Shelepov V.Ju. **Segmentation and diphone recognition of speech signals.**

The article describes the speech recognition technology, based on use of the speech signal segmentation mechanism which was developed by the authors, the technique of word patterns construction with diphones and modified dynamical time warping (DTW) algorithm. The last gives a significant gain in recognition speed and essential memory quantity.

The feature vector based on the relative frequencies of the oscillations lengths. Speech samples (etalons) for recognition vocabulary are formed from diphone samples of database, which contains approximately 1500 samples and created for each speaker. This approach allows in future to avoid the creation of any whole-word samples by voice.

All word samples are stored in memory as a hierarchical tree structure. Each diphone is represented in the tree with his index number, and each word sample is represented as a branch of the tree. If some branches have a common part, the calculation, filling the relevant part of DTW-matrix, are performed only once.

The offered recognition method, based on diphones, was tested on vocabularies containing up to 30000 words. The recognition accuracy not less than 90 %. Experiments show also that in the most cases speech segmentation errors do not decrease the recognition accuracy when diphone-based DTW-recognition used.