УДК 681.3

Грабовая В.А., Федоров Е.Е., Шелепов В.Ю.
Донецкий государственный институт искусственного интеллекта

О СИСТЕМЕ КОМПЬЮТЕРНОГО РАСПОЗНАВАНИЯ РУССКОЙ РЕЧИ С АВТОМАТИЧЕСКИМ ПОСТРОЕНИЕМ ЭТАЛОНОВ

В основе системы распознавания изолированных русских слов, описанной ниже, лежит оригинальная система признаков, используемая в отделе распознавания речи ИПИИ, – относительные частоты длин полных колебаний [1] и использование известного алгоритма DTW (Dynamic Time Warping) [2]. При частоте дискретизации 20050 Гц сигнал в 10 тысяч отсчетов разбивается на отрезки длиной 368 отсчетов (удвоенная длина квазипериода основного тона для мужского голоса средней высоты). На каждом отрезке вычисляется 29-ти мерный вектор признаков. Таким образом, слово представляется в виде набора 27-ми векторов. Такое представление строится и при распознавании, и в процессе обучения - при создании эталонов, когда каждому слову распознаваемого словаря ставится в соответствие эталон - набор 27-ми векторов.

В данной работе рассматривается система распознавания, которая построена на базе так называемой "пофонемной кодовой книги", то есть совокупности кодовых векторов, каждый из которых является усреднением множества векторов, отвечающих определенному аллофону [3].

Идеальный пофонемный распознаватель должен был бы представлять собой систему автоматической сегментации речевого сигнала, которая разбивает его на участки, отвечающие последовательности произносимых аллофонов. Затем каждый из этих участков должен был бы сравниваться с множеством эталонов аллофонов. Следовательно, база эталонов состояла бы лишь из упомянутых эталонов аллофонов. Однако, при отсутствии на сегодняшний день абсолютно надежной системы сегментации, можно пойти другим путем. Имея для распознавания текстовый файл произвольного размера, можно каждому слову заранее соотнести транскрипцию, описывающую аллофоны, подлежащие распознаванию. Если каждому из аллофонов транскрипции сопоставить соответствующий кодовый вектор, повторив его необходимое количество раз, получим эталон слова. Далее применяется обычная процедура распознавания слов по эталонам (алгоритм DTW и т.д.).

Данная работа предполагает именно этот компромиссный подход. При этом в описываемой программе осуществлено автоматическое построение пригодных для компьютера транскрипций и автоматическое построение эталонов. Система автоматической транскрипции (транскриптор) написана так, что она способна строить довольно подробную фонетическую транскрипцию в соответствии с законами русской фонетики [4,5]. Перечень используемых аллофонов можно восстановить, пользуясь соответствующими колонками табл. 1. Затем, в соответствии с возможностями системы распознавания, производится отождествление ряда аллофонов и, в результате, используется полученная адаптированная транскрипция.

Обучение системы для конкретного диктора представляет собой наговаривание "пофонемной кодовой книги" [3]. Это процесс несравненно более короткий и менее трудоемкий, чем наговаривание эталона каждого слова.

РЕЗУЛЬТАТЫ АДАПТАЦИИ ТРАНСКРИПТОРА К ВОЗМОЖНОСТЯМ РАСПОЗНАВАТЕЛЯ

В ходе экспериментов и исследования амплитудно-временного представления (АВП) были выявлены следующие группы аллофонов, которые при наших методах распознавания приходится отождествлять:

а) [б], [г], [д] - твердые звонкие смычные взрывные;

б) [б'], [г'], [д'] - мягкие звонкие смычные взрывные;

в) [п], [к],[т], [п'], [к'] - глухие смычные взрывные.

Это аллофоны, выдержка которых включает два момента: во-первых, органы речи образуют полную смычку (в случае звонких звуков голосовые связки работают); во-вторых, струя воздуха ее прорывает. Это хорошо заметно на АВП этих звуков (рис. 1, 2). На рис. 1 представлены АВП для следующих слогов: да, ба, га. Как видно, общая длительность звука (б, г или д) состоит из длительности тональной смычки и длительности значащей части звука. Аналогичная ситуация наблюдается на рис. 2 при произнесении слогов: ата, апа, ака.

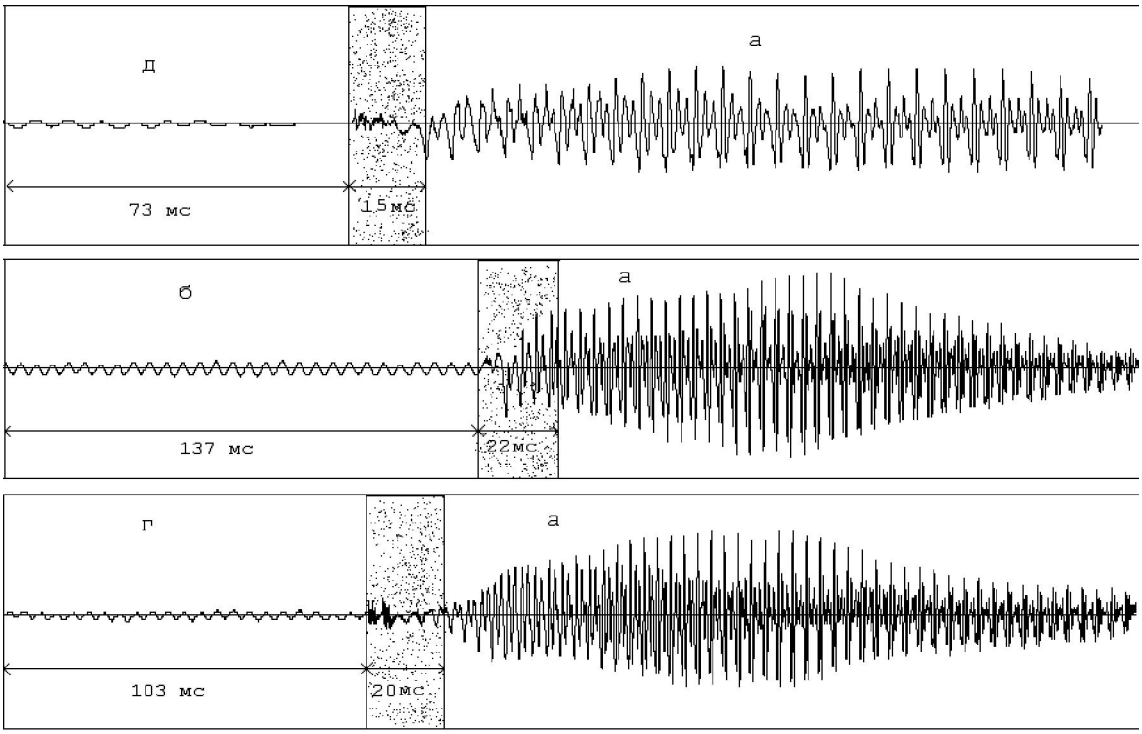

Рис. 1. АВП слогов да, ба, га.

Учитывая, что при построении кодового вектора используется, как минимум, два однородных отрезка длительностью по 368-м отсчетов, получаем, что различение вышеприведенных аллофонов в принятых нами подходах не представляется возможным. Участок, отличающий их, слишком мал и неоднороден, чтобы существенно повлиять на построение кодовых векторов, соответствующих данным аллофонам, а также на эталон слова в целом. Поэтому принято решение отождествить аллофоны указанных групп.

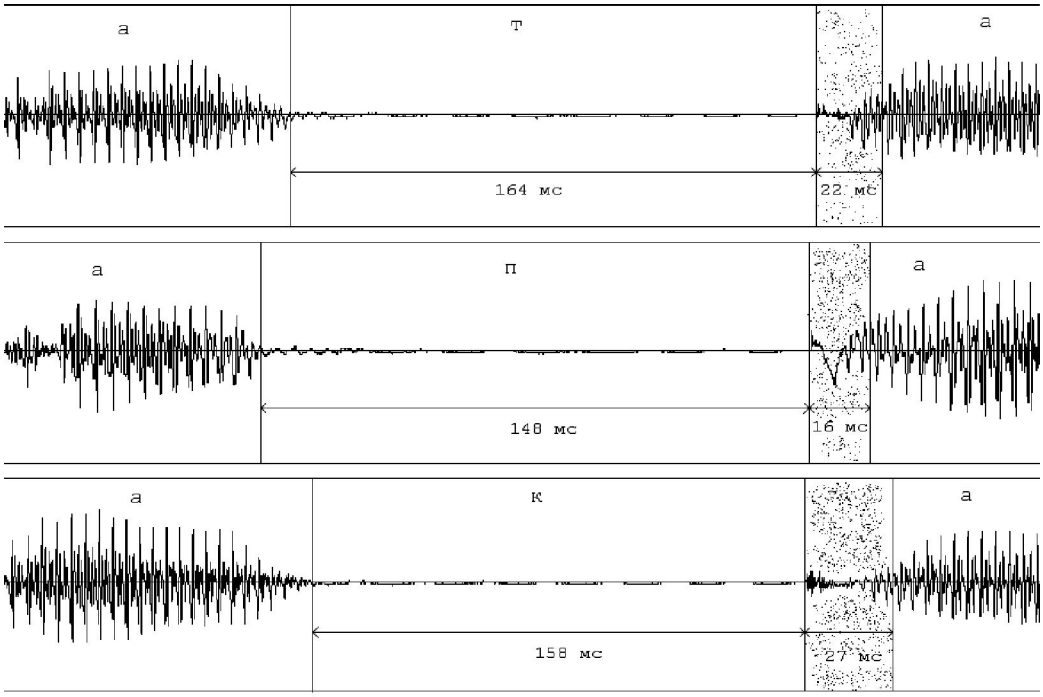

Рис. 2. АВП слогов ата, апа, ака.

При анализе АВП слова шесть (рис.3) установлено, что участок сигнала, соответствующий отличной от паузы части аллофона [т'], имеет достаточную протяженность для того, чтобы использовать его при построении кодового вектора. Исследования показали, что это явление распространяется на фонему <т'> в том случае, если она находится в середине или в конце слова. Поэтому звук " т' " в середине и в конце слова в транскрипции обозначается символом [-т'], где "-" обозначает паузу перед значащей частью звука.

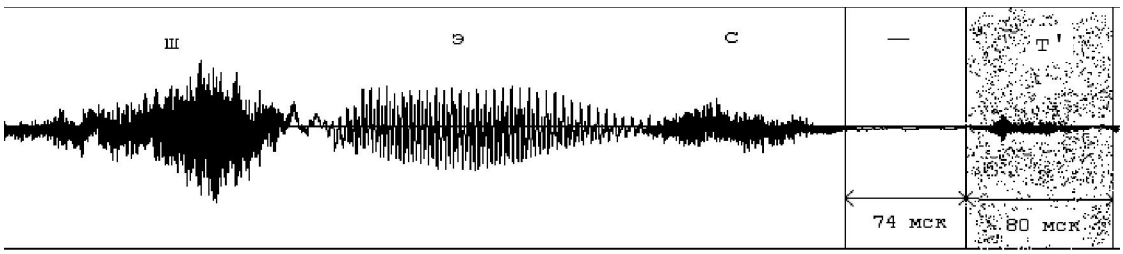

Рис. 3. АВП слова шесть.

Особую группу смычных согласных составляют аффрикаты-фонемы <ц> и <ч'>. Они характеризуются тем, что в экскурсии эти согласные имеют смычку, а в рекурсии – щель. Под экскурсией понимается переход органов речи к положению, необходимому для произношения данного звука. Рекурсия – это положение органов речи в тот момент, когда после прохождения струи воздуха они возвращаются в нерабочее состояние или переходят к артикуляции другого звука. Переход от смычки к щели происходит незаметно, и эту границу человеку установить невозможно. Однако на АВП сигнала (рис. 4, 5) можно точно выделить участок РС, соответствующий смычке (пауза) и участок, соответствующий звучащей части <ч'> или <ц>.

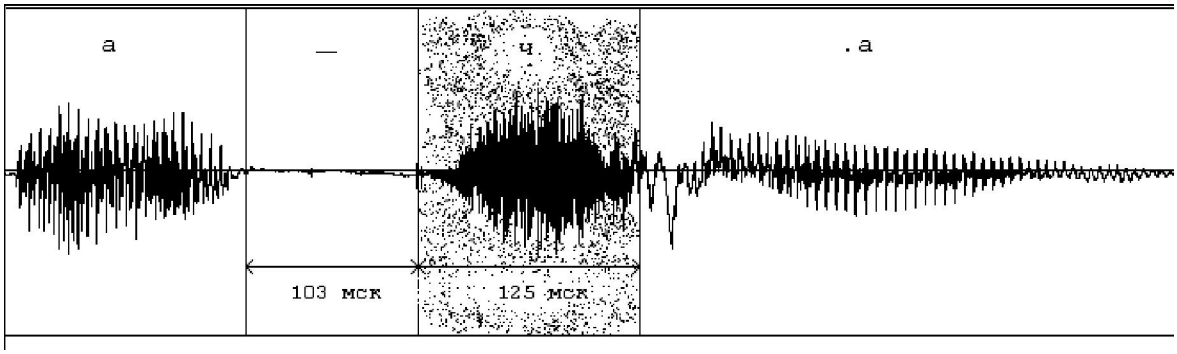

Рис. 4. АВП слога ача.

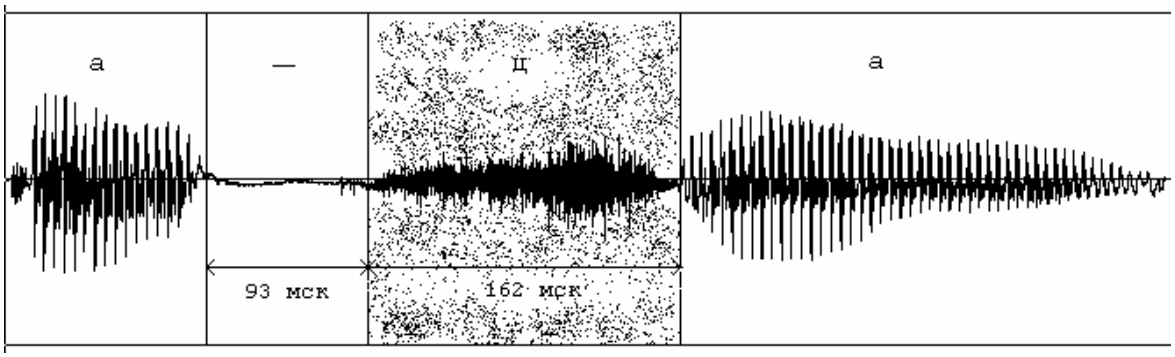

Рис. 5. АВП слога аца.

На основании полученных результатов в транскрипции перед звуками "ц" и "ч' " вставляется пауза: [-ц] и [-ч'].

Итак, начальный алфавит, используемый для построения автоматического транскриптора, претерпевает следующие изменения:

а) аллофоны [т], [п], [к], [п'], [к'] объединяются в одну группу, которую будем обозначать [-] (пауза), а звук "т' " описывается транскрипцией [-т'];

б) аллофоны [б], [г] и [д] отождествляем и обозначаем через [д], аллофоны [б'], [г'], [д'] отождествляем и обозначаем через [д'];

в) звуки " ч' " и "ц" имеют такие транскрипционные описания: [-ц] и [-ч'];

Для удобства работы введем свои обозначения для аллофонов, применительно к возможностям стандартной клавиатуры. Соотношение знаков русской транскрипции и знаков, принятых в данной работе, сведем в табл. 1. В русском языке если перед гласным и после него стоят мягкие согласные, то качество гласного изменяется значительно; это обозначается двумя точками. Если мягкий согласный только спереди или только сзади (он влияет менее значительно), то используем точку слева или справа соответственно [5].

Таким образом, транскрипция представляется как цепочка из перечисленных знаков. Процедура автоматического построения эталонов описана выше.

О ПРОБЛЕМЕ ДЛИТЕЛЬНОСТИ АЛЛОФОНОВ

В описываемой работе в качестве меры близости между объектами использовалось DTW-расстояние, определяемое при помощи DTW-алгоритма [1]. Целесообраз-

ность применения такого метода заключается в том, что DTW-алгоритм обеспечивает выравнивание различных по длительности акустически подобных кусков сигнала и далее производит сравнение.

Табл. 1. Соотношение знаков русской транскрипции и знаков, принятых в данной работе.

Знаки транскрипции			
рус. яз.	проекта	рус. яз.	проекта
у′	y	и′	и
ы′	ы	э′	э
Q	q	и^{э}	i
ы^{э}	ъ	у	u
.а, .а .	я	.о, .о.	ё
.э, .э.	е	.у, .у.	ю
а .	a	о.	o
э.	э	у.	y
с	c	с'	s
з	з	з'	z
п, т, к	-	п', к'	-
б, г, д	д	б', г', д'	d
в	в	в'	v
ф	ф	ф'	f
ш	ш	ш:'	щ
х	х	х'	h
-ц	-ц	ч'	-ч
ж	ж	т'	-t
р	р	р'	r
л	л	л'	l
м	м	н'	n
н	н	м'	m
j	j		

Табл. 2. Данные о средней длительности согласных звуков

Аллофон	Средняя длительность аллофона		
	Начало	Середина	Конец
-	0	4	4
f	0	7	6
j	6	4	5
l	4	4	7
m	4	4	6
n	4	4	6
r	5	4	4
s	7	5	7
t	0	4	4
v	4	4	6
z	5	7	-
в	4	4	-
д	4	4	-
ж	4	4	7
з	7	6	-
л	4	5	4
м	4	4	5
н	4	4	5
р	4	5	5
с	6	7	6
ф	0	8	8
х	4	7	8
ц	5	5	7
ч	4	4	6
ш	5	7	8
щ	8	9	9

Однако в ходе экспериментов было установлено, что при построении эталона из кодовых векторов варьирование длительности аллофонов допустимо лишь в весьма ограниченных пределах, превышение которых приводит к резкому ухудшению качества распознавания. Иначе говоря, эталон "работает" лишь при определенных значениях длин и близких к ним значениях. В связи с этим, было принято решение о том, чтобы каждому аллофону слова была приписана его средняя длительность в данной позиции (количество отрезков по 368 отсчетов, содержащихся в соответствующем участке речевого сигнала). Тогда. транскрипция будет иметь следующий вид: <транскрипция>::={<аллофон><длительность>}, где: <длительность> - количество вышеупомянутых отрезков речевого сигнала, приходящееся на аллофон.

Например, для слова Саша транскрипция будет следующей: [с6а10ш7q4].

Была разработана программа, определяющая среднюю длительность каждого аллофона в различных позициях. Для согласного звука длительность подсчитывалась в зависимости от позиции в слове: длительность в начале слова, длительность в середине слова, длительность в конце слова (табл. 2).

Аллофоны [а], [я], [о], [ё], [ы], [у], [ю], [э], [е] имеют среднюю длительность 10, [и] – 6, [q], [i] – 4, [ъ], [u] – 5 отрезков по 368-м отсчетов.

В результате использования при построении эталонов слов информации о длительности аллофонов качество распознавания значительно улучшилось.

Процесс распознавания в описываемой системе строится следующим образом. Распознаваемое слово записывается в виде набора 27-ми произвольных (некодовых) векторов [1]. Затем строится таблица расстояний этих векторов до всех векторов "кодовой книги". Далее вычисляются DTW-расстояния рассматриваемого слова до всех эталонов и за результат распознавания принимается слово, эталон которого оказывается ближайшим.

При этом расстояния между векторами берутся из упомянутой таблицы, а не вычисляются каждый раз, как это было, когда не использовалась "кодовая книга". Это требует значительно меньше времени. Таким образом, достигается значительный выигрыш как в скорости распознавания, так и в объеме необходимой памяти. Для пользователя отпадает необходимость в процессе обучения системы наговаривать каждое из слов распознаваемого словаря, достаточно лишь создание кодовой книги.

ЛИТЕРАТУРА

1. Дорохин О.А., Засыпкин А.В., Червин Н.А., Шелепов В.Ю. О некоторых подходах к проблеме компьютерного распознавания устной русской речи // Международная конференция "Знания Диалог Решение". Сборник научных трудов. - Ялта, 1997.- Т.1.- С.234-240.
2. L.Rabiner, B.H. Juang. Fundamentals of Speech Recognition. Prentice Hall PTR, 1993, 507 p.
3. Дорохин О.А., Федоров Е.Е., Шелепов В.Ю. Некоторые подходы к пофонемному распознаванию русской речи и распознаванию больших словарей // Искусственный интеллект. - Донецк, 1999. - № 2. - С. 329-333.
4. Панов М.В. Современный русский язык. Фонетика.-М.: Высш. школа, 1979.-256 с.
5. Аванесов Р.И. Фонетика современного русского литературного языка.-М.: Изд-во Московского государственного университета, 1956.- 240с.