УДК 004.89:004.93

О возможностях алгоритма DTW при распознавании речевых сигналов

А.В. Ниценко, В.Ю. Шелепов,
Институт проблем искусственного интеллекта
nav_box@mail.ru

Ниценко А.В.,Шелепов В.Ю. О возможностях алгоритма DTW при распознавании речевых сигналов.*В статье рассмотрены варианты применения для различных задач метода распознавания на основе алгоритма DTW, использующего фонемную сегментацию и эталоны слов, автоматически синтезируемые из эталонов дифонов. Метод отличается тем, что позволяет применять DTW -распознавание к словарям большого объема (более десяти тысяч слов), а также для распознавания слитной речи. Работа ведется с произвольными словарями, задаваемыми в текстовом виде, без необходимости выполнения предварительного обучения эталонов всех слов.*

Ключевые слова: *распознавание речи, динамическое программирование, алгоритм DTW, сегментация, дифон, антиэталон.*

Введение

Одной из актуальных проблем искусственного интеллекта является проблема преобразования устной речи в текст. За многие годы исследований был разработан широкий спектр методов и компьютерных программ, направленных на решение этой проблемы. Сегодня получены многообещающие результаты и созданы действующие коммерческие системы, в основном, для английского языка, а также испанского, французского, японского, китайского и арабских языков [1-8]. В числе последних достижений появились достаточно успешно работающие системы распознавания слитной речи с большими словарями, например голосовой ввод в поисковых интернет-системах Google и Yandex [9]. Однако их применение связано с работой в сети Internet и использованием облачных технологий. Проблема же распознавания речи на локальных устройствах остается открытой.

Одним из первых методов распознавания, которому уделяется внимание и сейчас, является метод динамической трансформации временной шкалы (DTW)[10,11], позволяющий найти оптимальное соответствие между двумя временными последовательностями. Сравнение DTW с другими методами показывает, что при распознавании команд небольшого словаря DTW даёт лучшие результаты. Первоначально DTW был доминирующей парадигмой при распознавании речи. Однако впоследствии предпочтение стали отдавать методам, основанным на использовании Скрытых Марковских Моделей, ссылаясь на невозможность использования DTW для распознавания отдельных слов при большом словаре и распознавания слитной речи [11,12].

В связи с этими проблемами в рамках DTW-парадигмы авторами предложен синтез эталонов слов из более мелких единиц – эталонов дифонов, на основе создаваемой в процессе обучения дифонной базы эталонов [13,14]. Опыт показывает, что использование дифонов, которые содержат межфонемные переходы, дает лучшие результаты по сравнению с использованием стационарных частей звуков речи. Также преимуществом является простота процедуры начального обучения, в которой разметка речевого сигнала и создание базы исходных эталонов дифонов, а затем и эталонов слов, являются полностью автоматизированными и не требуют участия эксперта. Цель настоящей обзорной статьи – показать, что предложенный подход позволяет распознавать большие словари отдельных слов и слитную речь.

Усовершенствованный метод распознавания речевых сигналов на основе алгоритма DTW

Используется 8-битная запись с частотой дикретизации 22050 Гц. Используются вектора признаков, связанные с относительными частотами длин полных колебаний на интервалах анализа длиной 368 отсчетов (удвоенный период основного тона для голоса средней высоты). Модификация классического метода распознавания с помощью алгоритма DTW заключается в использовании для распознавания эталонов слов, которые автоматически синтезируются из эталонов дифонов. Полная база последних в объеме около 1700 создается для каждого диктора заранее. Создание такой базы в дальнейшем избавляет пользователя от необходимости создавать любые образцы голосом и дает возможность применять этот метод для систем распознавания с большим объемом словаря.

Под дифоном, который соответствует

межфонемному переходу внутри слова, будем понимать участок стандартной длины: 3 окна в 368 отсчетов слева от метки между звуками и 3 таких же окна справа от той же метки. Эталон дифона – набор из шести соответствующих векторов признаков. Кроме того, используется участок в 3 окна в начале слова и участок в 3 окна в конце слова, условно называемый соответственно начальным и конечным полудифонами (переход от молчания к речи и наоборот). Все вектора, входящие в эталоны дифонов, играют роль кодовых векторов и образуют кодовую книгу. Все эталоны дифонов нумеруются, нумеруются также все вектора признаков, входящие в эталоны.

Авторами разработан простой автоматический транскриптор, управляющий файл которого содержит набор правил, каждое из которых записано в виде двух частей, соединенных знаком равенства. Слева стоят символы буквенной записи, справа – символы, которыми они заменяются в транскрипции. Машина, транскрибируя слово, последовательно ищет вхождения левой части очередного правила, и если таковое обнаруживается, заменяет его правой частью.

Каждое слово из словаря распознавания автоматически транскрибируется, по транскрипции строится цепочка имен дифонов, например,

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

В соответствии с этой цепочкой из эталонов дифонов склеивается эталон слова. Далее для простоты будем говорить о словах, хотя более точно нужно говорить о транскрипциях.

Словарь эталонов слов компактно представляется в виде дерева, использование которого существенно ускоряет процесс распознавания. Эталон каждого слова представляется в виде ветви этого дерева или ее части. Если несколько путей имеют общую часть, то вычисления, заполняющие соответствующую часть DTW-матрицы, выполняются только один раз. Уровни дерева соответствуют позициям дифонов в слове (рисунок 1).

Каждый узел в рамках каждого уровня представляет собой номер дифона, что находится в слове на соответствующей позиции. Узлы, соответствующие конечным дифонам слов, обозначаются 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 пути (выраженной в количестве дифонов) для соответствующего слова в словаре.

Процесс распознавания организован следующим образом. Исследуемый речевой сигнал преобразовывается в последовательность N векторов признаков и строится таблица D расстояний этих векторов до всех векторов эталонов. Далее вычисляются DTW-расстояния от распознаваемого слова до всех эталонов слов путем рекурсивного обхода дерева эталонов «в глубину». Сначала просматривается корень дерева, а затем смежные узлы вглубь дерева, пока не достигнут узел, помеченный как конец слова. После того, как достигнут конец слова, происходит возврат назад вдоль пройденного пути пока не найден узел, у которого есть еще не просмотренный потомок. Затем движемся в новом обнаруженном направлении. Процесс завершается, когда просмотрены все узлы дерева.

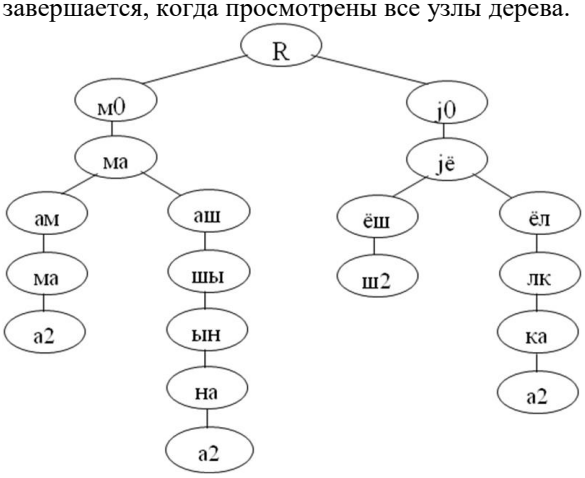

Рисунок 1 – Схема дерева синтеза эталонов для простого словаря

При прохождении ветвей дерева, по номерам дифонов строится цепочка соответствующих им номеров векторов, образующих эталон слова. При движении в глубину, в цепочку добавляются номера, соответствующие пройденным узлам, а при движении назад они удаляются из нее. Достигнув узла, являющегося концом очередного слова, вычисляется DTW-расстояние от построенной цепочки векторов (эталона данного слова) до цепочки векторов распознаваемого сигнала. Расстояния между векторами берутся из таблицы D. В процессе вычисления расстояний матрица DTW не пересчитывается полностью, а обновляются только столбцы, соответствующие новым кодовым векторам, номера которых добавлены в цепочку после возврата назад по окончании предыдущего этапа.

Таким образом, достигается значительный выигрыш, как в скорости распознавания, так и в объеме необходимой памяти. Дерево эталонов строится при загрузке в

программу словаря для распознавания в виде текстового списка.

Применение метода к распознаванию изолированных слов по частям

Распознавание между собой словоформ одного и того же слова представляет более трудную задачу, чем распознавание словоформ различных слов. Это вызвано тем, что они, как правило, отличаются окончаниями, которые чаще всего безударны и редуцируются при произношении. С целью увеличения надежности распознавания словоформ предлагается начинать распознавание с распознавания окончаний. В пользу этого можно привести следующие качественные соображения. Две словоформы достаточно длинного слова имеют общую основу и, следовательно, имеют больше общего, чем различий, что может служить источником ошибок. Если же ограничиться распознаванием одних окончаний, то их отличия относительно больше, чем отличия полных словоформ. Поэтому ошибки в их распознавании должны быть менее частыми. С другой стороны качество распознавания с помощью DTW возрастает при увеличении длины распознаваемых речевых отрезков. Поэтому целесообразно присоединять к окончанию часть суффикса, и работать с этими объектами, которые естественно назвать квазифлексиями. Соответственно оставшуюся часть слова будем называть квазиосновой.

Использование квазифлексий приводит также к сокращению размеров распознаваемых словарей. Квазифлексии, очевидно, являются общими для больших групп слов. Если имеется m квазиоснов и n квазифлексий, то их комбинации образуют $m \times n$ словоформ и, при распознавании словоформы как целого, словарь для распознавания составил бы $m \times n$ объектов. При распознавании же квазиосновы и квазифлексии отдельно, количество распознаваемых объектов составляет $m+n$. В результате время распознавания значительно сокращается, а надежность распознавания имеет тенденцию к увеличению.

Итак, для решения указанных проблем предлагается распознавать словоформы в два этапа: вначале распознавая изменяющуюся часть слова (квазифлексию), затем неизменяющуюся часть (квазиоснову) из множества, соответствующего распознанной квазифлексии [15]. Введенное понятие квазиосновы родственно используемому в лингвистике понятию основы слова, которая при простейшем описании определяется как его неизменяемая часть (приставка+корень+суффикс), то есть является результатом отбрасывания окончания. Понятие квазиосновы введено потому, что распознавание тем надежнее, чем длиннее распознаваемые речевые отрезки. Поэтому короткие словоформы (состоящие менее чем из 5 звуков) включаются в число квазиоснов целиком.

Исходя из того, что русский язык является флективным языком (т.е. синтаксическое управление осуществляется с использованием словоформ, образуемых при помощи флексий), слова языка моделируются в виде комбинации постоянной и переменной составляющих:

$x = c(x)\&f(x)$,

где c(x) – часть лексемы x, которая в процессе словоизменения остается неизменной (квазиоснова), f(x) – ее переменная составляющая (квазифлексия), & – знак конкатенации.

Так как распознавание будет вестись с использованием эталонов дифонов, то для каждой квазиосновы и квазифлексии используется транскрипция и по ней создается цепочка соответствующих дифонов. Например, для слова *«вокализация»*:

вокализа → вакаlиза → в0-ва-ак-ка-al-lи-из-за-а2

ция → цыjя → ц0-цы-ыj-jя-я2.

Для распознавания применяется алгоритм на основе DTW, описанный выше.

Определение квазифлексии производится по принципу минимума DTW-расстояния [15] (см. также следующий раздел). Он заключается в последовательном распознавании с помощью алгоритма DTW заключительных частей речевого сигнала, начиная с двух конечных сегментов: вначале берется два последних сегмента, затем три, четыре и так далее до задаваемого заранее максимального количества фонетических сегментов. При этом запоминается минимальное значение DTW-расстояния среди всех эталонов и соответствующая этому эталону квазифлексия, и далее производится сравнение эталонов со следующим участком сигнала. Таким образом, получается список гипотетических квазифлексий и DTW-расстояний от их эталонов до рассматриваемых речевых отрезков. Из этого списка выбирается квазифлексия с наименьшим расстоянием. Затем для выделенной таким образом квазифлексии происходит обращение к словарю соответствующих квазиоснов, и в пределах этого словаря осуществляется DTW-распознавание участка сигнала от начала слова до начала участка, соответствующего распознанной квазифлексии.

Применение алгоритма DTW к распознаванию слитно произнесенных фраз

Пусть имеется несколько слитно произнесенных фраз. Программа автоматически транскрибирует их и создаст для каждой из них

эталон из дифонов, игнорируя пробелы между словами. После этого их можно распознавать между собой теми же методами, что и отдельно произносимые слова. Но если рассматривать множество произвольных фраз, то их бесконечно много и, очевидно, следует добиваться их распознавания путем распознавания слов, из которых они состоят. Тогда основная сложность – выделение в речевом сигнале отрезков, отвечающих отдельным словам. Иначе говоря, необходимо научиться определять, где заканчивается одно слово и начинается другое.

Предлагаемый ниже метод основан на использовании фонемной сегментации [13,14]. Весь рассматриваемый речевой отрезок автоматически разбивается на сегменты, отвечающие отдельным звукам, и границы между словами следует искать среди конечного множества полученных границ между звуками.

Первоначально было проведено исследование распознавания пар слитно произносимых слов. При распознавании отрезка от начала до первой метки, а затем от первой метки до конца, как результат, получалась пара слов из словаря распознавания. Затем выполнялось распознавание от начала до второй метки и от второй метки до конца и так далее. Заключительным шагом было распознавание всего речевого отрезка от начала до конца как одного слова. В результате получалась последовательность гипотетических пар слов (на последнем месте – одно слово). Для каждой из этих пар автоматически строился эталон как для слитно произносимой, и результатом распознавания объявлялась пара, до которой DTW-расстояние минимально. Этот алгоритм показал высокую надежность. Но он включал целый набор актов распознавания отдельных гипотетических слов и в результате оказывался слишком долго работающим. Попытка применить аналогичный алгоритм к распознаванию большего числа слитно произнесенных слов приводила к экспоненциальному росту числа распознаваний гипотетических слов, и от нее пришлось отказаться.

Тогда было решено, двигаясь от начала до очередной метки, выводить только последовательность гипотез для первого слова, но с указанием DTW-расстояния до каждой из них [16,17]. Оказалось, что при выполнении некоторого условия на состав словаря гипотеза, соответствующая истинному первому слову (и соответствующему истинному отрезку от начала) имеет минимальное расстояние.

На рисунке 3 показан результат распознавания слитно произнесенной фразы *«доза мала»*. В левом верхнем поле находится список результатов распознавания последовательных отрезков сигнала, выделенных на рисунке 2. В списке выделена строка, отвечающая истинному отрезку слова *«доза»*.

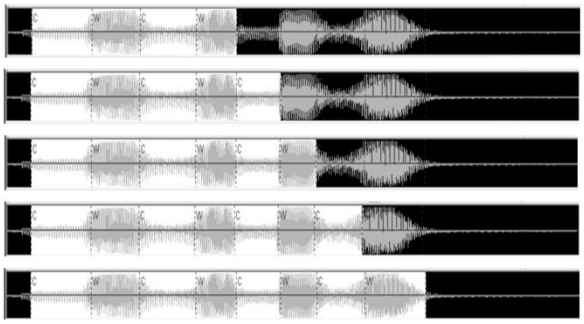

Рисунок 2 – Последовательное выделение распознаваемых отрезков сигнала с использованием сегментации

Итак, можно сформулировать следующий «принцип минимума»: по крайней мере, для словарей, удовлетворяющих некоторому ограничению, первое слово определяется с использованием меток сегментации из условия минимума DTW-расстояния. Понятно, что для распознавания второго слова фразы следует применить описанный метод к части сигнала от конца первого слова до конца речевого отрезка и так далее. Смысл этого принципа в следующем: DTW направлен на минимизацию расстояния сказанного слова до эталона того же слова. Остальные слова в полученном списке не соответствуют выделенным отрезкам речевого сигнала и то, что их расстояния до соответствующих эталонов оказались больше, представляется естественным.

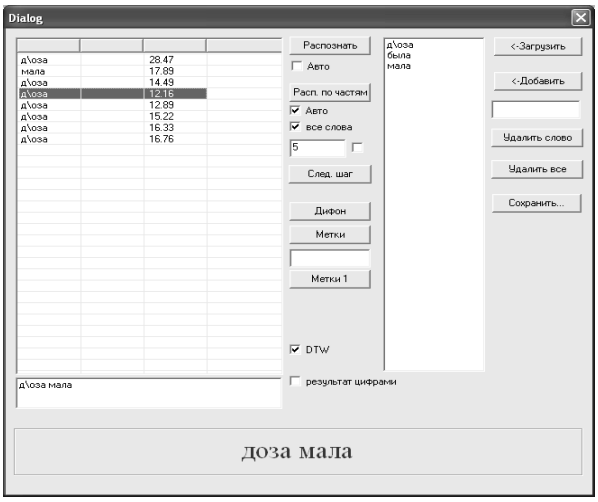

Рисунок 3 – Результат распознавания фразы *«доза мала»*

Распознавание слитных русских фраз на основе использования глухих фрагментов

Предлагаемый алгоритм опирается на принадлежащий авторам метод сегментации, то есть автоматического разбиения сигнала на участки, отвечающие отдельным звукам русской речи, с одновременной классификацией этих участков в рамках широкой фонетической классификации (W – гласный звук, C – звонкий согласный, F – глухой фрикативный, P – глухой взрывной) [13-14]. При распознавании используется описанный выше принцип минимума DTW-расстояния.

Предлагается метод распознавания слитных фраз с использованием частичных списков слов, формируемых из общего словаря на основании количества глухих фрагментов [18]. Для каждого слова словаря распознавания по автоматически создаваемой транскрипции определяется количество глухих фрагментов в его звучании. Все слова с n глухими фрагментами помещаются в текстовый файл словаря n.txt, при этом возникают словари 0.txt, 1.txt,…, N.txt. Предположим, что распознаваемый сигнал не начинается с глухого фрагмента. Если первое слово фразы вообще не содержит глухих фрагментов, его следует распознавать в словаре 0.txt, причем искать на отрезке от начала фразы до начала первого глухого фрагмента. Будем это делать, последовательно увеличивая интервал распознавания: от начала до первой метки сегментации, затем от начала до второй метки и так далее, до тех пор, пока не дойдем до левой границы первого глухого фрагмента. Далее продолжаем искать первое слово в словаре 1.txt, последовательно добавляя к интервалу распознавания, начинающемуся в начале сигнала, отрезки сегментации, пока не дойдем до начала второго глухого фрагмента. И так далее. Формируется список результатов всех этих распознаваний с указанием DTW-расстояний до них. В этом списке выбирается строка, где упомянутое DTW-расстояние минимально (рисунок 4). Этим заканчивается первый цикл распознавания. Его результат – распознавание первого слова и места его окончания. Далее, начиная с этого места, распознается второе слово фразы и определяется его конец и так далее. Ясно, что, если распознаваемый сигнал начинается с глухого фрагмента, то распознавание следует начинать со словаря 1.txt. В описанном методе записанный сигнал в значительной степени управляет выбором частичных списков для распознавания, формируемых из общего словаря.

Если слово из словаря n.txt заканчивается глухим звуком, а следующее начинается одним из звуков Б,Г,Д,Ж,З, то в слитной речи упомянутый глухой озвончается. Поэтому к словарю n.txt добавляется словарь n-1,V.txt с теми же словами, для которых при создании транскрипций используется модифицированный транскриптор, заменяющий глухие звуки в конце слов парными звонкими. Число глухих фрагментов в словах из n-1,V.txt на единицу меньше, чем в словах из n.txt. Далее, в каждом из словарей m.txt оставляются те слова, которые начинаются не на Б,Г,Д,Ж,З, а из остальных образуется словарь mD.txt. Алгоритм должен на соответствующих этапах вслед за словарем n-1.txt проработать со словарем n-1,V.txt, а вслед за словарем m.txt проработать со словарем mD.txt. Правильный результат обеспечивается принципом минимума DTW-расстояния. Возможен случай, когда фраза произнесена так, что озвончения в положенном месте не произошло. Тогда соответствующее распознанное слово найдется не в словаре n-1,v.txt, а в словаре n.txt.

Простейшим подтверждением работоспособности предлагаемого метода является разработанная нами программа, которая позволяет вводить в текстовом виде через специальное окно несколько фраз, автоматически создает текстовые файлы словарей и далее позволяет распознавать слитную речь в пределах общего образовавшегося словаря, произвольно меняя слова и их порядок. Здесь результат, как правило, стопроцентный. В качестве экспериментов со словарями большего объема мы реализовали распознавание пар ПРИЛАГАТЕЛЬНОЕ-СУЩЕСТВИТЕЛЬНОЕ, где фигурируют несколько сот существительных и прилагательных из тысячи наиболее частотных, приведенных в частотном словаре русского языках [19]. В этих экспериментах результат верного распознавания свыше 90%.

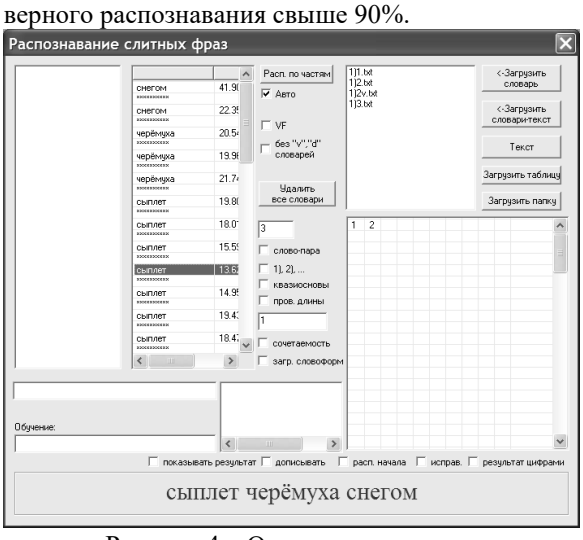

Рисунок 4 – Окно программы со списком гипотез и результатом распознавания

Распознавание слитно произносимых имен и отчеств

Рассматривается 150 русских мужских имен и столько же соответствующих отчеств, а также 88 женских имен. Общее количество имен: 238. Общее количество имен-отчеств: 35700. Их также можно было бы распознавать целиком, не членя на составляющие слова. Однако количество их уже достаточно велико, так что такой способ не обеспечивает достаточной скорости распознавания. С целью увеличения скорости (а также повышения надежности) применяется автоматическое разбиение словаря фраз на части путем использования VF-транскрипции. Эта транскрипция представляет собой последовательность чередующихся символов V и F, где V обозначает (максимальный) отрезок из соседствующих голосовых (гласных и звонких согласных) звуков, F – аналогичный отрезок из глухих (фрикативных и взрывных) звуков. Такая обобщенная транскрипция ранее использовалась нами для ускорения распознавания больших словарей отдельно произносимых слов. Она легко получалась автоматически из полной транскрипции и помещалась в конечных вершинах дерева транскрипций, соответствующих словам словаря. В рассматриваемой ситуации для каждой из фраз словаря «Имя-Отчество» также заранее создается транскрипция и VF-транскрипция. Затем автоматически формируются частичные списки фраз с одинаковой VF-транскрипцией и именами соответствующих файлов вида VmV.txt, VnF.txt, FkV.txt, FlF.txt. Символы V, F в начале и в конце имени показывают, с чего начинается и чем заканчивается VF-транскрипция, m, n, k, l – число символов F, входящих в VF-транскрипцию.

При записи фразы и ее априорной сегментации сразу же определяется ее VF-транскрипция и далее ведется распознавание в соответствующем частичном списке. Размеры этих списков за исключением файла V3F.txt таковы, что входящие в них фразы можно распознавать как целое с помощью обычного дифонного распознавания. Список V3F содержит 6923 фразы и из-за его величины распознавание в нем происходит достаточно медленно. Для ускорения мы применяем алгоритмы определения начального звука [20].

Тестирование распознавателя для диктора, который использует свою дифонную базу, дает не более 5% ошибочных распознаваний.

Поиск ключевых слов в слитной речи методом DTW-распознавания

Авторами предлагается способ построения антиэталонов и их использование для анализа произвольной слитно произнесенной русской фразы с целью выяснить, содержит ли она наперед заданное слово, которое в этой связи именуется ключевым. При этом если ключевое слово во фразе есть, то должен быть обнаружен лишь факт его наличия. Точная локализация этого слова не требуется.

Частным случаем рассматриваемой проблемы является задача определения ключевого слова среди отдельно произносимых русских слов. Различать неключевые слова при этом не нужно. Поэтому целесообразно использовать для них небольшое количество усредненных эталонов, которые уместно назвать антиэталонами ключевого слова. Ввиду малого количества антиэталонов это заведомо ускорит процедуру.

Есть S_0 – исходный список слов, из которого выбирается ключевое слово. Для него создается голосовой эталон, который также будем называть ключевым. Все остальные эталоны синтезируются из эталонов дифонной базы. Путем усреднения полученных синтетических эталонов создаются антиэталоны для ключевого слова. Многочисленные эксперименты показывают, что в большинстве случаев все неключевые слова классифицируются правильно, то есть оказываются ближе к одному из антиэталонов. Если ошибки и возникают, то даже для больших словарей S_0 их количество составляет несколько единиц. При анализе произвольных фраз с целью обнаружения ключевого слова это делает ошибку ложной тревоги малой.

Далее выполняется сегментация исследуемого речевого сигнала на отдельные звуки. Опираясь на сегментацию, производится распознавание с полученными эталонами всех речевых отрезков записанной фразы, содержащих столько глухих фрагментов, сколько их в ключевом слове. При выборе интервалов распознавания учитывается также близость количества отрезков сегментации к числу звуков в ключевом слове. Если результатом хотя бы одного из распознаваний является ключевое слово, то это слово считается найденным и программа выдает соответствующее сообщение. В противном случае считается, что ключевое слово не обнаружено. Малое число ошибок пропуска цели обеспечивается использованием голосового эталона для ключевого слова. Рисунки 5, 6 иллюстрируют результат поиска ключевого слова.

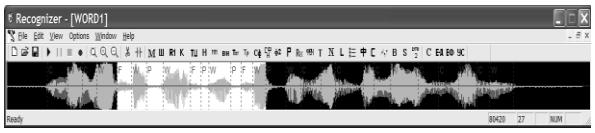

Рисунок 5 – Визуализация фразы *«Моя сокурсница уже сдала экзамен»*

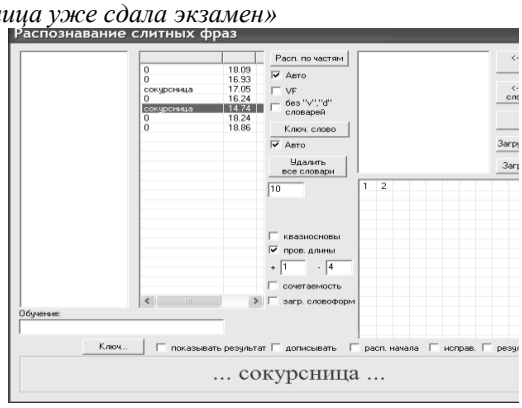

Рисунок 6 – Окно программы с результатом распознавания ключевого слова

Распознавание слитно произносимых сложных количественных числительных с поиском ключевого слова в потоке речи

Программно сформируем цифровой список чисел от 1 до 999 и заменим каждое из них его словесным выражением. Последнее представляет собой фразу из одного, двух или трех слов. Будем вначале работать со словарем, состоящим только из этих фраз с добавлением фраз, заканчивающихся словом «один», в которых это слово заменено словом «одна». В словарь добавляется также фраза (слово) «тыся». Далее будем именовать этот словарь как «Словарь 0».

Для каждой из фраз Словаря 0 автоматически создается транскрипция и строится эталон путем склеивания соответствующих эталонов дифонной базы. Из этих эталонов формируется дерево. Каждую сказанную фразу будем распознавать с помощью алгоритма DTW как цельный звуковой файл (не выделяя в нем отдельные слова). Такое распознавание обеспечивает надежный результат, получаемый достаточно быстро даже на весьма устаревшем компьютере со следующими параметрами: одноядерный процессор с тактовой частотой 2.4 ГГц и 1 Гб оперативной памяти.

Переходя к остальным числительным, заметим, что они состоят из двух частей: уже рассмотренное числительное от 1 до 999 и такое же числительное, которому предшествует нужная словоформа слова «тысяча». При этом одна из этих частей может быть пустой. Кроме того, вторая часть может состоять только из словоформы слова «тысяча». Далее будем использовать слово «часть» только в указанном смысле. Слово «один», если оно непосредственно предшествует слову «тысяча», заменяется словом «одна», а слово «два», предшествующее слову «тысячи», - словом «две».

Каждая из указанных частей выделяется и распознается отдельно, если разделяющая (открывающая, завершающая) словоформа слова «тысяча» заранее обнаруживается в сказанной фразе как ключевое слово в потоке речи. Это достигается путем использования вышеупомянутой априорной сегментации. А именно, ищется отвечающая звукосочетанию «тыся» последовательность сегментов WFW, открывающая фразу, или последовательность PWFW в любом другом месте фразы, и в обоих случаях на отрезке WFW проводится распознавание со Словарем 0. Если найденных отрезков более одного, выбирается тот, для которого DTW-расстояние до звукосочетания «тыся» минимально.

Исследование эффективности метода при распознавании изолированных слов

С помощью разработанного программного обеспечения было проведено исследование эффективности метода распознавания путём сравнения эффективности распознавания отдельно произносимых слов с методом на основе скрытых марковских моделей (в качестве тестовой платформы был использован HTK Toolkit) и коммерческой программой распознавания речи Voco. Для оценки качества распознавания речи использовался показатель процента корректно распознанных слов (WCR – Word Correctly Recognized).

В экспериментах участвовали 5 дикторов. Был сформирован словарь объемом 100 слов. Для каждого диктора был создан банк речевых сигналов – результатов произнесения слов словаря. Банк был записан в 5 версиях: одна версия предназначалась для режима обучения, остальные 4 – для режима тестирования.

Запись производилась в условиях низкого уровня фонового шума (отношение сигнал-шум примерно 50 дБ. Под отношением сигнал-шум в данном случае подразумевается отношение средних мощностей сигнала и шума в полосе частот от 0 до 11 кГц). Параметры записи наборов слов: частота дискретизации – 22050 Гц; разрядность квантования – 8 бит; средняя длительность записанного слова – 2 с (включая окружающие слово паузы, длительностью не менее 0,3 с каждая).

В качестве тестовой платформы для

метода скрытых марковских моделей был использован HTK Toolkit, чтобы построить базовую систему с MFCC коэффициентами. Для каждого диктора на обучающей выборке была обучена своя акустическая модель.

По результатам тестирования при распознавании отдельных слов с помощью DTW и синтетических эталонов (использования эталонов слов, синтезированных из эталонов дифонов), качество распознавания на тех же аудиоданных повышается на 3 – 18% по сравнению с распознаванием методом скрытых марковских моделей, и на 9 – 15% по сравнению с системой Voco (таблица 1).

Было проведено также исследование эффективности метода обработки речевых данных для распознавания изолированных слов на большом словаре. В процессе исследования из грамматического словаря русского языка А.А. Зализняка [21] объемом около 100 тыс. (точная цифра: 94604) слов в начальных формах случайным образом было отобрано 2 тысячи слов – тестовый словарь для распознавания. Далее из них случайным образом выбиралось 10 слов, которые произносил диктор. Этот последний этап повторялся 50 раз. Исследование показало высокую эффективность разработанного метода: доля корректно распознаваемых слов составляла не менее 90%.

Таблица 1. Результаты сравнительного тестирования качества распознавания

иктор	DTW +дифоны (WCR,%)	HTK Toolkit (WCR,%)	Voco (WCR,%)
	98%	95%	89%
	98%	93%	87%
	97%	92%	82%
	96%	78%	85%
	99%	89%	90%

Отметим, что указанные результаты распознавания отдельных слов программой Voco объясняются тем, что она предназначена для распознавания слитной речи с использованием n-граммной языковой модели. Для отдельных слов языковая модель практически не работает, в этом случае система распознавания опирается только на встречаемость слова в тренировочных данных, что, по сути, не несет никакой полезной информации, и распознавание ведется только за счет акустической модели. Таким образом, можно сделать вывод, что собственно акустическое распознавание отдельных слов в данной работе реализовано лучше.

Для исследования эффективности работы алгоритма на большом словаре, из грамматического словаря русского языка А.А. Зализняка объемом около 100 тыс. (точная цифра: 94604) слов в начальных формах случайным образом было отобрано 2 тысячи слов – тестовый словарь для распознавания. Далее из них произвольно отбиралось и произносилось 10 слов. Этот последний этап повторялся 50 раз.

Аналогичные результаты получались для остальных 44 испытаний, так что можно было сделать вывод: для словаря объемом 2000 произвольно выбранных начальных форм доля корректно распознаваемых слов составляет не менее 90%.

Аналогичная процедура тестирования применялась также на случайным образом сформированном словаре в 30000 начальных форм. Результат был аналогичным: доля корректно распознаваемых слов составляла не менее 90%.

Выводы

В статье рассмотрены варианты применения для различных задач метода распознавания на основе алгоритма DTW, использующего фонемную сегментацию и эталоны слов, автоматически синтезируемые из эталонов дифонов. Метод отличается тем, что позволяет применять алгоритм DTW для распознавания со словарем большого объема (более десяти тысяч слов), а также для распознавания слитной речи. Использование дифонной базы эталонов диктора вместе с автоматическим транскриптором позволяет за счет синтеза эталонов слов из эталонов дифонов работать с произвольными словарями, задаваемыми в текстовом виде, без необходимости выполнения предварительного обучения эталонов всех слов.

С помощью разработанного программного обеспечения проведено исследование эффективности алгоритмов сегментации и обработки речевых данных с использованием дифонов путем сравнения эффективности распознавания отдельно произносимых слов с методом на основе скрытых марковских моделей. По результатам тестирования, при распознавании отдельных слов с помощью DTW и синтетических эталонов, качество распознавания на тех же аудиоданных повышается на 3 – 18% по сравнению с

распознаванием методом скрытых марковских моделей.

Было проведено также исследование эффективности при распознавании изолированных слов на большом словаре, которое показало высокую эффективность разработанного метода: доля корректно распознаваемых слов составляла не менее 90%.

Литература

1. Andreas T., Ghosh P., Georgiou P., Narayanan S. Robust Word Boundary Detection in Spontaneous Speech Using Acoustic and Lexical Cues // IEEE International Conference on Acoustics, Speech, and Signal Processing, Taipei, 2009. – p. 4785 – 4788.
2. Zhijian O., Ji X. A study of large vocabulary speech recognition decoding using finite-state graphs // Chinese Spoken Language Processing (ISCSLP), 7th International Symposium, 2010. – p. 123-128.
3. Hong K., Tan T., Tang E., Cheah Y. Linguistic stem concatenation for malay large vocabulary continuous speech recognition // Research and Development (SCOReD), 2010 IEEE Student Conference on, 2010. – p. 144-148.
4. Karpov A.A., Kipytkova I.S., Ronzhin A.L. Very large vocabulary ASR for spoken Russian with syntactic and morphemic analysis // Proceedings of INTERSPEECH' 2011, Florence, 2011. – p. 3161-3164.
5. Susman D., Kopru S., Yazici A. Turkish Large Vocabulary Continuous Speech Recognition by using limited audio corpus // Signal Processing and Communications Applications Conference, 2012. – p. 1-4.
6. Saon G., Chien J. Large-Vocabulary Continuous Speech Recognition Systems: A Look at Some Recent Advances // Signal Processing Magazine, Vol. 29, № 6, 2012. – p.18-33.
7. Stas J., Hladek D., Juhar J., Zlacky D. Analysis of morph-based language modeling and speech recognition in Slovak // Information and communication technologies and services, Vol. 10, № 4, 2012. – p. 291-296.
8. Furui S. Recent progress in corpus-based spontaneous speech recognition // In IEICETRANS. INF. and SYST, Tokio, Japan, 2005. – p. 366-375.
9. Schalkwyk J., Beeferman D., Beaufays F., Byrne B., Chelba C., Cohen M., Kamvar M., Strope B. Google search by voice: a case study // Advances in Speech Recognition: Mobile Environments, Call Centers and Clinics, 2010. – p. 61-90.
10. Винцюк Т.К. Распознавание слов устной речи методами динамического программирования // Кибернетика, №. 1, 1968. – с. 81-88.
11. Sakoe H., Chiba S. Dynamic programming algorithm optimization for spoken word recognition // IEEE Trans. on Acoust., Speech an Signal Processing, Vol. 26, № 1, 1978. – p. 43-49.
12. Jelinec F. Statistical methods for speech recognition. – Cambridge, Mass.: MIT Press., 1998. – 305 p.
13. Бурибаева А.К., Дорохина Г.В., Ниценко А.В., Шелепов В.Ю. Сегментация и дифонное распознавание речевых сигналов // Труды СПИИРАН. – 2013. – № 31. – С. 20–42.
14. Шелепов В.Ю., Ниценко А.В. Сегментация и дифонное распознавание речи. – Донецк: ГУ ИПИИ, 2015. – 231 с.
15. Nitsenko A.V. A «by part» method of Russian word speech recognition // Eurasian Journal of Mathematical and Computer Applications, Vol.1, Iss. 2, 2014. – p. 102-109.
16. Шелепов В.Ю., Ниценко А.В. К проблеме распознавания слитной речи // Искусственный интеллект. – 2012. – №4 – С.272 – 281.
17. Шелепов В.Ю., Ниценко А.В. О некоторых вопросах, связанных с дифонным распознаванием и распознаванием слитной речи // Искусственный интеллект. – 2013. – №3 – С. 209 – 216.
18. V.Ju Shelepov, A.V.Nicenko. Recognition of the continuous-speech russian phrases using their voiceless fragments // Eurasian journal of mathematical and computer applications, Vol. 4., №4., 2016. — P. 19-24.
19. О. Н. Ляшевская, С. А. Шаров, Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). – М.: Азбуковник, 2009. – 1112 с.
20. Шелепов В.Ю., Ниценко А.В. О распознавании первого звука в слитном речевом отрезке. // Проблемы искусственного интеллекта. – 2015. – № 0(1). – С. 116 – 122.
21. Зализняк А.А. Грамматический словарь русского языка. Словоизменение. / А.А. Зализняк. – М.: Аст-пресс, 2008. – 880 с.

Ниценко А.В., Шелепов В.Ю О возможностях алгоритма DTW при распознавании речевых сигналов. *В статье рассмотрены варианты применения для различных задач метода распознавания на основе алгоритма DTW, использующего фонемную сегментацию и эталоны слов, автоматически синтезируемые из эталонов дифонов. Метод отличается тем, что позволяет применять алгоритм DTW для распознавания со словарем большого объема (более десяти тысяч слов), а также для распознавания слитной речи. Использование дифонной базы эталонов диктора вместе с автоматическим транскриптором позволяет за счет синтеза эталонов слов из эталонов дифонов работать с произвольными словарями, задаваемыми в текстовом виде, без необходимости выполнения предварительного обучения эталонов всех слов.*
Ключевые слова: *распознавание речи, динамическое программирование, алгоритм DTW, сегментация, дифон, антиэталон.*
Nitsenko A.V., Shelepov V.Ju., About DTW-algorithm capabilities for speech recognition. *In the article are considered variants of application for various problems of the speech recognition method based on the DTW algorithm using phonemic segmentation and word templates automatically synthesized from the diphone templates. The method advantage is that it allows to apply the DTW algorithm for large vocabulary speech recognition (more than ten thousand words), as well as for continuous speech recognition. The use of the speaker's diphone samples database together with the automatic transcription makes it possible to work with arbitrary dictionaries specified as the text by synthesizing the word templates from diphone templates without the need to perform the preliminary training of the whole word templates.*
Keywords: *speech recognition, dynamical programming, DTW algorithm, speech segmentation, diphone, anti-sample.*

Статья поступила в редакцию 20..2017
Рекомендована к публикации д-ром физ.-мат.. наук А.С. Миненко