

УДК 004.89:004.93

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

Институт проблем искусственного интеллекта МОН Украины и НАН Украины, г. Донецк
Украина, 83048, г. Донецк, ул. Артема, 118 б
Донецкий национальный технический университет, Украина
Украина, 83000, г. Донецк, ул. Артема, 58

О некоторых вопросах, связанных с дифонным распознаванием и распознаванием слитной речи

A.V. Nicenko, V.Ju. Sheleпов, G.V. Dorohina

*Institute of Artificial Intelligence MES of Ukraine and MAS of Ukraine, c. Donetsk,
Ukraine, 83048, c. Donetsk, Artema st., 118 b
Donetsk national technical University
Ukraine, 83000, Donetsk, Artema st., 58*

On Some Questions of Diphone Recognition and Recognition of Continuous Speech

А.В. Ніценко, В. Ю. Шелепов, Г.В. Дорохіна

Институт проблем штучного інтелекту МОН України і НАН України, м. Донецьк
Україна, 83048, м. Донецьк, вул. Артема, 118 б
Донецький національний технічний університет, Україна
Україна, 83000, м. Донецьк, вул. Артема, 58

Про деякі питання, пов'язані з дифонним розпізнаванням та розпізнаванням зв'язного мовлення

В статье обсуждается дифонное распознавание с использованием и без использования межфонеменной обработки, методы ускорения распознавания, способы быстрого создания дифонной базы, модификация эталонов дифонов в случае ошибки при распознавании, использование второго минимума при распознавании слитной речи, распознавание слов по частям, текстовый редактор с автоматически добавляемой парадигмой нового слова и голосовым вводом.

Ключевые слова: дифон, алгоритм DTW, ускорение распознавания, создание дифонной базы, модификация дифонов, слитная речь, второй минимум, распознавание слов по частям, текстовый редактор.

The subject of the article is: diphone recognition with and without interphone processing, recognition acceleration methods, fast creation diphone-base way, modification diphone-patterns in the case of recognition error, using of the second minimum for continuous speech, recognition words by part, text editing program with automatic adding of new word paradigm and voice inputing.

Key words: diphone, DTW- algorithm, diphone-base creation, modification of diphones, continuous speech, the second minimum, recognition words by part, text editing program.

У статті обговорюється дифонне розпізнавання з використанням та без використання міжфонемною обробки, методи прискорення розпізнавання, способи швидкого створення дифонної бази, модифікація еталонів дифонів у випадку помилки в розпізнаванні, використання другого мінімуму під час розпізнавання зв'язного мовлення, розпізнавання слів за частинами, текстовий редактор з автоматичним додаванням парадигми нового слова та голосовим вводом.

Ключові слова: дифон, алгоритм DTW, прискорення розпізнавання, створення дифонної бази, модифікація еталонів дифонів, зв'язне мовлення, другий мінімум, розпізнавання слів за частинами, текстовий редактор

1 Сравнение распознавания без использования и с использованием межфонемной обработки

Использованию дифонов при распознавании отдельно произносимых слов и слитной речи посвящены работы [1], [2]. В работе [1] было отмечено, что DTW-распознавание слова с эталонами, построенными из эталонов дифонов, возможно как для сигнала, в котором удалены стационарные части звуков (межфонемная обработка), так и для исходного сигнала. На рис. 1 и 2 в верхнем поле слева представлен некоторый словарь для распознавания. В среднем верхнем поле для произнесенного слова «ЗАПИСАТЬ» приведен список кандидатов на распознавание с указанием DTW-расстояний.

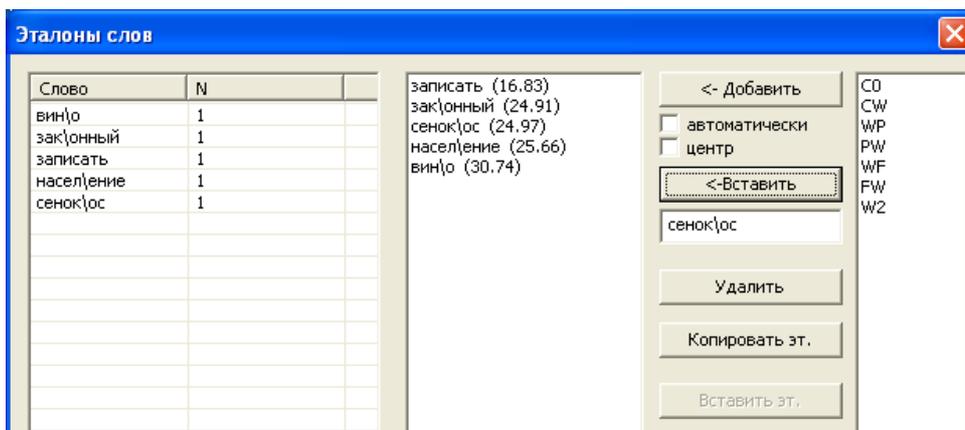


Рисунок 1 – Результат распознавания без межфонемной обработки сигнала

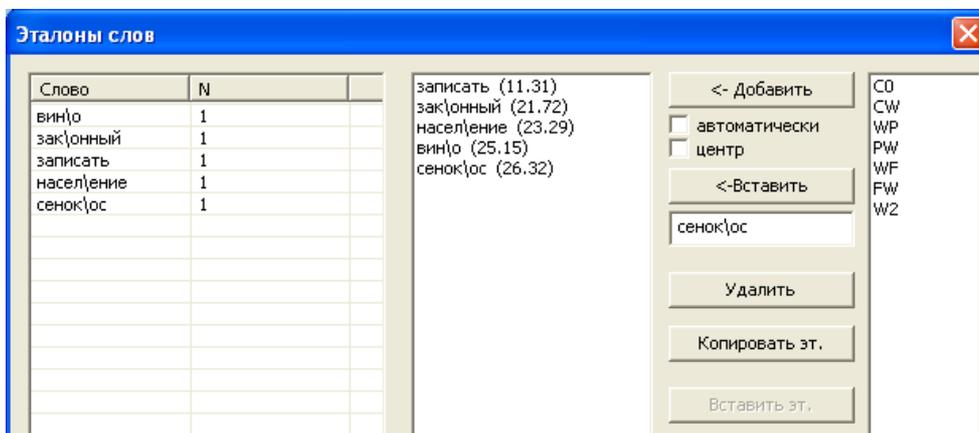


Рисунок 2 – Результат распознавания сигнала после межфонемной обработки

Рисунок 1 соответствует распознаванию исходного сигнала, а рис. 2 – распознаванию сигнала после межфонемной обработки. В первом случае отношение двух первых расстояний в списке есть

$$\frac{16,83}{24,91} \approx 0,67,$$

а во втором случае оно равно

$$\frac{11,31}{21,72} \approx 0,52.$$

Аналогичный результат прослеживается во всех других наших экспериментах: распознаваемое слово лучше отделяется от следующего за ним по DTW-расстоянию после межфонемной обработки, нежели без нее. Вывод: при дифонном распознавании записанный речевой сигнал целесообразно подвергать предварительной межфонемной обработке.

2 Проблемы ускорения и оптимизации распознавания

При распознавании больших и сверхбольших словарей чрезвычайно актуальной становится проблема скорости работы системы. Ниже предлагается три способа ускорения распознавания.

1. Использование классификации по длине слова.

Поскольку наша система сразу после записи сегментирует слово, мы получаем представление о количестве входящих в него дифонов (длина сказанного слова). С другой стороны, при создании дерева транскрипций, мы получаем точную информацию о длине каждого слова словаря. Очевидно, нет смысла искать результат распознавания среди слов, которые сильно отличаются от сказанного по длине. Учитывая возможные ошибки сегментации, мы ищем результат распознавания среди слов, которые отличаются от сказанного по количеству дифонов не более чем на два. А именно, прорабатывая очередную ветвь дерева эталонов, мы начинаем с подсчета количества входящих в нее дифонов, и переходим к вычислению расстояния до сказанного только в том случае, когда упомянутое количество отличается от длины сказанного не более чем на 2.

2. Использование VF-транскрипции.

Напомним, что при сегментации наша система осуществляет также широкую фонетическую классификацию составляющих звуков. С целью увеличения надежности ограничимся делением этих звуков на звонкие (идентификатор V) и глухие (в наших обозначениях идентификатор F). Таким образом, мы получаем обобщенную VF-транскрипцию сказанного. Ясно, что результат распознавания следует искать лишь среди слов словаря с такой же VF-транскрипцией. Чтобы после такого сокращения множества кандидатов на распознавание не создавать дерево заново, мы с самого начала записываем в конечном узле каждого слова его VF-транскрипцию. После этого процесс распознавания строится аналогично тому, как это делается при классификации по длине: вычисление DTW-расстояния осуществляется только для слов с нужной VF-транскрипцией.

3. Классификация по первому звуку слова.

Указанная классификация является лингвистически наиболее естественной. На сегодняшний день мы можем надежно классифицировать первый звук слова, если он является гласным, фрикативным или глухим взрывным. Звонкие согласные мы пока предпочитаем классифицировать, не распознавая их между собой. Таким образом, после сегментации мы используем первые полудифоны $a\theta$, $u\theta$, ..., $e\theta$, $и\theta$, $и\theta$, $ка\theta$, $ки\theta$, ..., $С\theta$. Символ $С\theta$ соответствует начальному участку произвольного звонкого согласного. Эти полудифоны мы распознаем заранее и при дальнейшей работе с деревом ограничиваемся ветвями, которые начинаются с распознанного полудифона.

3 Быстрое создание дифонной базы

Наглядная программа для создания эталонов дифонов может быть описана следующим образом. Записанный сигнал сразу автоматически сегментируется (рис. 3).

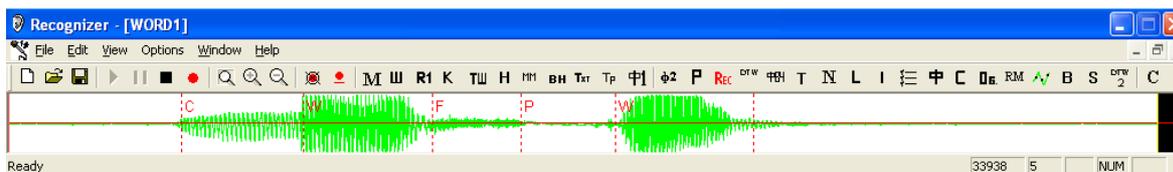


Рисунок 3 – Визуализация сигнала для слова «ласка» с сегментацией

При этом в окне программы создаётся список имен межфонемных переходов в терминах широкой фонетической классификации (W, C, F, P). Выделение элемента этого списка сопровождается выделением в сигнале соответствующего дифона (рис. 4, 5).

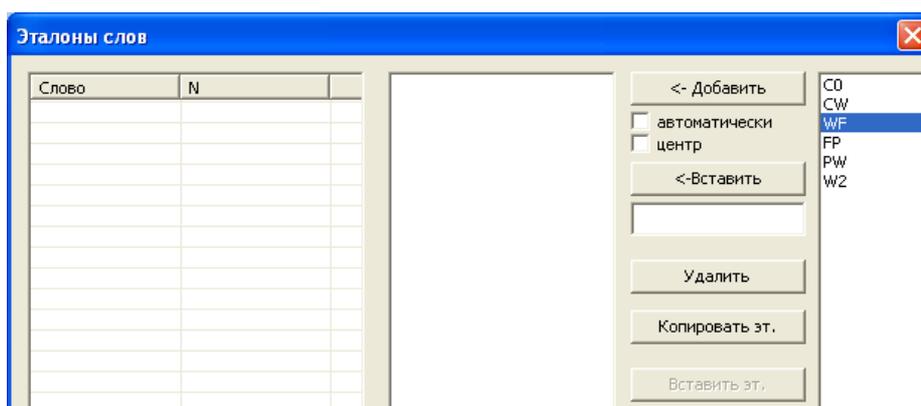


Рисунок 4 – Фрагмент программы: в правом верхнем поле список межфонемных переходов для слова «ласка»

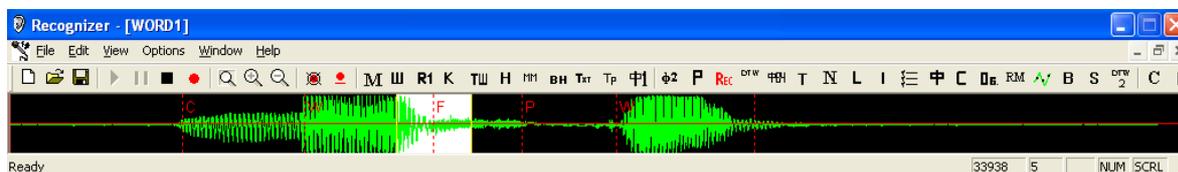


Рисунок 5 – Фрагмент программы: показан результат автоматического выделения дифона при выделении элемента списка на рис. 4

При нажатии кнопки «вставить» создается эталон дифона, включаемый в базу эталонов.

На самом деле все указанные операции автоматизируются, что позволяет получить программу быстрого создания базы дифонов. Ниже приведен перечень звуко-сочетаний, произнесение которых обеспечивает создание дифонов для распознавания произвольных слов со строгим чередованием гласных и согласных звуков:

абавагада, ажазакала, аманапара, басатафахацаша, аб\е, ав\е, аг\е, ад\е, а\е, аз\е, ак\е, ал\е, ам\е, ан\е, ап\е, ар\е, ас\е, ат\е, аф\е, ах\е, ач\е, аш\е, леб\ев\ег\ед\е, лез\ек\ел\ем\е, \ен\еп\ер\ес\ет\е, б\еф\ех\еч\ещ\е, леба, лева, лега, леда, ле\е, лежа, леза, лека, лела, лема, лена, лепа, лера, леса, лета, лефа, леха, леца, леша, ёб\ёв\ёг\ёд\ё, ёз\ёк\ёл\ём\ё, ён\еп\ёр\ес\ёт\ё, в\ёф\ёх\ёч\ёщ\ё, ёб\о, ёв\о, ёг\о, ёд\о, ёё, ёж\о, ёз\о, ёк\о, ёл\о, ём\о, ён\о, ёп\о, ёр\о, ёс\о, ёт\о, ёф\о, ёх\о, ёц\о, ёш\о, ибивигиди, изикилими, инипирисити, гифихичищи, ибу, иву, игу, иду, и\е, ижу, изу, ику, илу, иму, ину, ипу, иру, ису, иту, ифу, иху, ицу, ишу, \об\ов\ог\од\о, \ож\оз\ок\ол\о, \ом\он\оп\ор\о, \ос\от\оф\ох\оц\ош\о, \об\ё, \ов\ё, \ог\ё, \од\ё, \оз\ё, \оё, \ок\ё, \ол\ё, \ом\ё, \он\ё, \оп\ё, \ор\ё, \ос\ё, \от\ё, \оф\ё, \ох\ё, \оч\ё, ощё, убувугуду, ужузукулу, умунупуру, гусутуфухуцушу, уби, уви, уги, уди, у\е, узи, уки,

ули, уми, уни, упи, ури, уси, ути, уфи, ухи, учи, ущи, ыбывыгыды, ыжызыкылы, ымыныпыры, дысытыфыхыщышы, ыбю, ывю, ыгю, ыдю, ызю, ылю, ыкю, ылю, ымя, ыню, ыпо, ырю, ысю, ытю, ыфю, ыхю, ычю, ыщю, эбэвэгэдэ, эжэзэкэлэ, эмэнэпэрэ, жэсэтэфэхэцшэ, эб\я, эв\я, эг\я, эд\я, эз\я, эя, эк\я, эл\я, эм\я, эн\я, эп\я, эр\я, эс\я, эт\я, эф\я, эх\я, эч\я, эщ\я, юбююгюдю, юзюкюлюмя, юнюпюрюсютю, дюфюхючющю, юю, юбы, ювы, югы, юды, южы, юзы, юкы, юлы, юмы, юны, юпы, юры, юсы, юты, юфы, юхы, юцы, юшы, \яб\яв\яг\яд\я, \яз\як\ял\ям\я, \ян\яп\яр\яс\ят\я, з\яф\ях\яч\ящ\я, \ябэ, \явэ, \ягэ, \ядэ, \яе, \яжэ, \язэ, \якэ, \ялэ, \ямэ, \янэ, \япэ, \ярэ, \ясэ, \ятэ, \яфэ, \яхэ, \яцэ, \яшэ, за, лал, мам, нан, рар, сас, цац, шаш, л\яль, м\ямь, н\янь, р\ярь, с\ясь, ч\яч, щ\ящ, как, пап, тат, к\о, ку, кы, кэ, п\о, пу, пы, пэ, т\о, ту, ты, тэ, кякь, пяпь, тять, кё, кю, ки, к\е, пё, пю, пи, п\е, тё, тю, ти, т\е, фаф, хах, цац, ф\яфь, х\яхь, ч\яч, ой.

На рис. 6 представлено окно программы обучения.

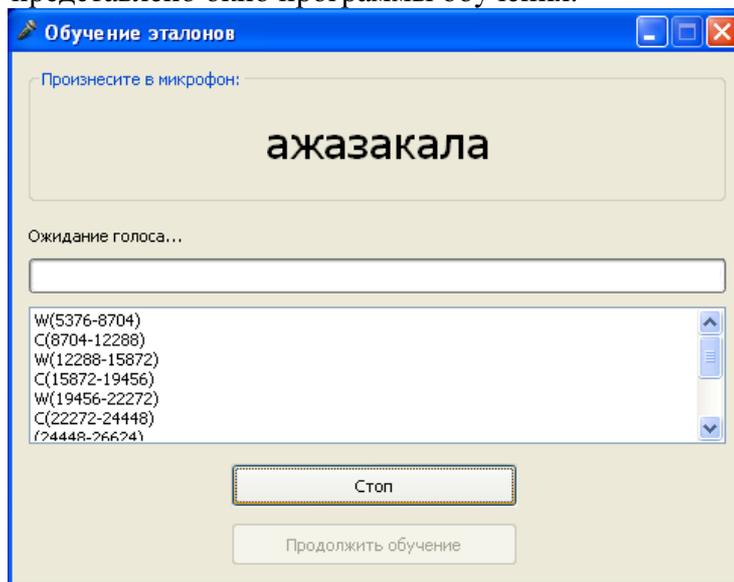


Рисунок 6 – Окно программы обучения:

После нажатия кнопки «Начать обучение – Стоп» программа предлагает очередное звукосочетание. Когда слово произнесено, она автоматически выделяет все входящие в него дифоны, создает их эталоны и предлагает следующее звукосочетание. Звукосочетания подобраны так, чтобы обеспечивать надежную сегментацию и выделение дифонов. В случае, когда все же полученная сегментация не соответствует ожидаемому количеству сегментов, программа просит пользователя повторить произнесение. Вся процедура обучения по указанному списку занимает 10 – 15 минут.

Аналогичным образом создается список обучающих звукосочетаний полной базы дифонов для распознавания произвольных слов и слитной речи. В этом случае обучение занимает около часа.

4 Модификация эталонов дифонов в случае ошибки при распознавании

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 доучивания: в случае ошибки пользователь указывает мышкой в списке или вводит с клавиатуры правильное слово; про-

грамма, сегментируя сигнал, создает эталоны прозвучавших дифонов и с их помощью модифицирует эталоны базы путем усреднения. Здесь ошибки в сегментации становятся важными, и программа, зная слово, в большинстве случаев исправляет их:

- лишние удвоения сегментов С, Р, F заменяются одним сегментом;
- не разделенные при сегментации С, Р, F – участки разделяются на два сегмента (метка ставится посередине);
- последовательность сегментов WCW заменяется на WW, если в транскрипции слова на этом месте присутствуют WW (метка ставится посередине отрезка WW);
- удаляются лишние метки в сочетаниях FP и PF, если в транскрипции на этом месте один сегмент F или P;
- добавляются метки в сегменты F и P, если в транскрипции на этом месте сочетание FP или PF (соответствующая метка ставится посередине);
- добавляются метки в сегмент W, если в транскрипции на этом месте WW (метка ставится посередине);
- добавляются пропущенные метки W и C перед сегментами F и P или после сегментов F и P.

5 Использование второго минимума при распознавании слитной речи

В работе [2] сформулирован принцип минимума DTW-расстояния для определения первого слова слитно произносимой фразы при определенном ограничении на состав распознаваемого словаря. А именно, в словаре не должно быть таких пар слов, что транскрипция одного из них получается из транскрипции другого приписыванием в конце дополнительных транскрипционных символов. В противном случае, при произнесении более длинного слова такой пары, DTW-расстояние до слова с более короткой транскрипцией может оказаться меньше.

Для того чтобы программа правильно работала при наличии слов с такими фонетическими вложениями, предлагается использовать «МЕТОД ВТОРОГО МИНИМУМА». Пусть распознается слитно произносимая фраза из двух слов. Найдем минимум-гипотезу для первого слова (гипотеза 1) и, распознав оставшуюся часть сигнала, получим пару слов. Для этой пары слов построим эталон, как для слитно произносимой фразы, и вычислим до него DTW-расстояние d_1 исходного сигнала. Рассмотрим все гипотезы, следующие за гипотезой 1. Среди них выберем ту, которой соответствует минимальное DTW-расстояние (второй минимум) и, действуя с ней далее так же, как с гипотезой 1, найдем расстояние d_2 . Мы получили два варианта распознавания исходной фразы. Из них следует выбрать тот, для которого величина d_i ($i = 1, 2$) меньше.

Если фраза состоит из k слов, то описанный алгоритм обобщается следующим образом. Находим для первого слова два варианта, используя первый и второй минимумы. В первом случае от конца первого слова аналогичным образом ищем второе слово, что в свою очередь дает два варианта. То же делаем во втором случае (еще два варианта). Продолжая действовать таким же образом, найдем 2^k наборов слов. Для каждого из этих наборов построим эталон, как для слитно произнесенной фразы, и найдем расстояние исходного сигнала до каждого из этих эталонов. Результатом распознавания объявляется тот набор слов, расстояние до которого минимально. Количество вычислений при увеличении k растет не слишком сильно, так как многие из упомянутых наборов слов совпадают между собой и вычисления для них, как для слитно произносимых фраз повторять не надо.

6 Распознавание слов по частям

Предлагаемую технику распознавания слитной речи можно применить также для распознавания словоформ одного и того же слова, выделяя в них общую часть (квазиоснова) и изменяющиеся части (квазифлексии). Рассматривая множество словоформ различных слов, объединяем квазиосновы в один словарь, а квазифлексии – в другой. Произнеся словоформу, находим квазиоснову по принципу минимума DTW-расстояния, при этом выделяется соответствующая часть речевого сигнала. Оставшуюся часть распознаём, используя словарь квазифлексий. Квазифлексии, очевидно, являются общими для большой группы слов. Если у нас есть m квазиоснов и n квазифлексий, то их комбинации образуют $m \times n$ словоформ и, распознавая словоформы как целое, мы имели бы словарь для распознавания из $m \times n$ объектов. Распознавая же квазиосновы и квазифлексии отдельно, мы распознаем $m + n$ объектов. Правда, работая с квазиосновой, согласно алгоритму определения первого слова в слитной речи, приходится совершать не одно, а несколько распознаваний (от начала до первой метки, от начала до второй метки и так далее). Однако в рассматриваемом случае можно начинать с распознавания от начала до предполагаемой конечной метки основы, которая отвечает началу самой длинной квазифлексии. В результате процесс распознавания квазиосновы сильно сокращается.

7 О текстовом редакторе с автоматически добавляемой парадигмой нового слова и голосовым вводом

Результаты, полученные в [1], позволяют заняться разработкой своеобразного текстового редактора с голосовым вводом. Предполагается первоначальный словарный запас в несколько десятков тысяч русских словоформ, отвечающих словам из «Нового частотного словаря русской лексики», составленного С.А. Шаровым и О.Н. Ляшевской на основе Национального корпуса русского языка [3]. При этом используется тысяча наиболее часто употребляемых глаголов, тысяча наиболее часто употребляемых существительных и так далее. Словарь в несколько десятков тысяч получается при включении всех словоформ упомянутых слов. Программа должна давать пользователю возможность с самого начала набирать произвольный текст. Все упомянутые словоформы пользователь имеет возможность вводить голосом. Если же слово не входит в первоначальный словарь, оно вводится с клавиатуры. И при голосовом вводе и при вводе с клавиатуры результат первоначально содержится в отдельном поле, что дает пользователю возможность контролировать и при необходимости исправлять его. По нажатию пробела результат передается в текст. Для того чтобы исключить ошибки при ручном наборе проверяется наличие введенного слова в обширном словаре русских словоформ *Checker*, используемом для такой проверки. Если слово все-таки не попало в текст, но пользователь убедился в его правильном написании, он нажимает «Enter», после чего слово вставляется в текст и вставляется в *Checker* вместе с его полной парадигмой, то есть набором всех его словоформ.

Если слово отсутствовало в первоначальном словаре для распознавания и набрано с клавиатуры, то в момент нажатия пробела в словарь для распознавания вводится его полная парадигма и создается новое дерево эталонов, так что в дальнейшем любую из этих словоформ можно вводить голосом.

Наконец, если результат голосового ввода оказался ошибочным и пользователь заменяет его с клавиатуры нужным словом, программа «знает», какие дифоны участ-

вуют в нужном эталоне и автоматически модифицирует их путем усреднения дифонов, существовавших до этого в базе, и дифонов прозвучавшего слова, добиваясь тем самым правильного распознавания.

Все текстовые словари задаются в виде деревьев, что обеспечивает быстрый поиск. Работа с парадигмами слов основана на использовании большого декларативного морфоанализатора, разработанного в отделе распознавания речевых образов Института проблем искусственного интеллекта НАН и МОН Украины [4].

Таким образом, описываемый редактор должен

- 1) позволять с самого начала набирать нужные тексты;
- 2) автоматически пополнять словарь для распознавания парадигмами новых слов;
- 3) совершенствовать по ходу дела дифонную базу, улучшая качество распознавателя.

Все это становится возможным благодаря процедуре автоматического создания из эталонов дифонов эталонов новых слов, появляющихся в процессе работы только в текстовом виде.

Литература

1. Шелепов В.Ю. О распознавании речи на основе межфонемных переходов / В.Ю. Шелепов, А.В. Ниценко, Г.В. Дорохина // Искусственный интеллект. – 2012. – № 1. – С. 132-139.
2. Шелепов В.Ю. К проблеме распознавания слитной речи / В.Ю. Шелепов, А.В. Ниценко // Искусственный интеллект. – 2012. – № 4. – С. 272-281.
3. Режим доступа : <http://dict.ruslang.ru/freq.php>
4. Дорохина Г.В. Модуль морфологического анализа слов русского языка / Г.В. Дорохина, А.П. Павлюкова // Искусственный интеллект. – 2004. – № 3. – С. 636-642.

Literatura

1. Shelepov V.Ju. O raspoznavanii rechi na osnove mezhfonomnyh perehodov / V.Ju. Shelepov, A.V. Nicenko, G.V. Dorohina // Iskusstvennyj intellekt. – 2012. – № 1. – С. 132-139.
2. Shelepov V.Ju. K probleme raspoznavanija slitnoj rechi / V.Ju. Shelepov, A.V. Nicenko // Iskusstvennyj intellekt. – 2012. – № 4. – С. 272-281.
3. Rezhim dostupa : <http://dict.ruslang.ru/freq.php>
4. Dorohina G.V. Modul' morfologicheskogo analiza slov russkogo jazyka / G.V. Dorohina, A.P. Pavljukova // Iskusstvennyj intellekt. – 2004. – № 3. – С. 636-642.

A.V. Nicenko, V.Ju. Shelepov, G.V. Dorohina

On Some Questions of Diphone Recognition and Recognition of Continuous Speech

The article lies in course of authors approach to speech recognition by dynamic programming method with patterns which automatically created from diphone patterns using transcription. It contains some ideas which are important for realization of this approach in large vocabularies recognition and continues speech: fast creation diphone-base way, recognition acceleration methods, modification diphone-patterns in the case of recognition error, continues speech recognition with vocabulary which have phonetic inclusions, recognition of word-forms like continuous speech segments with stem and ending classification. The last section contains description of structure of important application: text editing program with automatic adding of new word paradigm and voice inputing.

Статья поступила в редакцию 05.07.2013.