УДК: 004.934.2

В.Ю. Шелепов, А.В. Ниценко

Использование малых речевых единиц при распознавании речи с помощью алгоритма *DTW*

Описана разработанная авторами техника автоматической сегментации речевых сигналов и *DTW*-распознавания русской речи с использованием малых речевых единиц. Предложено использовать дифоны, содержащие межфонемные переходы. Предлагаемые подходы реализованы в реальных программах распознавания, которые показывают достаточно высокий уровень надежности.

Ключевые слова: автоматическая сегментация, дифон, стационарная часть звука, *DTW*-алгоритм.

Описано розроблену авторами техніку автоматичної сегментації мовних сигналів і *DTW*-розпізнавання російської мови з використанням малих мовних одиниць. Запропоновано використовувати дифони, що містять міжфонемні переходи. Запропоновані підходи реалізовано в реальних програмах розпізнавання, які показують досить високий рівень надійності.

Ключові слова: автоматична сегментація, дифон, стаціонарна частина звуку, *DTW*-алгоритм.

Введение. Большинство современных работ по распознаванию речи развивает алгоритмы, использующие Скрытые Марковские Модели (*Hidden Markov Models – HMM*) (например, [1–5]). В то же время внимание уделяется другому традиционному методу, идущему от Т.К. Винцюка. Это так называемый алгоритм *DTW* [6, 7]. Результаты настоящей статьи лежат в русле этого подхода. В основе – принадлежащий авторам метод автоматической сегментации речевого сигнала [8–10].

Разработанные алгоритмы

Используется восьмибитная запись с частотой дискретизации 22050 Гц. В работах [11, 17, 19] предложена технология распознавания речи, основанная на автоматической сегментации речевого сигнала и использовании дифонов. Под сегментацией подразумеватся разбиение сигнала на отрезки, соответствующие отдельным звукам речи. Одновременно осуществляется автоматическое отнесение звуков к классам широкой фонетической классификации: *W* – согласный, *C* – звонкий согласный, *F* – глухой фрикативный, *P* – глухой взрывной (паузообразный) звуки. На рис. 1 приведена визуализация сигнала, соответствующего слову «событие» с автоматической сегментацией.

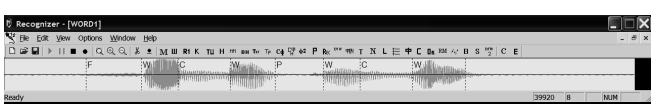

Рис. 1

В качестве основного инструмента сегментации используется численный аналог полной вариации, вычисляемый для последовательных отрезков по 256 отсчетов:

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

Дифон в первоначальном смысле этого слова – отрезок сигнала между серединами участков двух соседних звуков. Используется формализованное понятие дифона – это три окна по 368 отсчетов слева от автоматически проставляемой метки между звуками и три таких же окна справа от этой метки (368 – длина удвоенного квазипериода для мужского голоса средней высоты). Имя дифона – пара символов, обозначающих соответствующие звуки. На рис. 2 представлен результат автоматического выделения дифона *ма* в слове *сама*.

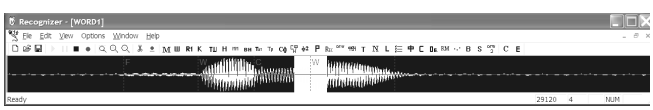

Рис. 2

Полным колебанием называем отрезок сигнала между двумя соседними локальными максимумами (если максимумы не строгие, подразумевается отрезок от начала первого до начала второго). Длина полного колебания – количество содержащихся в нем отсчетов. Каждому отрезку в 368 отсчетов ставится в соответствие 29-мерный вектор относительных частот длин полных колебаний [11]. В процессе обучения для определенных ранее дифонов создаются представления – последовательности шести описанных векторов, которые рассматриваются как эталоны

соответствующих дифонов. Различные экземпляры эталонов одного и того же дифона можно усреднять, повышая их универсальность [11]. Таким образом, создается база эталонов дифонов. В расчете на слитную речь она насчитывает около 1700 единиц. Помимо «полных» дифонов используются начальные и конечные полудифоны длиной в три окна по 368 отсчетов. Их имена снабжаются знаками *ноль* и *два* соответственно.

Использование дифонов при распознавании представляется перспективным ввиду явления коартикуляции – взаимовлияния соседних звуков во время произнесения.

В определенном словаре для распознавания для каждого слова автоматически создается его транскрипция [18], и по последней строится цепочка имен соответствующих дифонов. Например:

далеко → да*l*еко → д0-да-а*l*-*l*е-ек-ко-о2 (2)

(здесь твердые согласные обозначаются кириллическими, а мягкие – соответствующими латинскими символами).

Ввиду автоматической сегментации можно было бы предположить возможность, сказав слово, распознавать соответствующую последовательность дифонов и по ней восстанавливать распознаваемое слово. Однако дифоны – слишком короткие речевые единицы, и при упомянутом их количестве неизбежны многочисленные ошибки. Ситуация становится значительно более выигрышной, если из эталонов дифонов склеивать эталоны целых слов и распознавать последние с использованием известного алгоритма *DTW* [12], а также [11, 15].

Таким образом, отпадает необходимость создавать голосом в процессе обучения эталон для каждого слова распознаваемого словаря. Эти словари достаточно задавать в текстовом виде. Совокупность автоматически создаваемых транскрипций слов организуется в виде дерева, работа с которым существенно ускоряет процесс *DTW*-распознавания [19].

Создание эталонов всех дифонов – достаточно долгий и трудоемкий процесс. Для этого имеется специальная подпрограмма, предлагающая осуществлять его путем произнесения набора звукосочетаний типа абба, абду, абгэ,…, сегментация которых может осуществляться практически безошибочно. При таком подходе создание полной базы дифонов занимает около часа.

Авторами разработана программа дифонного распознавателя, которая тестировалась на случайным образом сформированных списках слов объемом 30 тыс., взятых из словаря [13]. Число правильных распознаваний – не менее 90 процентов.

Далее ясно, что по порядку величины количество дифонов равно квадрату количества звуков. Поэтому количество эталонов должно сократиться во много раз, если попытаться использовать вместо дифонов стационарные части звуков. При этом очевидно, этого нельзя делать относительно звонких взрывных Б, Д, Г в твердом и мягком вариантах, ибо они различаются между собой только на переходе к следующему звуку. То же относится к глухим взрывным К, П, Т. Для всех остальных звуков такой подход представляется возможным.

Результаты

Авторами реализована программа распознавателя, работающего с такими речевыми единицами. База эталонов включает в себя эталоны дифонов, для которых первый звук – звонкий или глухой взрывной, а также эталоны стационарных частей гласных, невзрывных звонких согласных и глухих фрикативных звуков: *а*, *и*, *о*, … …*в*, *л*, *м*, … *с*, *ш*, … В связи с этим в общем случае отказываемся от начальных и конечных полудифонов. Однако сохраняются полудифоны с глухими взрывными в начале слова (с именами полных дифонов) и используются также в середине слова вместо полных дифонов. В базу включены также эталоны для квазипериодических частей звонких взрывных под именами б, г, д, *b*, *g*, *d* (совпадающие эталоны), и эталоны серединных паузообразных частей глухих взрывных под именами к, п, т, *k*, *@*, *t* (совпадающие эталоны). Здесь @ – используемый транскрипционный символ для звука п-мягкое. Аналогом цепочки (2) теперь будет последовательность далеко – да*l*еко – д-да-*l*-е-к-ко-о.

На рис. 3 изображено окно описываемой программы–распознавателя. Левое вертикальное поле содержит слова из словаря для распознавания в 1000 слов, правое – список кандидатов на распознавание с указанием *DTW*-расстояний до сказанного слова *высокомерие*, которое непосредственно перед этим добавлено в словарь в текстовом виде (левое верхнее горизонтальное поле) и для которого создан синтетический эталон после нажатия кнопки *Синтез фон*.

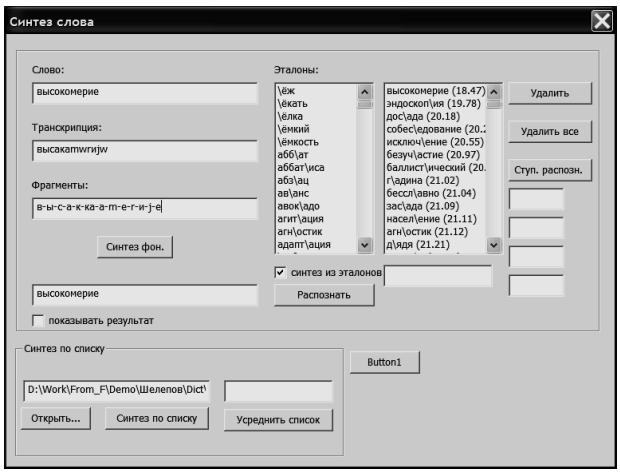

Рис. 3

Отметим, что описанные программы распознавания содержат функцию *дообучение*. В случае ошибочного распознавания пользователь вместо распознанного слова вводит правильное. Тогда компьютер корректирует, в случае необходимости, сегментацию и, зная прозвучавшие составляющие речевые единицы, усредняет их и соответствующие единицы базы.

Подчеркнем две ранее используемые идеи: использование дифонов и *DTW*-распознавание слов (или слитных фраз) по эталонам, склеенным из эталонов дифонов. Вторая идея может быть сформулирована в более общем виде как использование ограниченного числа малых речевых единиц так, что они распознаются не сами по себе, а применяются для создания эталонов смысловых речевых единиц (слова, фразы), которые можно распознавать, на основе алгоритма *DTW* со всеми его преимуществами. В настоящей статье описана система распознавания, использующая в качестве таких единиц чистые звуки и лишь некоторые дифоны. Это позволяет в разы сократить количество эталонов и время, затрачиваемое на обучение системы. Теперь время создания базы эталонов для конкретного диктора – менее 10 мин.

Заключение. Изложенное будет особенно существенным при обучении базы эталонов на больших речевых банках с целью создания дикторонезависимого или многодикторного распознавателя. При этом можно отметить также следующий факт: в эталонах слов все эталоны безударных гласных в твердых сочетаниях можно заменить одним нейтральным, например э, а в мягких – например, *и*. Тогда автоматически создаваемая транскрипция, например, для слова *обновление* будет выглядеть как *эбнэвlепии*. Как показывают эксперименты, распознавание при этом остается успешным. Таким образом, отпадет необходимость создания *дикторонезависимых* эталонов для разных безударных гласных.

1. *Робейко В.В.*, *Сажок М.М.* Розпізнавання спонтанного мовлення на основі акустичних композитних моделей слів у реальному часі // Штучний інтелект. – 2012. – № 4. – С. 253–263.
2. *Яценко В.В.*, *Сажок Н.Н.* Система устного перевода спонтанных высказываний в рамках предметных областей // УСиМ. – 2013. – № 4. – С. 63–70.
3. *Deshmukh S.D.*, *Bachute M.R.* Automatic Speech and Speaker Recognition by MFCC, HMM and Vector Quantization // Int. J. of Engineering and Innovative Technology. – 2013. – **3**, № 1. – P. 93–98.
4. *Zarrouk E.*, *Ayed Y.*, *Gargouri F.* Hybrid continuous speech recognition systems by HMM, MLP and SVM: a comparative study // Int. J. of Speech Technology. – 2014. – P. 1–11.
5. *Mulik V.*, *Mane V.*, *Jamadar I.* Hidden Markov Model Based Robust Speech Recognition // Int. J. of Innovative Research in Advanced Engineering (IJIRAE). – 2015. – **2**, № 2. – P. 262–271.
6. *Muda L.*, *Begam M.*, *Elamvazuthi I.* Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient and Dynamic Time Warping (DTW) Techniques // Journal of computing. – 2010. – **2**, № 3. – P. 138–143.
7. *Nandyala S.P.*, *Kumar T.K.* Hybrid HMM/DTW based Speech Recognition with Kernel Adaptive Filtering Method // Int. J. on Computational Sciences & Applications (IJCSA). – 2014. – **4**, № 1. – P. 11–21.
8. *Wang D.*, *Lu L.*, *Zhang H.* Speech segmentation without speech recognition // Proc. (ICASSP '03). – 2003. – **1**. – P. 468–471.
9. *Gómez J.A.*, *Calvo M.* Improvements on Automatic Speech Segmentation at the Phonetic Level // Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. – 2011. – **7042**. – P. 557–564.

10. *Natarajan V.A.*, *Jothilakshmi S.* Segmentation of Continuous Speech into Consonant and Vowel Units using Formant Frequencies // Int. J. of Comp. Appl. – 2012. – **56**, № 15. – P. 24–27.
11. *Сегментация* и дифонное распознавание речевых сигналов / А.К. Бурибаева, Г.В. Дорохина, А.В. Ниценко и др. // Тр. СПИИРАН, 2013. – Т. 31. – С. 20–42.
12. *Винцюк Т.К.* Анализ, распознавание и интерпретация речевых сигналов. – Киев: Наук. думка, 1987. – 262 с.
13. *Зализняк А.А.* Грамматический словарь русского языка. – М.: Русский язык, 1977. – 879 с.
14. *Шелепов В.Ю.*, *Ниценко А.В.* Структурная классификация слов русского языка. Новые алгоритмы сегментации речевого сигнала и распознавания некоторых классов фонем, // Искусственный интеллект. – 2007. – **17**. – С. 223–233.
15. *Шелепов В.Ю.* Лекции о распознавании речи, ІПШІ. – Донецк: Наука і освіта, 2009. – 192 с.
16. *Шелепов В.Ю.*, *Жук А.В.*, *Ниценко А.В.* Построение системы голосового управления компьютером на примере задачи набора математических формул // Искусственный интеллект. – 2010. – **3**. – С. 259–267.
17. *Шелепов В.Ю.*, *Ниценко А.В.*, *Дорохина Г.В.* О распознавании речи на основе межфонемных переходов // Там же. – 2012. – **1**. – С. 132–139.
18. *Шелепов В.Ю.*, *Ниценко А.В.* К проблеме распознавания слитной речи // Там же. – **4**. – С. 272–281.
19. *Шелепов В.Ю.*, *Ниценко А.В.*, *Дорохина Г.В.* О некоторых вопросах, связанных с дифонным распознаванием и распознаванием слитной речи // Там же. – 2013. – **3**. – С. 209–216.

Тел. для справок: +38 062 311-3424 (Донецк)
E-mail: *vladislav.shelepov*2012*@yandex.ua*,
*frost*109*@yandex.ua*, *nav_box@mail.ru*

UDC: 004.934.2

V.Ju. Shelepov, A.V. Nitsenko

Using Minor Language Units for Speech Recognition with the Help of DTW Algorithm

Keywords: automatic segmentation, diphone, stationary part of sound, DTW-algorithm.

Introduction: The article describes a technique of automatic speech segmentation and DTW-recognition using minor language units, developed by the authors for Russian speech. The main tool for segmentation is a numerical analogue of the total variation. In [11, 17,19] the authors suggest using the diphones containing interphoneme transitions as the minor language units. The templates for these are used to synthesize the templates of the semantic units, i.e. words and phrases. Then the DTW algorithm (with its advantages) is applied to the recognition of a word as a whole. As the result of this procedure there is no need to pronounce the words of the vocabulary under recognition during the training, thus a possibility to set the vocabulary just in text form is created.

Purpose: The purpose of the research is to reduce the size of the reference template database, and, as a consequence, the training time for a particular speaker.

Results: An innovation of this research is the use of exclusively diphones, whose first sound is one of explosive (b, g, d, k, p, t), and the stationary parts of other sounds. A set of automatically generated vocabulary words' transcriptions is organized in a tree structure, which considerably speeds up the process of recognition. The proposed approach is implemented in real-recognition software, demonstrating the high reliability.

●

Внимание !

Авторы статей **обязательно** должны подать структурированную (*Introduction*, *Purpose*, *Methods*, *Results*, *Conclusion*) расширенную аннотацию на английском до одной стр. текста через два интервала, информацию об авторах на английском и, кроме пристатейного списка литературы (на языке статьи), список литературы в транслитерации (с указанием в скобках перевода на англ. названия ссылки).