Ш

УДК 004.89:004.93

В. Ю. Шелепов, А. В. Ниценко
Государственное учреждение «Институт проблем искусственного интеллекта», г. Донецк
83048, г. Донецк, ул. Артема, дом 118 б

О РАСПОЗНАВАНИИ ПЕРВОГО ЗВУКА В СЛИТНОМ РЕЧЕВОМ ОТРЕЗКЕ

V. Ju. Shelepov, A. V. Nicenko
Public institution «Institute of Problems of Artificial intelligence», Donetsk
83048, Donetsk, Artema st., 118b

ON RECOGNITION OF THE FIRST SOUND IN CONTINUOUS SPEECH FRAGMENT

В. Ю. Шелєпов, А. В. Ниценко
Державна установа «Інститут проблем штучного інтелекту», м. Донецьк
83048, м. Донецьк, вул. Артема, буд. 118 б

ПРО РОЗПІЗНАВАННЯ ПЕРШОГО ЗВУКУ У ЗЛИТОМУ МОВНОМУ ВІДРІЗКУ

В статье предлагаются алгоритмы DTW-распознавания в слитном речевом отрезке первого звука или достаточно узкого класса, к которому он относится. Это важно в связи с проблемой ускорения распознавания для больших и сверхбольших словарей.
Ключевые слова: сегментация, участки первого звука, классы распознавания, переходы в зависимости от промежуточных результатов распознавания.

DTW-algorithms are proposed for recognition of the continuous speech fragment first sound or small enough class which contains it. This is important in connection with the problem of recognition speed-up for large and super large vocabularies.
Key words: segmentation, parts of the first sound, recognition classes, transitions depending on intermediate recognition results.

У статті пропонуються алгоритми DTW-розпізнавання у злітному мовному відрізку першого звуку або досить вузького класу, до якого він належить. Це важливо у зв'язку з проблемою прискорення розпізнавання для великих та надвеликих словників.
Ключові слова: сегментація, ділянки першого звуку, класи розпізнавання, переходи у залежності від проміжних результатів розпізнавання.

Будем вести изложение для случая распознавания русского слова как целого. Результаты целиком сохраняются для отрезков слитной речи.

Успехи последнего времени в распознавании речи, отражающиеся в первую очередь в поисковых системах Google и Yandex и распространяющиеся даже на мобильные устройства, сопряжены с работой в сети. В то же время проблема распознавания на малых локальных компьютерах остается актуальной.

При этом, помимо использования механизма скрытых марковских моделей (см., например, [1-5]), продолжают развиваться подходы, связанные с динамическим программированием (алгоритм DTW, см., например, [6], [7]). Данная работа лежит в этом русле. В основе – принадлежащий авторам метод автоматической сегментации речевого сигнала, основанный на использовании численного аналога полной вариации. Проблеме сегментации речевого сигнала посвящены также работы [8-10].

Очевидно, при распознавании больших и сверхбольших словарей неоценимую роль играло бы распознавание первого звука слова. Это наиболее естественный и эффективный способ ускорения распознавания в целом, которое чрезвычайно актуально для упомянутых словарей. Действительно, если компьютер очень быстро (ввиду небольшого числа звуков) определяет, что сказанное слово начинается на [и], то дальше он может вести распознавание только среди слов словаря, которые начинаются с этого звука. При этом первый звук обладает очевидными преимуществами: повышенная надежность в определении его начала и отсутствие влияния предшествующего звука, которого просто нет. Но отрицательную роль при работе с маленьким словарем таких звуков играет то, что они представляют собой очень короткие речевые единицы, а DTW-распознавание тем надежей, чем больше длина распознаваемого отрезка.

На сегодняшний день мы в состоянии распознавать первый звук, если он является гласным или одним из звуков [j], [ж], [ш], [щ] (здесь [j] – согласный, в звучании слов, которые при написании начинаются с букв «е, ё, ю, я»). В остальных случаях мы опознаем класс, содержащий помимо данного некоторые другие близкие звуки (см., ниже список (1)). Однако класс этот достаточно узок, чтобы сократить описанным образом количество кандидатов при распознавании слов большого словаря во много раз.

При 8-битной записи и частоте дискретизации 22 050 Гц, используется разбиение сигнала на окна по 368 отсчетов (удвоенный квазипериод основного тона для мужского голоса средней высоты). На каждом из них вычисляется вектор признаков, связанных с относительными частотами длин полных колебаний (см., например, [11]). Ниже перечислены классы, с которыми мы работаем. Каждому из них ставится в соответствие эталон – набор описанных векторов признаков [11].

Далее, мы опираемся на автоматическую сегментацию – разбиение слова на отдельные звуки [11].

Будем обозначать гласные и твердые согласные звуки соответствующими русскими буквами в квадратных скобках. Мягкие согласные будем обозначать соответствующими латинскими буквами в квадратных скобках. Исключение для мягкого [п’], которое будет обозначаться через [@].

Занимаясь дифонным распознаванием, мы ввели понятие формализованного дифона – отрезок в 3 окна слева и 3 окна справа от метки между соседними звуками. Имя дифона – пара символов, соответствующих звукам. Мы используем также начальный полудифон – 3 окна от начала (его имя снабжается в конце символом 0) и конечный полудифон звука – 3 окна слева от заключительной метки звука.

Обозначим через (%) отрезок первого звука слова. В свое время нами замечено, что гласные звуки в начале слова хорошо распознаются между собой по начальному

полудифону (эта идея развита А. К. Бурибаевой и А.А. Шарипбаевым для начальных гласных казахских слов [12]). То же относится к звуку [ш] и классу звуков {[с], [s]}. В то же время мы обнаружили, что этот результат не распространяется на ряд звонких согласных.

Далее, глухие взрывные звуки в начале слова не выделяются при сегментации и должны распознаваться в сочетании с последующими звонкими звуками. Поэтому звукосочетания типа [па], [ki] с глухим взрывным в начале слова вынужденно распознаются по начальному полудифону. Но таких сочетаний слишком много и это слишком короткие речевые единицы для того, чтобы было обеспечено их надежное DTW-распознавание. Однако оказывается есть возможность при заранее определенном гласном распознавать между собой элементы пары (а, *Р*а), где на месте *Р* может стоять любой из звуков [к], [п], [т]. То же относится к паре (и, *Р*и), где на месте *Р* может стоять любой из звуков [k], [@], [t] и ко всем остальным парам такого вида. Отсюда вытекает необходимость заранее распознавать гласные, причем уже не по первому полудифону. При выборе участка для распознавания звуков между собой методом DTW нужно исходить из того, что DTW-распознавание, при прочих равных условиях, тем надежней, чем больше длина распознаваемого отрезка. Учитывая также, что требуется распознавание первого звука вне зависимости от влияния на него последующего звука, приходим к тому, что целесообразно использовать отрезок (*), получаемый из (%) отбрасыванием участка заключительного полудифона (последний как раз и является основным носителем упомянутого влияния).

Итак, при распознавании первого звука или его класса мы работаем со следующими объектами.

Участки 1-го звука, используемые при распознавании

(%) – отрезок всего звука в целом;
(0) – участок начального полудифона;
(*) – отрезок, получаемый из (%) отбрасыванием участка заключительного полудифона.

Результирующие классы распознавания

а, и, о, у, э, ж, ж1, ш, щ, *Р*а, *Р*е, *Р*ё, *Р*и, *Р*о, *Р*у, *Р*ы, *Р*э, *Р*ю, *Р*я, D, j, L, N, R, S, Z, Z1 (1)

Каждый из классов а,…, ж, ж1, …,щ, j состоит из всевозможных реализаций звука, соответствующего буквенному символу. Класс *Р*а состоит из всех начальных полудифонов вида Ра0, где на месте Р может стоять любой из звуков [к], [п], [т]. Аналогично устроены классы, *Р*о, *Р*у, *Р*ы, *Р*э. Класс *Р*и состоит из полудифонов вида Ри0, где на месте Р может стоять любой из звуков [k], [@], [t]. Аналогично устроены классы, *Р*е, *Р*ё, *Р*ю, *Р*я. При этом под [е], [я] понимаются только соответствующие ударные звуки. Это связано с тем, что в безударном варианте они произносятся крайне неопределенно.

Далее (речь идет о всевозможных реализациях звуков),
D– класс звуков [б], [b], [г], [g],[д], [d], [в], [v];
L– класс звуков [л], [l];
N – класс звуков [м], [m], [н], [n];
R– класс звуков [р], [r];
S – класс звуков [с], [s];
Z– класс звуков [з], [z],
Z1– класс звуков [з], [z].

Промежуточные классы распознавания

А – множество всех пар вида (а, *Р*а);
…….
Е – множество всех пар вида (е, *Р*е);
…….
Э – множество всех пар вида (э, *Р*э);
……...
Я – множество всех пар вида (я, *Р*я).

Подчеркнем еще раз, что звуки [е], [я] здесь ударные.

Отметим, что, работая с транскрипциями, можно было бы использовать информацию о твердости или мягкости начального согласного. Однако когда начальные звуки распознаются в интересах распознавания слов при ориентации только на их буквенную запись, нет необходимости различать звуки внутри классов Z, L, R, S.

Распознаватели с перечнем классов распознавания:

DTW(*): А, Е, Ё, И, О, У, Ы, Э, Ю, Я, ж, ж1, ш, щ, j, R, S D, L, N, Z, Z1

$$\begin{array}{c} DTW_А(0)\text{: а, } Р\text{а} \\ DTW_И(0)\text{: и, } Р\text{и} \\ DTW_О(0)\text{: о, } Р\text{о} \\ DTW_У(0)\text{: у, } Р\text{у} \\ DTW_Э(0)\text{: э, } Р\text{э} \end{array} \qquad (2)$$

В круглых скобках указан отрезок, по которому ведется распознавание.

Первым прорабатывает распознаватель DTW(*). Если результатом его работы является распознавание одного из классов R, S, Z, Z1, ж, ж1, ш, щ, j, то это, в соответствие с (1), –окончательный результат.

Если результатом работы распознавателя DTW(*) является распознавание одного из классов А, И, О, У, Э, то происходит переход к следующему распознавателю по правилам, указанным ниже.

Переходы в зависимости от результатов распознавания

$$А\rightarrow DTW_А(0),\ И\rightarrow DTW_И(0),\ О\rightarrow DTW_О(0),\ У\rightarrow DTW_У(0),\ Э\rightarrow DTW_Э(0),$$

Результатами работы распознавателей, стоящих в этих соотношениях справа, являются, в соответствии с (1) и (2), результирующие классы.

Далее, по аналогии с предыдущим вслед за распознаванием класса Ы должно было бы следовать распознавание по отрезку (0) с перечнем классов ы, *Р*ы. Но русское слово не может начинаться со звука [ы], так что результат *Р*ы такого распознавания известен заранее. Слова же, начинающиеся при написании с букв «е, ё, ю, я», при произнесении начинаются со звука [j], так что результат распознавания по отрезку (0) с перечнем классов е, *Р*е также заранее известен: это *Р*е. Аналогично для «ё, ю, я». Таким образом, если результатом работы распознавателя DTW(*) является распознавание одного из классов Ы, Е, Ё, Ю, Я, то это позволяет без дополнительных операций записать в качестве окончательного результата *Р*ы или *Р*е… или *Р*я соответственно.

Скажем, наконец, чем вызвано наличие двух классов ж, ж1 (и аналогично классов Z, Z1) с одинаковым звуковым составом. Очевидно, все классы распознавания привязаны к отрезкам сигнала, на которых работают соответствующие распознаватели. Введение дополнительного класса связано с недостаточно надежным выделением соответствующих шумных звуков в начала слова при сегментации. Как отмечено выше, сегментация использует численный аналог полной вариации. Звуки [ж], [з], [z] ввиду существенной шумной компоненты (артикуляция как для [ш], [с], [s] при одновременной работе голосовых связок) могут иметь участки с повышенной в сравнении с другими звонкими согласными вариацией. Из-за этого на участках этих звуков в начале слова могут возникать лишние, якобы межзвуковые, метки. В результате (*) может оказаться коротким участком в начале сигнала, слишком удаленным от реального следующего звука. В этом случае целесообразно ввести дополнительный аналог класса со своим эталоном.

Рисунок 1 представляет окно программы распознавателя. В левом верхнем поле находится список результирующих классов распознавания с указанием числа выполненных обучений. Список меняется при снятии флажка «авто» и вводе в поле «Идент. отрезка» нужного идентификатора отрезка распознавания: 0 или *. Результат распознавания выводится в нижнем поле. Кнопка «Вставить» предназначена для создания эталона класса, имя которого вводится в правом верхнем поле. Программа снабжена автоматической функцией дообучения, основанной на процедуре усреднения эталонов [11]. Пользователь запускает ее (в случае ошибки в распознавании первого звука) по копке «Добавить» при выбранной строке нужного класса и включенном флажке «усреднять». При выключенном флажке и нажатии кнопки эталон создается заново. Вместо кнопки можно использовать двойной щелчок мыши на нужной строке.

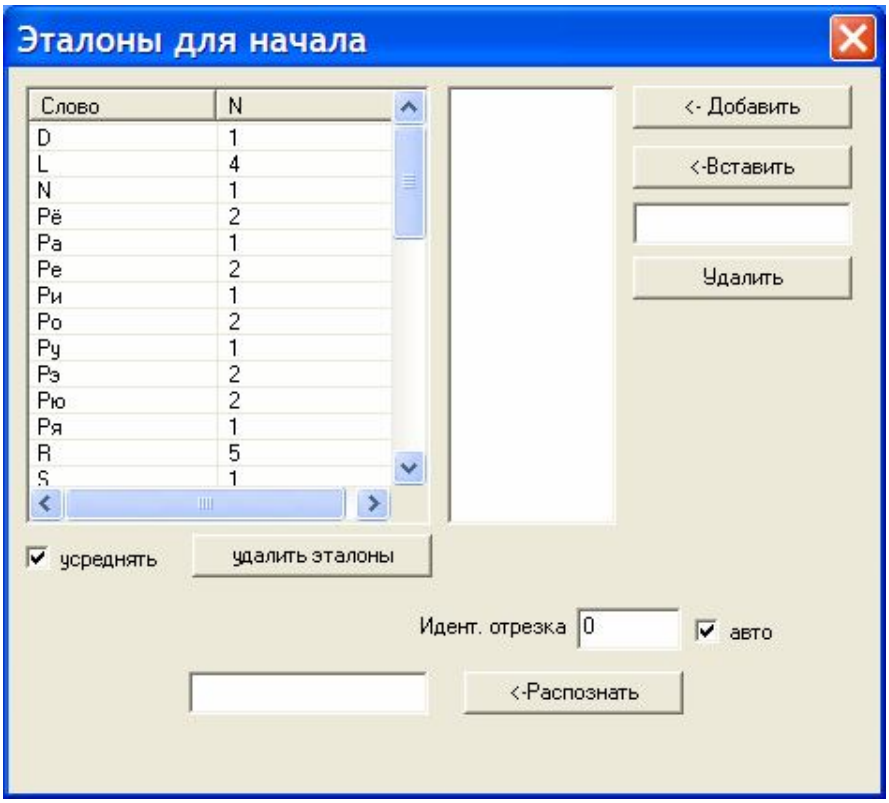

Рисунок 1 – Окно программы распознавателя

Выводы

Предложен новый метод распознавания первого звука в слитном речевом отрезке. Этот метод использует сравнение разных участков звука с эталонами на основе алгоритма DTW. Предварительное определение первого звука позволяет сильно ускорить процесс распознавания всего рассматриваемого речевого отрезка. К достоинствам данного метода следует отнести невысокую вычислительную сложность и простоту обучения. Данный метод может быть использован при распознавании слитной речи и отдельно произносимых слов с большим и сверхбольшим словарем.

Список литературы

1. Сажок Н.Н. Розпізнавання спонтанного мовлення на основі акустичних композитних моделей слів у реальному часі / В.В. Робейко, М.М. Сажок // Штучний інтелект. – 2012. – № 4. – С. 253-263.
2. Сажок Н.Н. Система устного перевода спонтанных высказываний в рамках предметных областей / Н.Н. Сажок, В.В. Яценко // Управляющие системы и машины. – 2013. – № 4. – С. 63-70.
3. Deshmukh S.D. Automatic Speech and Speaker Recognition by MFCC, HMM and Vector Quantization / S.D. Deshmukh, M.R. Bachute // International Journal of Engineering and Innovative Technology. – 2013. – Vol. 3, № 1. – P. 93-98.
4. Zarrouk E. Hybrid continuous speech recognition systems by HMM, MLP and SVM: a comparative study / E Zarrouk, Y. Ayed, F. Gargouri // International Journal of Speech Technology. – 2014. – Vol. 17, № 3. – P. 223-233.
5. Mulik V. Hidden Markov Model Based Robust Speech Recognition / V. Mulik, V. Mane, I. Jamadar. // International Journal of Innovative Research in Advanced Engineering (IJIRAE). – 2015. – Vol. 2, № 2. – P. 262-271.
6. Muda L. Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient and Dynamic Time Warping (DTW) Techniques / L. Muda, M. Begam, I. Elamvazuthi // Journal of computing. – 2010. – Vol. 2, № 3. – P. 138-143.
7. Nandyala S. P. Hybrid HMM/DTW based Speech Recognition with Kernel Adaptive Filtering Method / S.P. Nandyala, T.K. Kumar // International Journal on Computational Sciences & Applications (IJCSA). – 2014. – Vol.4, № 1. – P. 11-21.
8. Wang D. Speech segmentation without speech recognition / D. Wang, L. Lu, H. Zhang // Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). – 2003. – Vol. 1. – P. 468-471.
9. Gómez J.A. Improvements on Automatic Speech Segmentation at the Phonetic Level / J.A. Gómez, M. Calvo // Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. – 2011. – Vol. 7042. – P. 557-564.
10. Natarajan V.A. Segmentation of Continuous Speech into Consonant and Vowel Units using Formant Frequencies / V.A. Natarajan, S. Jothilakshmi // International Journal of Computer Applications. – 2012. – Vol. 56, № 15. – P. 24-27.
11. Сегментация и дифонное распознавание речевых сигналов / А.К. Бурибаева, Г.В. Дорохина, А.В. Ниценко, В.Ю. Шелепов // Тр. СПИИРАН. – Вып. 31 (2013). – С. 20-42.
12. Aigerim K. Buribayeva. Kazakh Vowel Recognition at the Beginning of Words1 / Aigerim K. Buribayeva, Altynbek A. Sharipbay // Mediterranean Journal of Social Sciences, MCSER Publishing, Rome-Italy. – April 2015. – Vol. 6, № 2. – S. 4.

References

1. Sazhok N. N. Rozpiznavannja spontannogo movlennja na osnovi akustichnih kompozitnih modelej sliv u real'nomu chasi / V. V. Robejko, M. M. Sazhok // Shtuchnij intelekt. – 2012. – № 4. – S. 253-263.
2. Sazhok N. N. Sistema ustnogo perevoda spontannyh vyskazyvanij v ramkah predmetnyh oblastej / V. V. Jacenko, N. N. Sazhok // Upravlja'uschie sistemy i mashiny. – 2013. – № 4. – S. 63-70.
3. Deshmukh S.D. Automatic Speech and Speaker Recognition by MFCC, HMM and Vector Quantization / S.D. Deshmukh, M.R. Bachute // International Journal of Engineering and Innovative Technology. – 2013. – vol. 3, №1. – P. 93-98.
4. Zarrouk E. Hybrid continuous speech recognition systems by HMM, MLP and SVM: a comparative study/ E. Zarrouk, Y. Ayed, F. Gargouri // International Journal of Speech Technology. – 2014. – Vol. 17, № 3. – p. 223-233.

5. Mulik V. Hidden Markov Model Based Robust Speech Recognition / V. Mulik, V. Mane, I. Jamadar. // International Journal of Innovative Research in Advanced Engineering (IJIRAE). – 2015. – Vol. 2, № 2. – P 262-271.
6. Muda L. Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient and Dynamic Time Warping (DTW) Techniques / L. Muda, M. Begam, I. Elamvazuthi // Journal of computing. – 2010. – vol. 2, № 3. – P. 138-143.
7. Nandyala S. P. Hybrid HMM/DTW based Speech Recognition with Kernel Adaptive Filtering Method/ S. P. Nandyala, T.K. Kumar // International Journal on Computational Sciences & Applications (IJCSA). – 2014. – vol. 4, № 1. – P. 11-21.
8. Wang D. Speech segmentation without speech recognition / D. Wang, L. Lu, H. Zhang // Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). – 2003. – vol. 1. – P.468-471.
9. Gómez J.A. Improvements on Automatic Speech Segmentation at the Phonetic Level / J.A. Gómez, M. Calvo // Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. – 2011. – Vol. 7042. – P. 557-564.
10. Natarajan V.A. Segmentation of Continuous Speech into Consonant and Vowel Units using Formant Frequencies / V.A. Natarajan, S. Jothilakshmi // International Journal of Computer Applications. – 2012. – Vol 56, № 15. – P. 24-27.
11. Segmentacija i difonnoe raspoznavanie rechevyh signalov / A. K. Buribaeva, G. V. Dorohina, A. V. Nicenko, V. Ju. Shelepov // Tr. SPIIRAN. – Vyp. 31 (2013). – S. 20-42.Aigerim K. Buribayeva. Kazakh Vowel Recognition at the Beginning of Words1 / Aigerim K. Buribayeva, Altynbek A. Sharipbay // Mediterranean Journal of Social Sciences, MCSER Publishing, Rome-Italy, Vol 6 No 2 S4, April 2015.
12. Aigerim K. Buribayeva. Kazakh Vowel Recognition at the Beginning of Words1/ Aigerim K. Buribayeva, Altynbek A. Sharipbay // Mediterranean Journal of Social Sciences, MCSER Publishing, Rome-Italy, Vol 6 No 2 S4, April 2015

RESUME

V. Ju. Shelepov, A. V. Nicenko

On Recognition of the First Sound in Continuous Speech Fragment

Background: Recent advances in speech recognition, reflected primarily in the search engines like Google and Yandex, are based on web services. At the same time, the problem of speech recognition on local computers remains relevant. Performance of recognition is still one of the big problems especially in the case of continuous speech and large vocabulary. This paper proposes a method to speed up large vocabulary Russian speech recognition procedure by preliminary recognition of the first sound or small enough class of sounds which contains it.

Materials and methods: The method uses DTW-based pattern matching algorithm for recognition of the first sound in continuous speech fragment. It differs from others since there is no previous sound to influence it. Identification of the first sound is the way to speed up the recognition of whole fragment. The problem has complete solution for vowels and sounds [j], [ж], [ш], [щ]. In others cases small enough class of sounds which contains present sound is recognized (for toneless this is pare with the next voiced sound) and thus increases recognition speed. Short line recognizers are used in general case. Every recognizer works with own part of the sound.

Results: A new method for recognition of the first sound in continuous speech fragment was proposed. This method uses the authors' automatic segmentation and pattern matching based on the dynamic time warping algorithm between the different parts of the sound and the pre-trained reference templates.

Conclusion: Preliminary recognition of the first sound allows to speed up greatly the whole speech fragment recognition process. The main advantages of this method are low computational complexity and simple training procedure. This method can be used for large and extra-large vocabulary continuous and isolated Russian speech recognition. The proposed approach is implemented in real recognition software, demonstrating high reliability.

Статья поступила в редакцию 01.07.2014.