УДК 004.934

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

СЕГМЕНТАЦИЯ РЕЧЕВОГО СИГНАЛА НА ОСНОВЕ ПРЕДПОЛОЖЕНИЯ О ЕГО ФОНЕТИЧЕСКОМ СОСТАВЕ

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

SEGMENTATION OF SPEECH SIGNAL ON THE ASSUMPTION ABOUT ITS PHONETICAL STRUCTURE

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

СЕГМЕНТАЦІЯ МОВНОГО СИГНАЛУ НА ПІДСТАВІ ПРИПУЩЕННЯ ПРО ЙОГО ФОНЕТИЧНИЙ СКЛАД

В работе предлагаются алгоритмы сегментации речевого сигнала, соответствующего русскому слову или слитно произнесенной фразе в случае, когда они известны заранее. Часть алгоритмов исправляет и уточняет априорную сегментацию. Другой алгоритм, использующий путь выравнивания при DTW-распознавании, применяется к сигналу непосредственно. В этом случае источником меток служит эталон слова или фразы, синтезируемый путем склеивания (конкатенации) эталонов дифонов. В сочетании с предварительным распознаванием это обеспечивает безусловную апостериорную сегментацию.
Ключевые слова: априорная сегментация, условная сегментация, широкая фонетическая классификация, метка, путь выравнивания, модификация дифонной базы.

The article contains algorithms for segmentation of speech signal corresponding to the beforehand known word or continuous speech phrase. One part of the algorithms corrects and qualifies the a priori segmentation. The other algorithm, using the so-called DTW equalization-path, is applied directly to the signal. In this case the source of marks is the etalon of word or phrase which is synthesized from the diphone etalons. In combination with a priori recognition we obtain unconditional a posteriori segmentation.
Key words: a priori segmentation, conditional segmentation, wide phonetic classification, mark, equalization-path, diphone-base modification.

У роботі пропонуються алгоритми сегментації мовного сигналу, що відповідає російському слову або разом виголошеній фразі в разі, коли вони відомі заздалегідь. Частина алгоритмів виправляє і уточнює апріорну сегментацію. Інший алгоритм, який використовує шлях вирівнювання при DTW-розпізнаванні, застосовується до тону безпосередньо. У цьому випадку джерелом міток служить еталон слова або фрази, що синтезується шляхом склеювання (конкатенації) еталонів дифонів. У поєднанні з попередніми розпізнаванням це забезпечує безумовну апостеріорну сегментацію.
Ключові слова: апріорна сегментація, умовна сегментація, широка фонетична класифікація, метка, шлях вирівнювання, модифікація дифонної бази.

Введение

Частично результаты данной статьи (см. раздел 1) опубликованы в работе [11]. В последние годы сегментации речевых сигналов посвящены работы [1-5]. В работах [6], [7] описаны предложенные авторами методы 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). На нее опирается развиваемый авторами метод дифонного DTW-распознавания отдельно произносимых слов или слитной речи. В качестве основного инструмента сегментации используется численный аналог полной вариации, вычисляемый для последовательных отрезков по 256 отсчетов:

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

Поскольку во всех наших системах такая сегментация выполняется сразу после записи, и предшествует всем процедурам распознавания, ее естественно называть априорной.

В противоположность ей будем называть указанную в заглавии сегментацию **условной сегментацией**, имея в виду, что она выполняется при условии соответствия рассматриваемого речевого отрезка задаваемому слову или слитно произнесенной фразе. Это понятие введено в работе [8]. Для определенности в дальнейшем будем говорить о сегментации отдельно произнесенного слова.

Распознавание речи на всех этапах, за исключением автоматического транскрибирования слов распознаваемого словаря, связано со случайными процессами, что является основным источником возможных ошибок. Это относится и к априорной сегментации. Как отмечено в [7], в большинстве случаев ошибки сегментации не влияют на результат распознавания. Однако они становятся существенными в следующей ситуации. Если сказанное слово распознано ошибочно, то пользователь может ввести в соответствующее поле правильное слово, и программа будет знать имена дифонов базы, которые нужны для построения эталона этого слова. Если сегментация будет правильной, то можно правильно автоматически выделить и прозвучавшие дифоны. В этом случае распознающую систему можно дообучить, усреднив дифоны сказанного слова и соответствующие дифоны базы. Использование модифицированных дифонов при создании эталонов слов словаря будет приводить к улучшению распознавания данного диктора. Итак, ошибочную априорную сегментацию бывает необходимо исправлять, заменяя ее условной.

1. В работе [9] предложено осуществлять условную сегментацию путем модификации априорной сегментации. Описанные там процедуры систематизируются и дополняются следующим образом.

1-1. Прежде всего, программа должна выяснить имеются или нет ошибки в априорной сегментации (в случае их отсутствия коррекция, естественно, не нужна). Для этого по введенному слову строится его транскрипция, а затем обобщенная транскрипция в терминах широкой фонетической классификации (ШФК).

Пусть для примера сказано слово «пальма» и для него получилась сегментация рис. 1:

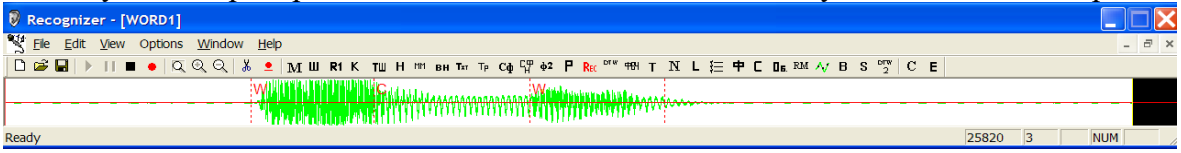

Рисунок 1 – Визуализация сигнала для слова «пальма»
с ошибочной априорной сегментацией

с сегментами

W,C,W (1)

Наш автоматический транскриптор затранскрибирует слово так: «пальма». Начальное *П* при сегментации не выделяется. Поэтому использование ШФК-транскрипции выделяет гласный, два согласных и еще один гласный звук:

WCCW (2)

Программа последовательно сопоставляет символы (2) и символы (1), идя слева направо. Она обнаруживает, что второму символу С последовательности (2) соответствует W цепочки (1). Это указывает наличие и место ошибки. Коррекция будет состоять в делении сегмента С пополам с помощью дополнительной С-метки (рис. 2):

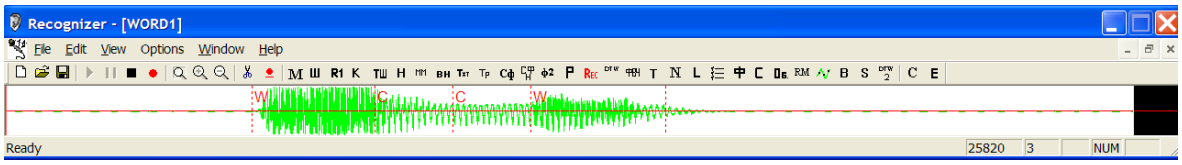

Рисунок 2 – Сегментация сигнала для слова «пальма» после коррекции

Аналогично осуществляется анализ и коррекция при обнаружении в априорной сегментации двух рядом стоящих С-сегментов там, где должен быть один (вторая С-метка убирается).

Аналогично осуществляется коррекция сегментации на участках глухих звуков. Отметим, что для *Ц* и *Ч* считаются правильными две сегментации: PF и F.

1-2. На рис. 3 представлена визуализация сигнала для слова «облом» с априорной сегментацией, содержащей гласную вставку между звуками [б] и [л].

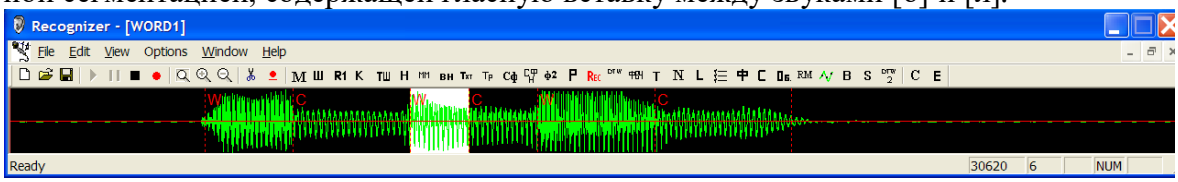

Рисунок 3 – Визуализация сигнала для слова «облом» с гласной вставкой между [б] и [л]

Программа обнаруживает ее описанным выше способом, сопоставляя цепочку

W,C,W,C,W,C

и ШФК-транскрипцию WCCWC. В результате коррекции средний W-сегмент убирается (вторая С-метка ставится на место второй W-метки).

1-3. Встречаются случаи, когда для слова, начинающегося звонким согласным, априорная сегментация ошибочно ставит в начале лишний гласный звук. Чаще всего это случается для слов, начинающихся на [з]. Пример на рис. 4:

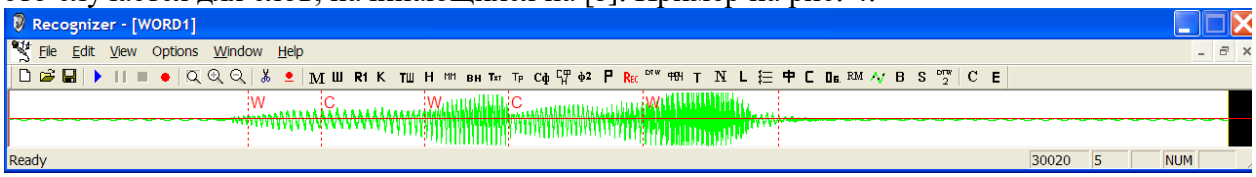

Рисунок 4 – Сегментация сигнала для слова «зима» с ошибочным гласным в начале

При коррекции маркировка начальной метки заменяется на С, а следующая С-метка убирается.

Результат на рис. 5:

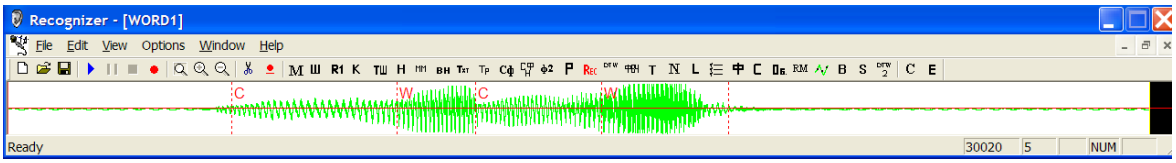

Рисунок 5 – Результат коррекции сегментации для слова «зима»

1-4. Остановимся на случае, когда при априорной сегментации не выделен С-сегмент перед глухим звуком (см. рис. 6). Наличие и местоположение этой ошибки определяется так же, как и выше. Коррекция осуществляется путем отдельной сегментации соответствующего W-отрезка; результат представлен на рис. 7. В случае если дополнительная метка при этом все же не появляется, используется искуственное разбиение отрезка «равномерными» метками: он делится на 3 равные части и последняя треть считается искомым С-сегментом.

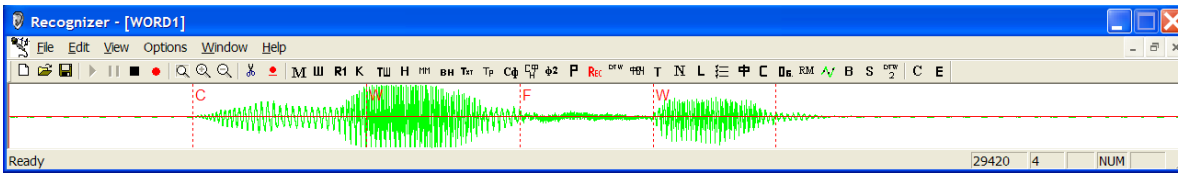

Рисунок 6 – Визуализация сигнала для слова «больше»
с ошибочной априорной сегментацией

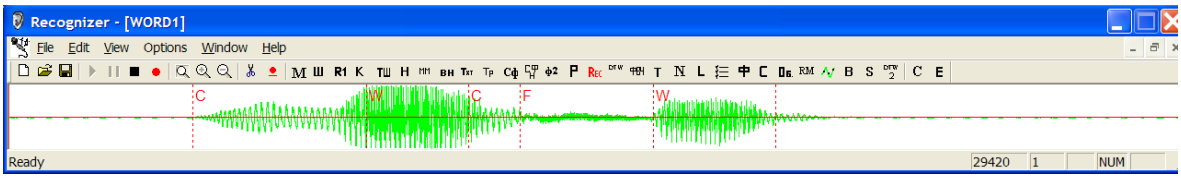

Рисунок 7 – Сегментация предыдущего сигнала после коррекции

Аналогично исчерпывается случай, когда при априорной сегментации не выделен С-сегмент после глухого звука.

1-5. В случае, когда слово заканчивается глухим звуком, но соответствующий заключительный отрезок в сегментации отсутствует, последний добавляется искусственно.

1-6. Достаточно частой является ошибка, когда при априорной сегментации не выделяется твердый или мягкий звук [р]. Это бывает, когда этот звук произносится не раскатисто, с неактивной артикуляцией. Здесь рассмотрим несколько отдельных случаев.

a) Звук [р] находится между двумя гласными и при сегментации его следует искать внутри самого длинного W-отрезка. В этом случае этот W-отрезок разбивается «равномерными» метками на 3 равные части и средняя треть выделяется как отрезок звука [р].

b) Звук [р] предшествует звонкому согласному (см. рис. 8).

В этом случае мы с помощью алгоритма, предложенного в [10], выделяем участки, соответствующие ударам языка о нёбо (р-удары, см. рис. 9) и отдельно сегментируем отрезок от первой р-метки до конца следующего С-отрезка. Результат представлен на рис. 10.

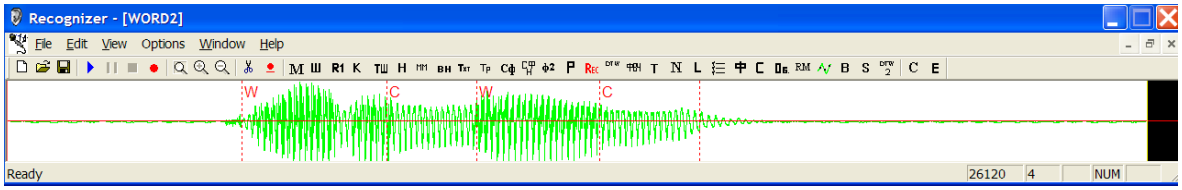

Рисунок 8 – Априорная сегментация для слова «карман» с отсутствующим сегментом [р]

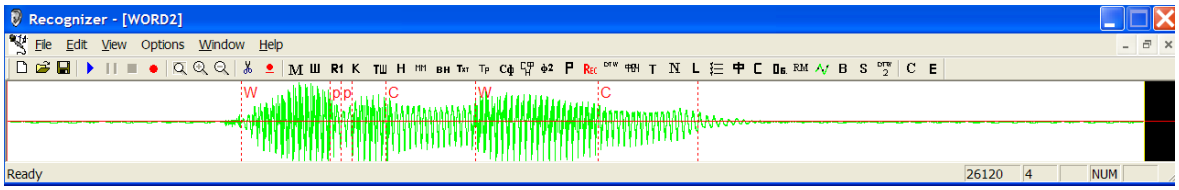

Рисунок 9 – Сегментация слова «карман» с выделением р-ударов

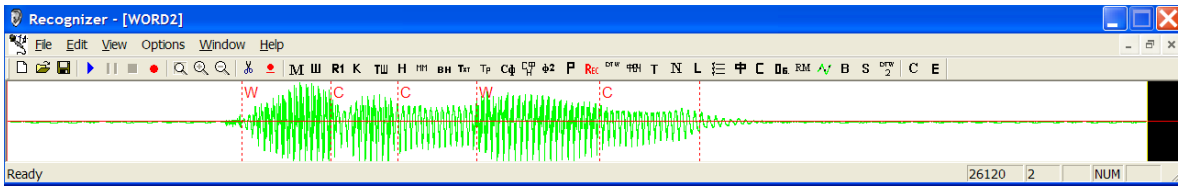

Рисунок 10 – Результат коррекции сегментации для слова «карман»

c) Звук [р] следует за звонким согласным. Этот случай исчерпывается аналогично предыдущему, только движение происходит не слева направо, а справа налево.

d) Звук [р] соседствует с глухим звуком. Здесь алгоритм коррекции такой же, как в случаях b) и c) с заменой звонкого согласного на глухой.

1-7. Случай двух рядом стоящих гласных.

Если участок, где находятся такие гласные выделен, то желаемая метка, которая будет центром соответствующего дифона, может быть получена следующим образом. Участок разбивается «равномерными» метками на 3 равные части, средняя треть удаляется, а оставшиеся две части всего сигнала склеиваются. Таким образом, в данном случае происходит не только коррекция сегментации, но и преобразование распознаваемого сигнала. Однако это преобразование находится в русле того, что мы делаем, когда производим при распознавании слова межфонемную обработку (см. [6], [7]).

Выделение участка, где находятся два соседних гласных, представляет наиболее трудную часть обсуждаемой проблемы, ибо при априорной сегментации он первоначально может как выделяться целиком, так и разбиваться на два или три разноименных отрезка. Например, участок звукосочетания *АИ* стабильно сегментируется как WC. Выделение и сегментация двух рядом стоящих гласных более успешно достигается с помощью алгоритма, описанного в следующем разделе.

2. Ниже предлагается прямой метод условной сегментации, использующий так называемый «путь выравнивания» и не связанный с модификацией априорной сегментации.

Наглядным результатом условной сегментации должны быть автоматически расставляемые метки, отделяющие в сигнале участки соседних звуков. На рис. 11 представлено окно соответствующей программы. Слово вводится в верхнее поле и сегментируется по нажатии кнопки «Усл. сегм» или пробела на клавиатуре.

На рис. 12 представлен результат выполнения этой программой условной сегментации для слова «*УСИЛИЕ*».

Источником упомянутых меток будет служить эталон слова, синтезируемый путем склеивания (конкатенации) эталонов дифонов.

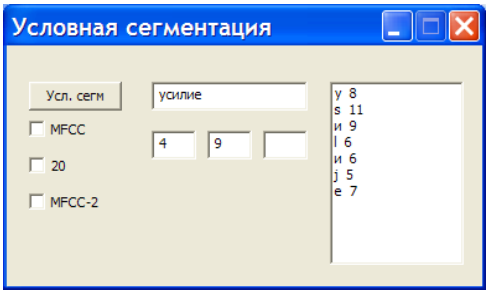

Рисунок 11 – Окно программы условной сегментации

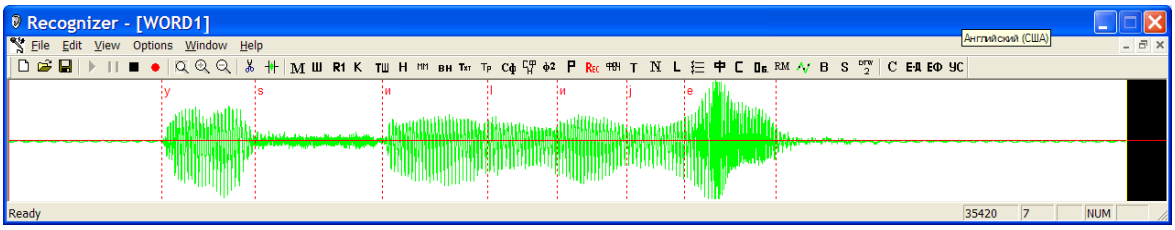

Рисунок 12 – Результат условной сегментации слова «усилие»

Эталон каждого дифона ab, где a, b – транскрипционные символы соответствующих звуков, представляет собой последовательность шести векторов признаков: три слева и три справа от границы между звуками. Эталон слова – последовательность таких шестерок, к которой добавлены начальный и конечный полудифоны. Эта последовательность строится в соответствии с транскрипцией слова. Таким образом, эталон слова – последовательность векторов признаков. Середины шестерок соответствуют границам между звуками.

При записи речевого сигнала для него непосредственно создается аналогичное параметрическое представление в виде последовательности векторов признаков. Для этого представления и вышеописанного эталона слова сроится DTW-матрица, как при DTW-распознавании (см., например, [6], [7]). Далее определяется отношение соответствия между векторами представления сказанного слова и векторами эталона (см. [6], [7]). Это соответствие графически иллюстрируется линией вида той, что представлена на рис. 13.

Будем называть ее путем выравнивания (она описывает результат растяжения-сжатия времени при DTW-распознавании).

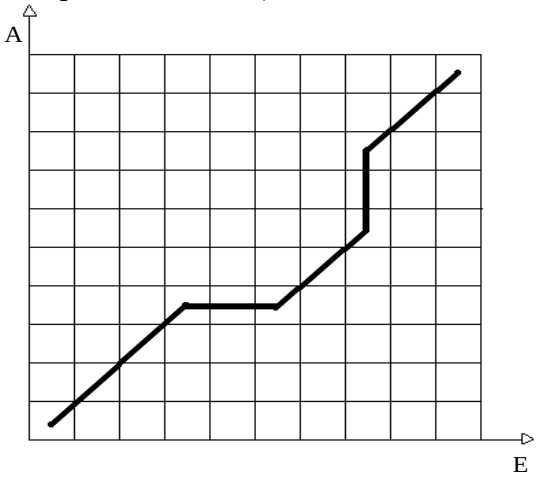

Рисунок 13 – Графическая иллюстрация соответствия между векторами

С помощью этого соответствия, зная границы между звуками в синтезированном эталоне, получаем аналогичные границы в записанном речевом сигнале. Таким образом, выполняется условная сегментация.

Подобный результат достигается более сложным способом с помощью алгоритма Витерби при распознавании на основе скрытых марковских моделей.

3. Описанная условная сегментация может применяться при коррекции дифонной базы, используемой при распознавании. В случае ошибки при распознавании слова, которая, возможно, сопровождаться ошибкой в априорной сегментации, пользователь вводит с клавиатуры правильное слово. Программа заменяет сегментацию условной и в соответствии с ней модифицирует нужные эталоны дифонов путем замены или усреднения прежних дифонов базы и вновь прозвучавших дифонов. Использование модифицированных дифонов при создании эталонов слов словаря будет приводить к улучшению распознавания данного диктора. На рис. 14 и 15 представлены ошибочная априорная сегментация и условная сегментация того же сигнала, которая является правильной.

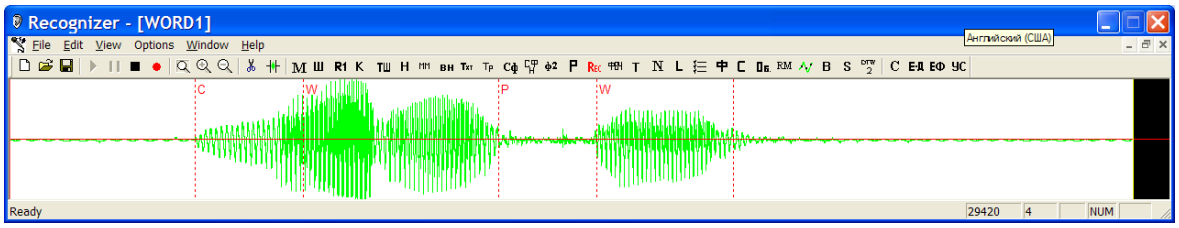

Рисунок 14 – Ошибочно отсегментированный сигнал,
соответствующий слову *МОРЯКИ*

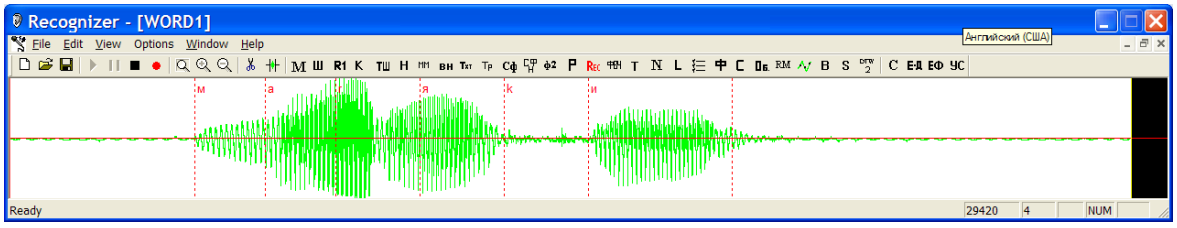

Рисунок 15 – Результат условной сегментации того же речового сигнала

Отметим, что, если ограничиваться априорной сегментацией, то при наличии в ней ошибок описанная модификация дифонной базы также будет ошибочной. Этим определяется место и значение более точной условной сегментации.

4. Программа условной сегментации, описанная в разделе 2, может быть совмещена с программой распознавания слов наперед заданного словаря. Результат распознавания сказанного слова передается программе сегментации и сигнал автоматически сегментируется с использованием пути выравнивания. В совокупности получается программа безусловной апостериорной сегментации.

Выводы

Из предложенных алгоритмов сегментации речевого сигнала, соответствующего заранее известному русскому слову или слитно произнесенной фразе предпочтительным является алгоритм, использующий путь выравнивания (раздел 2). Его сопряжение с распознавателем наперед заданного словаря обеспечивает безусловную апостериорную сегментацию.

Отметим способ контроля и улучшения условной сегментации. Эталоны дифонов для конкретного диктора создаются специальной программой, предполагающей произнесение обучающих словосочетаний вида *АБАВАГАДА,…* (см. [7]). Качество эталонов целесообразно проверить, произнося эти сочетания еще раз и наблюдая расстановку меток условной сегментации. Программа условной сегментации, описанная в разделе 2, снабжена функцией, которая позволяет в случае необходимости переместить метку с помощью мышки и, нажав клавишу «ENTER» создать вокруг полученной метки новый эталон соответствующего дифона или усреднить его с ранее существовавшим.

Список литературы

1. Mporas I. Speech segmentation using regression fusion of boundary predictions / I. Mporas, T. Ganchev and N. Fakotakis / I. Mporas // Computer Speech and Language. – 2010. – Vol. 24, № 2. – P. 273-288.
2. 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.
3. 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.
4. Automatic phonetic segmentation using boundary models / J. Yuan, N. Ryant, M. 4.Liberman [and all.] // Proceedings of Interspeech 2013. – 2013. – P. 2306-2310.
5. Patc Z. Phonetic Segmentation Using KALDI and Reduced Pronunciation Detection in Causal Czech Speech / Z. Patc, P. Mizera, P. Pollak // Text, Speech, and Dialogue. – 2015. – Vol. 9302. – P. 433-441.
6. Сегментация и дифонное распознавание речевых сигналов / А. К. Бурибаева, Г. В. Дорохина, А. В. Ниценко, В. Ю. Шелепов // Тр. СПИИРАН. – 31 (2013). – С. 20-42.
7. Шелепов В.Ю. Сегментация и дифонное распознавание речи / Шелепов В. Ю., Ниценко А. В. – Донецк : ГУ ИПИИ, 2015. – 231 с.
8. Козлов А. В. Система пофонемного распознавания отдельно произносимых слов / А. В. Козлов, Г. В. Саввина, Шелепов В. Ю. // Искусственный интеллект. – 2003. – № 1. – С. 156-165.
9. Шелепов В. Ю. О некоторых вопросах, связанных с дифонным распознаванием и распознаванием слитной речи / В. Ю. Шелепов, А. В. Ниценко, Г. В. Дорохина // Искусственный интеллект. – 2013. – № 3. – С. 209-216.
10. Шелепов В. Ю Обнаружение и выделение звука [р] в речевом сигнале / В. Ю. Шелепов, М. Х. Карабалаева, А. В. Ниценко // Искусственный интеллект. – 2011. – № 1. – С. 168-174.
11. Шелепов В. Ю. Сегментация речевого сигнала, соответствующего заранее известному слову / В. Ю. Шелепов, А. В. Ниценко // Искусственный интеллект. – 2014. – №4. – С. 202-207.
12. Шелепов В. Ю. О распознавании первого звука в слитном речевом отрезке / В. Ю. Шелепов, А. В. Ниценко // Проблемы искусственного интеллекта. – 2015. – № 0 (1). – С. 116-122.

References

1. Mporas I. Speech segmentation using regression fusion of boundary predictions / I. Mporas, T. Ganchev and N. Fakotakis / I. Mporas // Computer Speech and Language. – 2010. – Vol. 24, № 2. – P. 273-288.
2. 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.
3. 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.
4. Automatic phonetic segmentation using boundary models / J. Yuan, N. Ryant, M. 4.Liberman [and all.] // Proceedings of Interspeech 2013. – 2013. – P. 2306-2310.
5. Patc Z. Phonetic Segmentation Using KALDI and Reduced Pronunciation Detection in Causal Czech Speech / Z. Patc, P. Mizera, P. Pollak // Text, Speech, and Dialogue. – 2015. – Vol. 9302. – P. 433-441.
6. Segmentacija i difonnoe raspoznavanie rechevyh signalov / A. K. Buribaeva, G. V. Dorohina, A. V. Nicenko, V. Ju. Shelepov // Tr. SPIIRAN. – 31 (2013). – S. 20-42.

7. Shelepov V.Ju. Segmentacija i difonnoe raspoznavanie rechi / Shelepov V. Ju., Nicenko A. V. –Doneck : GU IPII, 2015. – 231 s.
8. Kozlov A. V. Sistema pofonemnogo raspoznavanija otdel'no proiznosimyh slov / A. V. Kozlov, G. V. Savvina, Shelepov V. Ju. // Iskusstvennyj intellekt. – 2003. – № 1. – S. 156-165.
9. Shelepov V. Ju. O nekotoryh voprosah, svjazannyh s difonnym raspoznavaniem i raspoznavaniem slitnoj rechi / V. Ju. Shelepov, A. V. Nicenko, G. V. Dorohina // Iskusstvennyj intellekt. – 2013. – № 3. – C. 209-216.
10. Shelepov V. Ju Obnaruzhenie i vydelenie zvuka [r] v rechevom signale / V. Ju. Shelepov, M. H. Karabalaeva, A. V. Nicenko // Iskusstvennyj intellekt. – 2011. – № 1. – S. 168-174.
11. Shelepov V. Ju. Segmentacija rechevogo signala, sootvetstvujushhego zaranee izvestnomu slovu / V. Ju. Shelepov, A. V. Nicenko // Iskusstvennyj intellekt. – 2014. – №4. – S. 202-207.
12. Shelepov V. Ju. On Recognition of the First Sound in Continuous Speech Fragment / V. Ju. Shelepov, A. V. Nicenko // Problems of Artificial Intelligence. – 2015. – № 0 (1). – P. 116-122.

RESUME

V. Ju. Shelepov, A.V. Nicenko

Segmentation of Speech Signal on the Assumption About its Phonetical Structure

Background: the article describes the segmentation of speech signal of a predetermined word or continuous speech phrase (conditional segmentation). Some recent works devoted to the segmentation of speech signals are mentioned in the literature list. Segmentation of speech fragment with beforehand known content is necessary for diphone-base modification in the case of correction of recognition error.

Materials and methods: the a priori segmentation developed by the authors on the basis of numerical analogue of the full variation is used. So-called “equalization-path” in the process of DTW-recognition is also applied.

Results: The first part of the article is the description of methods based on the authors’ a priori segmentation of any speech signal. Generalized transcription within the framework of wide phonetic classification is the control information. Algorithms are proposed for adding of wanting segments of vowels, consonants and unvoiced sounds, and eliminating of unnecessary ones. In the second part the direct conditional segmentation method is proposed. It uses the equalization-path and it is not connected with a priori segmentation modification. The third part describes application of conditional segmentation for diphone-base modification. The brief fourth part describes the unconditional a posteriori segmentation.

Conclusion: Among the proposed algorithms for segmentation of speech signal of a beforehand known word or continuous speech phrase, the preferential one is the algorithm using the equalization-path. Together with the DTW-recognition of a given vocabulary it provides the unconditional a posteriori segmentation.

Статья поступила в редакцию 05.03.2016.