УДК [004.934+004.056.5]:811.411.21

СЕГМЕНТАЦИЯ РЕЧИ НА ФОНЕТИЧЕСКИЕ ЭЛЕМЕНТЫ ДЛЯ СИСТЕМ ЗАЩИТЫ РЕЧЕВОЙ ИНФОРМАЦИИ

Е.Н. СЕЙТКУЛОВ[1], С.Н. БОРАНБАЕВ[1], А.В. ПОТАПОВИЧ[2], Г.В. ДАВЫДОВ[2]

[1]*Евразийский национальный университет им. Л.Н. Гумилева, Республика Казахстан*

[2]*Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь*

Поступила в редакцию 4 февраля 2019

Аннотация. Статья посвящена разработке алгоритма сегментации речи на фонетические элементы для синтеза речеподобных сигналов в системах защиты речевой информации. Основное внимание уделяется установлению границ фонетических единиц речи с учетом влияния этого фактора на качество синтезируемой речи компиляционным методом. Рассматриваются особенности установления границ фонем для слитной речи и влияние этого фактора на качество синтезируемой речи по базе фонем. Предлагается для обеспечения качественной синтезируемой речи начало и окончание фонем при сегментации устанавливать при переходе реализации сигнала через ноль, а при синтезе речеподобных сигналов использовать сплайн-функции на границах сегментов фонем.

Ключевые слова: сегментация речи, границы фонем, речеподобные сигналы, синтез, сплайн-функции.

Abstract. The article is devoted to the development of speech segmentation algorithm on phonetic elements for the synthesis of speech-like signals in speech information protection systems. The main attention is paid to establishing the boundaries of phonetic units of speech, taking into account the influence of this factor on the quality of the synthesized speech by the compilation method. The features of establishing the boundaries of phonemes for fused speech and the influence of this factor on the quality of synthesized speech on the basis of phonemes are considered. It is proposed to ensure the quality of synthesized speech beginning and ending phonemes at the segmentation set in the transition implementation of a signal through zero and in the synthesis of speech-like signals to use the spline function at the boundaries of segments phonemes.

Keywords: speech segmentation, phoneme boundaries, speech-like signals, synthesis, spline functions.

Doklady BGUIR. 2019, Vol. 123, No. 5, pp. 66-71
Segmentation of speech on phonetic elements for systems of speech information protection
Y.N. Seitkulov, S.N. Boranbayev, A.V. Patapovich, H.V. Davydau

DOI: http://dx.doi.org/10.35596/1729-7648-2019-123-5-66-71

Введение

Сегментация речи может выполняться методами, основанными на использовании априорной информации о сегментируемом речевом сигнале [1–4], и методами, не использующими сведений о сегментируемом сигнале, и предназначенными в большинстве случаев для распознавания речи, верификации диктора по голосу, распознавания языка, на котором говорит диктор [5–9]. Выбор метода сегментации речи определяется конечной целью. Вместе с тем в методах сегментации речи можно выделить методы сегментации на фонетические элементы, что необходимо как для задач распознавания речи, так и синтеза речи и речеподобных сигналов. Однако если при сегментации речи на такие структурные элементы, как предложения, фоноабзацы, слова, границы сегмента можно установить достаточно точно вне зависимости от эксперта при ручной сегментации и используя методы

автоматической сегментации, то при сегментации речи на фонемные элементы границы этих элементов установить весьма сложно. Эти границы будут определяться конкретными задачами и целями сегментации речи. Если сегментация речи на фонемные элементы выполняется с целью распознавания речи или решения задач, связанных с процессами распознавания, то границы сегмента для фонетического элемента следует устанавливать таким образом, чтобы повысить степень правильного определения фонемного образа вне зависимости от акустического звучания определяемого объекта. С другой стороны, границы фонемного структурного элемента речи при ее сегментации с целью получения базы фонемных структурных элементов речи для преобразования текста в речь или синтеза речи или речеподобных последовательностей будут совсем другими. Они должны обеспечивать высокое качество синтезируемой речи (синтезируемая речь не должна иметь металлический оттенок и по звучанию должна быть как можно ближе к голосу определенного диктора, из речи которого формировалась база фонемных структурных единиц речи). Одновременно с этим большое влияние на качество синтезируемой речи оказывает метод формирования просодики речи с учетом индивидуальных особенностей говорящего.

В системах активной защиты речевой информации в помещениях для переговоров в качестве маскирующих сигналов часто используются комбинированные маскирующие сигналы, состоящие из «белого» шума и речеподобных сигналов [10, 11]. При этом рекомендуется использовать в качестве речеподобных сигналов речевые последовательности, сформированные с учетом лингвистических особенностей языка и статистических характеристик встречаемости фонем в данном языке, а также длины слов и предложений. Формирование речеподобных сигналов выполняется компиляционным методом по базе структурных единиц речи. В результате сформированные таким образом речеподобные сигналы сохраняют все оттенки речи определенного диктора, и их весьма сложно отличить от информационных сигналов этого же диктора.

Речевой аппарат человека устроен таким образом, что нельзя установить его форму и динамику при произношении слитной речи. Одни фонемы переходят в другие без четко выраженных границ. При переходе от одной фонемы к другой речевой аппарат должен перестроиться и занять такое положение, при котором можно сформировать последующую фонему, поэтому установить точную границу перехода от одной фонемной структуры к другой невозможно. Более точно эта граница будет определяться последующими задачами, для решения которых и выполняется сегментация речи на фонемные структурные элементы.

При синтезе речеподобных сигналов компиляционным методом по базам фонем не всегда обеспечивается высокое качество синтезируемой речи, хотя применяются методы формирования просодики. Повысить качество синтезируемой речи при этом можно, используя экспоненциальные сплайн функции на границах перехода от одной фонемной структуры к другой и наложении при этом окончания одной фонемной структуры на начало второй фонемной структуры. При этом будет происходить некоторое более быстрое затухание амплитуд колебаний окончания одной фонемной структуры и увеличение амплитуды начала второй фонемной структуры. Такой механизм компиляционного синтеза речи позволит устранить скачки сигнала на границах фонемных структур. Для выполнения такого компиляционного синтеза речи нужна база фонемных структурных единиц речи с несколько увеличенными сегментами во временной области, так как при синтезе происходит наложение окончания одной фонемной структуры на начало второй фонемной структуры.

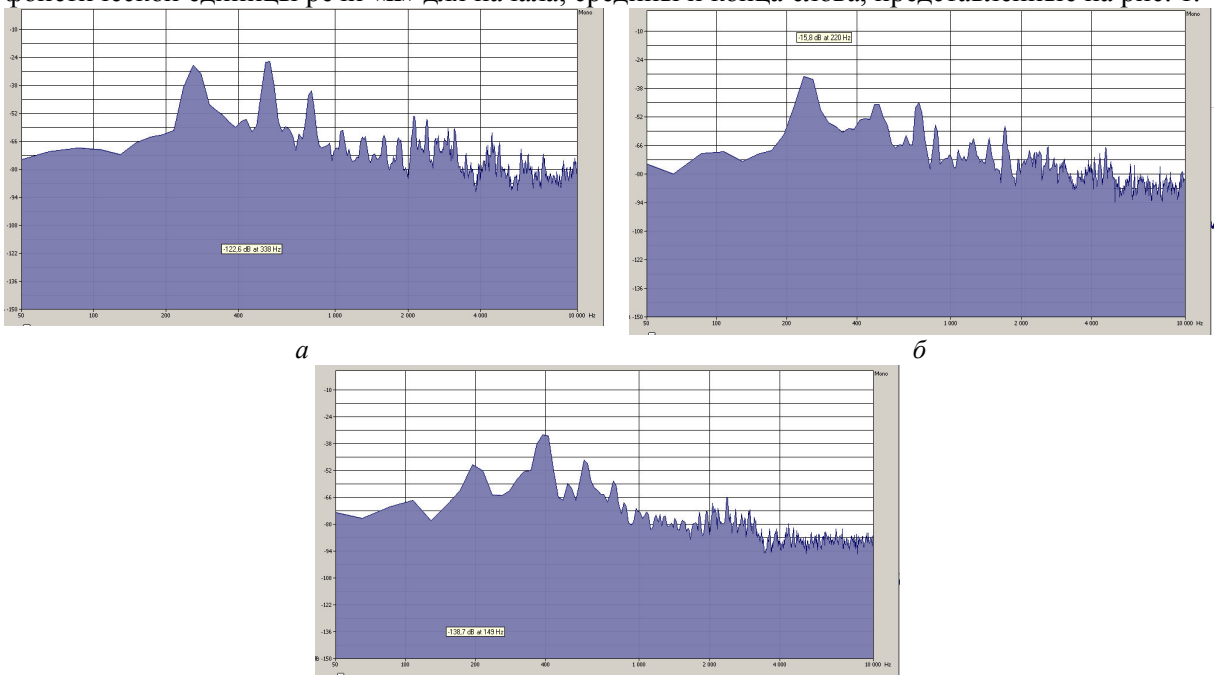
Анализ методов сегментации речи показал, что для формирования базы фонемных структурных единиц речи для синтеза речи компиляционным методом наиболее удобным является метод сегментации, использующий динамическое программирование. Для этого необходимо иметь фонетическую запись слитной речи, размеченной вручную на фонемные структурные элементы и содержащей все фонемные структурные единицы, необходимые для базы. Обычно это 300–400 аллофонов для русской, казахской, белорусской речи и около 1200 фонемных структурных единиц речи для китайского языка, так как он является тональным.

Требования к эталонной речи и разметке ее на фонемные структурные единицы вручную

Эталонная речь, предназначенная для сегментации ручным способом и создания по ней эталонной базы фонетических структурных элементов, должна представлять собой слитную речь. Кроме того, в эталонной речи должны присутствовать все фонетические элементы, необходимые для создания базы. В работе В. Н. Сорокина [12] указывается, что для сегментации речевого сигнала необходимо выполнять поиск границ квазистационарных и переходных процессов. Наиболее четко на фонограммах видны участки с гласными звуками.

Гласные звуки речи распределяются по той части языка, которая поднята при произношении данного звука. Движение языка по горизонтали приводит к образованию гласных трех рядов: переднего, среднего и заднего. Кроме того, они различаются по степени приподнятости той или иной части языка: гласные верхнего подъема, среднего и нижнего. Вместе с тем они делятся на губные, т. е. при образовании которых принимают участие губы (это, например, «**о**»), и неогубленные, т. е. при образовании которых губы не принимают участия.

Из-за эффектов редукции и коартикуляции точно установить границы фонем весьма трудно. Коартикуляция, это когда согласная фонема в значительной степени приобретает окраску последующей гласной фонемы, а гласная фонема в значительной степени приобретает окраску предшествующей согласной. Подтверждением этого являются спектрограммы фонетической единицы речи «**li**» для начала, средины и конца слова, представленные на рис. 1.



а *б* *в*

Рис. 1. Спектрограммы фонетической единицы речи «**li**»: *а* – для средины слова; *б* – для начала слова; *в* – для конца слова

Как видно из представленных на рис. 1 спектрограмм, в зависимости от положения в слове фонетической единицы речи «**li**» имеются существенные различия как по частотам формант, так и по их амплитудам. Это указывает на то, что при формировании базы фонемных структурных единиц речи для синтеза речеподобных сигналов и синтеза речи в базу необходимо включать для одной фонемной единицы ее фонетические реализации в зависимости от ее окружения. Необходимо учитывать предшествовавший фонетический элемент и последующий.

В большинстве работ по сегментации речи алгоритм сегментации начинается с определения энергетических параметров сегментов и их резких изменений, а далее происходит вычисление спектральных составляющих сигнала и нахождение участков его изменения, что обычно связывается с изменением артикуляции. [2, 5, 6]. Далее может проводиться анализ статических характеристик сегмента путем усреднения его спектра

или сравнения спектров в соседних кадрах, вычисления кепстральных коэффициентов. Предлагается для нахождения границ фонем использовать различия фаз в соседних кадрах для разных частотных областей.

Как показал анализ структуры фонетических единиц слитной речи, установить четко и однозначно границы фонем невозможно, так как есть участок, на котором фонема четко видна, и есть переходной участок. Эти участки имеются как перед фонемой, так и после нее. В зависимости от окружения длина переходных участков может быть различной. На рис. 2 представлена временная реализация фонетической структуры «**gali**», взятой из слитной речи русскоязычного диктора.

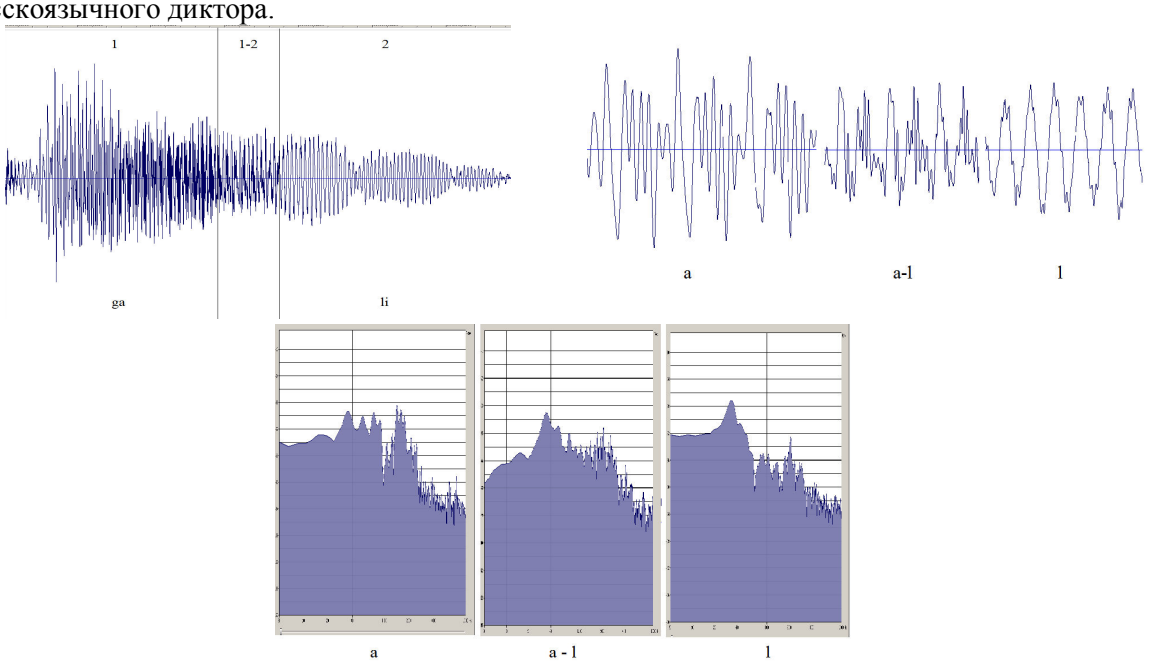


Рис. 2. Временная реализация фонетической структуры «**gali**»

На этой реализации были выделены три временных участка: участок 1 – это фонетическая структура (слог) «**ga**»; участок 2 – фонетическая структура «**li**»; участок 1–2 это переходной участок фонетической структуры «**ga**» в фонетическую структуру «**li**». Чуть ниже показаны участки реализаций с длительностями по 17 мс каждая: a – участок фонемы «**a**»; a-l – участок перехода фонемы «**a**» в фонему «**l**»; l – участок реализации фонемы «**l**». На участке реализации фонемы «**a**» видны четко повторяющиеся по форме три участка. Эти участки мы назвали доменами. Такие же домены можно отметить на переходном участке фонемы «**a**» в фонему «**l**» и на участке реализации фонемы. Их количество остается равным трем при длине реализации 17 мс. Это значит, что длина домена составляла 5,65 мс. Таким образом, можно сделать вывод о том, что частота основного тона была около 177 Гц. Практика показала, что при ручной сегментации, используя информацию об изменении формы доменов, можно достаточно точно устанавливать границы переходной области одной фонемы в другую. Для приведенной на рисунке реализации длина участка перехода фонемы «**a**» в фонему «**l**» составила 38,7 мс. На нижней части этого же рисунка приведены спектры для соответствующих участков фонем и переходной области. Следует отметить, что, используя в вейвлет преобразованиях (DWT) в качестве базовых функций форму доменов, можно с более высокой точностью определять границы фонем. Как показала практика, форма отдельных доменов меньше зависит от языка, а в большей степени от диктора.

Для других 30 дикторов величина участка перехода от фонемы «**a**» к фонеме «**l**» для фонетической структурной единицы «**gali**» лежала в диапазоне от 25 до 45 мс. Величина переходных участков между другими сочетаниями фонем может быть другой и колебаться от 11 до 50 мс. Это определяется темпом речи диктора, четкостью произношения и особенностями артикуляционного тракта. При сегментации речи на фонемы или аллофоны необходимо в ее длительность включать и переходные участки, т. е. переходной участок, предшествующий данной фонеме, участок, непосредственно соответствующий фонеме,

и переходной участок после фонемы – переход к другой фонеме. Так, было предложено базу аллофонов для синтеза речи по тексту для русского языка формировать из 498–552 элементов, а также систему индексов для обозначения аллофонов при автоматическом синтезе речи. При создании базы мультифонов число элементов увеличивалось от 6000 до 7000 [13]. Для синтеза речеподобных сигналов база аллофонов может быть сокращена до 320 элементов для русской и 342 для белорусской речи. Для казахского языка укороченная база аллофонов может быть сформирована из 245 элементов. При создании укороченных баз фонетических структурных единиц речи, предназначенных для синтеза речеподобных сигналов, в состав базы не включались элементы, имеющие малую вероятность их появления в речи на данном языке.

Требования к эталонным текстам для ручной сегментации заключаются в том, что они должны содержать все элементы базы из наиболее часто употребляемых слов в данном языке. Кроме того, речь должна носить слитный характер.

Сплайн функции при синтезе речеподобных сигналов

Так как фонетические базы структурных элементов речи вначале и в конце содержат переходные области, то при синтезе речи следует использовать сплайны для переходных областей. Рекомендуется использовать так называемое «сшивание» аллофонов при компиляционном синтезе речи. Переходной участок окончания предшествующего аллофона, умноженный на убывающую функцию, изменяющуюся от 1 до 0, накладывается на переходной участок последующего аллофона, умноженный на возрастающую функцию от 0 до 1. Если длины накладываемых друг на друга переходных участков не равны, то длина формируемой переходной области выбирается равной длине более длительного переходного участка. Недостающая часть меньшего переходного участка дополняется нулевыми значениями амплитуды.

В работе [13] предлагается выполнять умножение переходных участков на линейные функции, изменяющиеся от 1 до 0 и от 0 до 1. Однако в связи с тем, что чувствительность слуха имеет нелинейный характер, более эффективно применение сплайн функций более высокого порядка (до третьей степени).

Заключение

Установлено, что величина переходных участков одной фонемы в другую в сильной степени зависит от характерного темпа произношения слитной речи для конкретного диктора, вида соединяемых фонем и составляет от 10 до 50 мс.

Работа выполнена при поддержке грантового финансирования КН МОН РК, №AP 05130293.

Список литературы / References

1. Sakoe H., Chiba S. Dynamic Programming Algorithm Optimization for Spoken Word Recognition // IEEE Transactions on Acoustics, Speech, and Signal Processing. 1978. Vol. ASSP-26, No. 1. P. 43–49.
2. Scharenborg O., Wan V., Ernestus M. Unsupervised speech segmentation: An analysis of the hypothesized phone boundaries // The Journal of the Acoustical Society of America. 2010. Vol. 127, No. 2. P. 1084–1095.
3. Gomez J.A., Calvo M. Improvements on automatic speech segmentation at the phonetic level // Materials of 16th Iberoamerican CongressProgress in Pattern Recognition, Image Analysis, Computer Vision and Applikations. 2011. P. 557–564.
4. Bemdt D.J., Clifford J. Using Dynamic Time Warping to FindPatterns in Time Series // AAAI Proc. knowledge discovery in databases. 1994. P. 359–370.
5. A Review: Automatic Speech Segmentation / Sakran A.E. [et al.] // International Jornal of Computer Science and Mobile Computing. 2017. Vol. 6, No. 4. P. 308–315.
6. Makowski R., Hossa R. Automatic speech signal segmentation based on the innovation adaptive filter // International Journal of Applied Mathematics and Computer Science. 2014. Vol. 24, No. 2. P. 259–270.
7. Kamarauskas J. Automatic Segmetation of Phonemes using Artificial Neural Networks // Elektronika ir Elektrotechnika. 2006. Vol. 72, No. 8. P. 39–42.
8. Automatic Silence/Unvoiced/Voiced Classification of Bangla Velar Phonemes: New Approach / Syed Akhter Hossain [et al.] // 8th ICCIT. Dhaka, 2005.

9. Highly accurate phonetic segmentation using boundary correction models and system fusion / A. Stolcke [et al.] // 2014 IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP). IEEE, 2014. P. 5552–5556.
10. Method for protecting speech information / H.V. Davydau [et al.] // Doklady BGUIR. 2015. № 8 (94). P. 107–110.
11. Rationale for the method of formation of the combined speech masking signals / Y. Seitkulov [et al.] // IEEE 8th International Conference on Application on Information and Communication Technologies (AICT). Astana, Kazakhstan, 2014.
12. Sorokin V.N. Segmentation of the period of the fundamental tone of a voice source // Acoustical Physics. 2016. Vol. 62, No. 2. P. 244–254.
13. Algoritym of forming speech base units using the method of dynamic programming / Seitkulov Y.N. [et al.] // Journal of Theoretical and Applied Information Technology. 2018. Vol. 96, No 23. P. 7928–7941.

Сведения об авторах

Сейткулов Е.Н., к.ф.-м.н., директор НИИ информационной безопасности и криптологии Евразийского национального университета им. Л.Н. Гумилева.

Боранбаев С.Н., д.т.н., профессор Евразийского национального университета им. Л.Н. Гумилева.

Потапович А.В., с.н.с. НИЛ 5.3 НИЧ Белорусского государственного университета информатики и радиоэлектроники.

Давыдов Г.В., к.т.н., ведущий научный сотрудник НИЛ 5.3 НИЧ Белорусского государственного университета информатики и радиоэлектроники.

Information about the authors

Seitkulov Y.N., PhD, director of the institute of information security and cryptology of Eurasian national university named after L.N. Gumilyov.

Boranbayev S.N., D.Sci, professor of Eurasian national university named after L.N. Gumilyov.

Patapovich A.V., researcher of SRL 5.3 of R&D department of Belarusian state university of informatics and radioelectronics.

Davydau H.V., PhD, researcher of SRL 5.3 of R&D department of Belarusian state university of informatics and radioelectronics.

Адрес для корреспонденции

220013, Республика Беларусь,
г. Минск, ул. П. Бровки, 6,
Белорусский государственный университет
информатики и радиоэлектроники
тел. +375-29-670-30-40;
e-mail: nil53@bsuir.edu.by.by
Потапович Александр Владимирович

Address for correspondence

20013, Republic of Belarus,
Minsk, P. Brovka st., 6,
Belarusian state university
of informatics and radioelectronics
tel. +375-29-670-30-40;
e-mail: nil53@bsuir.edu.by.by
Patapovich Aleksandr Vladimirovich