Algorithm of Allophone Borders Correction in Automatic Segmentation of Acoustic Units

Janusz Rafalko

Algorithm of allophone borders correction in automatic segmentation of acoustic units

Janusz Rafałko

Faculty of Mathematics and Information Science, Warsaw University of Technology, Poland
j.rafalko[at]mini.pw.edu.pl

Keywords: Speech, speech synthesis, allophone, phoneme, text to speech, tts, acoustic units, speech processing, border correction

Abstract. In concatenative speech synthesis the fundamental factor with heavy influence on synthesized speech quality is the database of acoustic units. In case of bases received in automatic way, the key matter is suitable marking the borders of acoustic units. This article describes the algorithm of correction of acoustic units borders appointive in automatic way. It is based on two factors specified and tested here. It also describes worked out method of grade of acoustic units database, which allows to observe the influence of introduced correction on the base quality.

1 Introduction

The main goal in speech technology is to create a speech which is almost as real as a voice of a living person. One of the methods of speech synthesis based on the text (Text to Speech), which allows to reproduce the human personal speech characteristics is a concatenation method. It uses small and natural acoustic units, from which the speech is synthesized. These can be allophones, diphones or syllables. In presented system, synthesizer is based on bases composed of allophones. That type of system synthesizes the speech by joining the acoustic units in accordance to appropriate phonetic rules. The individual features of human voice are not included in this rules, but only in natural acoustic units. In order to synthesize the voice of a particular man, an acoustic units database must be created.

The acoustic units databases are the allophones bases. Preparing such base takes months and it is done by experts in a manual way. Therefore there is a demand to work out an automatic method of creating such bases e.g. presented in [1, 2]. As a result of working the system in automatic way, the borders of allophones are marked. In order to obtain the best quality of synthesized speech, the borders have to be marked in a precise way. Introduced in this paper algorithms allow to correct errors of marking these borders.

2 Acoustic units database

Different approaches to the speech synthesis from the text are described in [3, 4]. The basic feature of concatenative approach of speech synthesis is the use of elementary pieces of natural speech [5]. In the synthesizer, the signal compiled from natural speech segments is a subject of further modification which changes the prosodic parameters of the signal.
In the [6], study shows the basic assumptions of concatenative TTS system for Polish language, based on allophones in the context of multilingual synthesis. Natural elements from which the speech is synthesized may be allophones, diphones, multiphones as well as syllables. In concatenative method of speech synthesizing these type of basic speech units has much influence on obtaining individual speech characteristics. This paper refers to the databases of acoustic units, which include several context groups of particular phoneme, which may be identified with acoustic allophone, described in the study of Jassem [7]. The advantages of the choice of allophones as a basic units [8, 9, 10, 11] base on the fact that firstly - speech units remain the effects of sounds interference, and secondly - the number of basic units is relatively low and holds in the range of 400 - 2000 in different systems. The difficulty posing in this approach is a necessity of precise allophones marking during the segmentation of natural speech signal.

3 Automatic segmentation of natural speech signal

If in the speech synthesis the compilation elements contain only the phonetical and acoustical characteristics, the segmentation task is about to “cut” the basic segments from the speech stream and place them into the database. The main stages of this algorithms include:

1. Selection and preparation of text and acoustic corpuses.
2. The automation of creation of acoustic units databases of the particular speaker voice.

As a result of this work system, we get many of the same units, but we need only one piece of each element to our base. In order to create an acoustic database it is necessary to analyze received units in details and delete those phonetic units in which the acceptable error during reading or automatic segmentation was exceeded and save only the best of them. If there are more than one identical allophones, we choose the best one. Finally we must perform the control of quality of each element that left, marking deviations and perform the correction of segments parameters with noticed deviations.

4 Correction algorithm

The "correction" operation is performed for those segments, which are obtained in accordance to the ways mentioned above. The segments with inaccurately defined limits are subject to correction using proper procedures, involving removing the inaccurate periods of basic tone and inserting the missing terminal periods of basic tone. The diagnosis of determining the limits of units is performed by determining the level of time between periods and acoustic signal characteristics similarities on the first and the second period of basic tone, as well as on the penultimate and the last one. Both cases of periods: first, second and penultimate, the last, will be described as terminal and pre-terminal. The correction is performed only in the case of voiced units.
To determine the level of similarity of terminal and pre-terminal period of acoustic characteristics, the formula 1a) is used. The distance between time characteristics is calculated as a ratio of duration of these periods (formula 1b).

$$\text{a),}\ L_A = 1 - \frac{\sum_{i=1}^{N} \left| s_i^G - s_i^P \right|}{\sum_{i=1}^{N} \left(\left| s_i^G \right| + \left| s_i^P \right| \right)} \qquad \text{b),}\ L_T = 1 - \frac{\min\left(T^G, T^P\right)}{\max\left(T^G, T^P\right)} \tag{1}$$

where:

s_i^G – value of the signal at the "i" segment of terminal period
s_i^P – value of the signal at the "i" segment of pre-terminal period
T^G, T^P – duration of periods, both the terminal and the pre-terminal

These factors assume value of range 0 – 1. The factor of the time adjustment L_T is constructed in such a way, that it takes low value for similar duration of terminal and pre-terminal period of acoustic unit, achieving value 0 when these periods are identical. When the durations differ from each other, this factor grows up to value 1. On the contrary factor L_A vice versa, similar periods - value close to 1.
"Correction" consists of removing terminal period and duplicating pre-terminal one. In result the number of periods of segment basic tone does not change. The second case of "correction" is an absolute rejection of terminal period. Whereas in case of the correct marking of border unit, such period remains. In order to find the appropriate values of the factors by which terminal period should be remove or improve by replacing them with pre-terminal, it was analyzed various bases received as a result of automatic segmentation.
It was set experimentally, that when factor $L_T < 0{,}2$ it means, that terminal and pre-terminal periods have similar durations and terminal period might be left without changes. Example of such situation is presented on fig. 1.

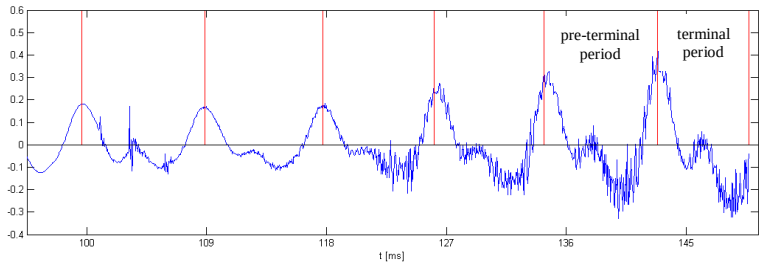

Fig. 1. Terminal and pre-terminal periods of "wi03" unit, $L_T = 0.19$.

If value of factor L_T is between 0.2 and 0.7 terminal period should be replaced by pre-terminal. Example of such situation is presented on next fig. 2. We can notice here that border was marked in mid-period, therefore it should be replaced by pre-terminal period.

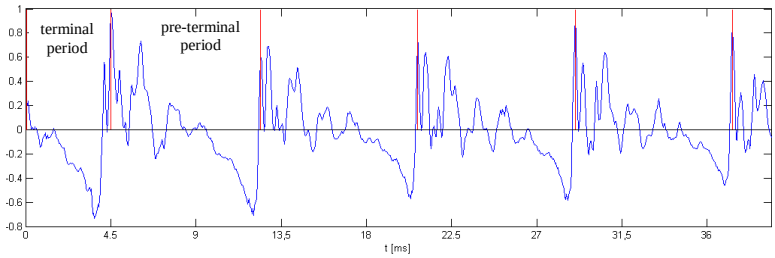

Fig. 2. Terminal and pre-terminal periods of "a0130" unit, $L_T = 0.44$.

Third case refers to factor $L_T > 0.7$. It means that terminal period is too short. Fig. 3 shows that the border of unit is marked just behind the border of pre-terminal period. In such cases terminal period is removed.

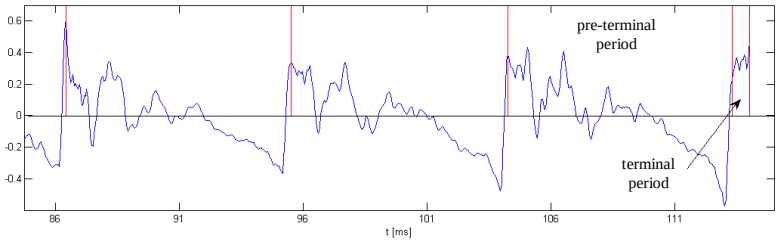

Fig. 3. Terminal and pre-terminal periods of "e1001" unit, $L_T = 0.92$.

In case of factor L_A we can also distinguish three cases. First of them, when L_A is larger ($L_A > 0.6$) means, that terminal and pre-terminal periods are similar in acoustic consideration to each other. In such case operation depends on ad valorem factor L_T. Examples are presented on previous fig. 3 illustrates case, when $L_A = 0.89$ but never-

theless, terminal period is removed because of high value of factor L_T. On fig. 2 terminal period is replaced at high value $L_A = 0.72$. In turn fig. 1 shows the case when both factors have very good value ($L_A = 0.79$, $L_T = 0.19$) and terminal period remains firm.
When L_A is in range $0.4 < L_A < 0.6$ terminal period is replaced by pre-terminal period. We can see it on fig. 4. The value $L_A = 0.57$ means that period will be replaced by pre-terminal. We can see (fig. 4) that terminal period is different from other periods of allophone unit therefore it should be replaced.

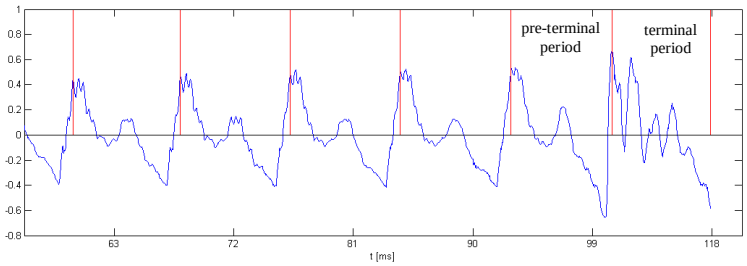

Fig. 4. Terminal and pre-terminal periods of "n03" unit, $L_A = 0.57$.

If $L_A < 0.4$ terminal period should be removed without any regard to value of factor L_T. Example is presented on next fig. 5. We can see that the algorithm of border marking is marked incorrectly in the last period.

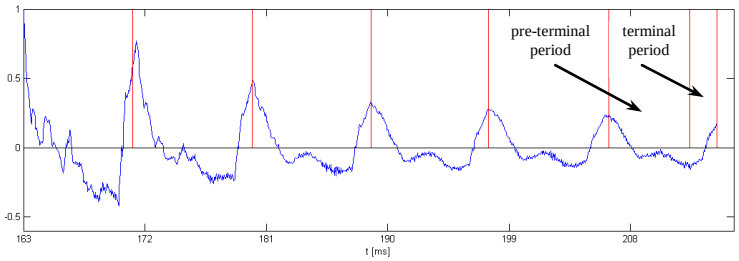

Fig. 5. Terminal and pre-terminal periods of "l02" unit, $L_A = 0.06$.

The analysis of all those cases for various bases led to defining required values of these factors and to determining when and how an allophone is a subject to a correction.
If:

- $L_T < 0.2$ and $L_A > 0.6$ – terminal period remains firm;
- $0.2 < L_T < 0.7$ and $L_A > 0.4$ – terminal period is replaced by pre-terminal;
- $L_T > 0.7$ or $L_A < 0.4$ – terminal period is removed;

The first and the third cases are trivial. The replacement of terminal period by pre-terminal is a second case of correction. Fig. 6 presents the allophone before correction

of primary border period and the same unit after correction. The correction coefficients of initial border periods of this element are $L_T = 0.3$ and $L_A = 0.63$.

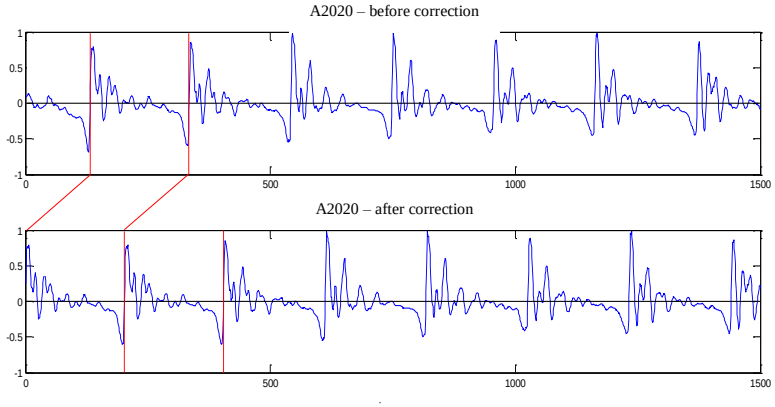

Fig. 6. Unit "a2020" before and after replacement terminal period by pre-terminal period.

5 The base correctness ratio

Taking into account parameters L_T and L_A it is possible to construct the ratio showing the correctness of cutting an acoustic unit, and after averaging - the ratio showing the correctness of whole base. Obviously, it will be showing only correctness of cutting voiced units. We can introduce the summary degree of similarity of terminal periods appointed as:

$$L = \frac{L_A + L_T}{2} \tag{2}$$

Factors L_A and L_T are constructed in such a way, that for correctly cut allophone L_A the value will be close to 1, but L_T close to 0. So total ratio for correctly cut segments should have the value approximate to 0.5. Upon introducing the average for this grade of similarity for all units in base we can receive the base correctness ratio:

$$L_{BCR} = \frac{1}{N}\sum_{i=1}^{N}\frac{L_A^i + L_T^i}{2} \tag{3}$$

Fig. 7 presents the ratio L histogram of acoustic units of voice standard base prepared manually. The average value that is the ratio of correctness for this base is $L_{BCR} = 0{,}48$. We can see that this ratio distribution is similar to Gaussian distribution.

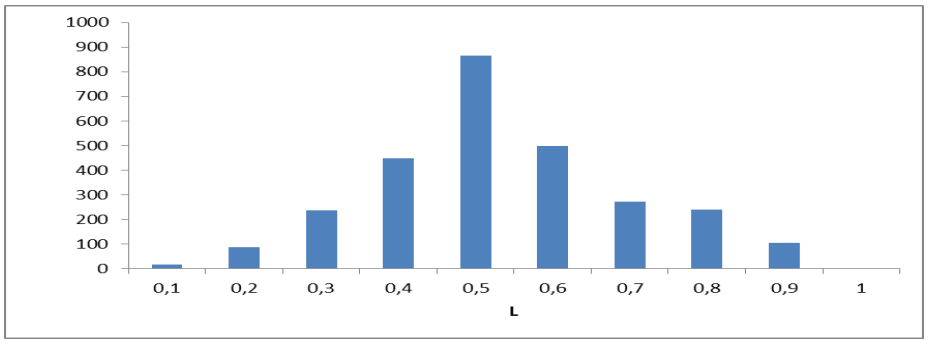

Fig. 7. Histogram of the factor L of the standard base.

Next graphs show histograms of ratio L before correction (fig. 8) and after correction (fig. 9). It is a base obtained in automatic way, for the same voice that standard base. In this case the base correctness ratio is about $L_{BCR} = 0.51$. There can be many reasons why the value of ratio L had deviated from 0.5 for the unit. The reason might be incorrectly marked borders of periods by implemented algorithm. However basic cause is the fact that terminal and pre-terminal periods in allophone need not and in practice cannot be identical.

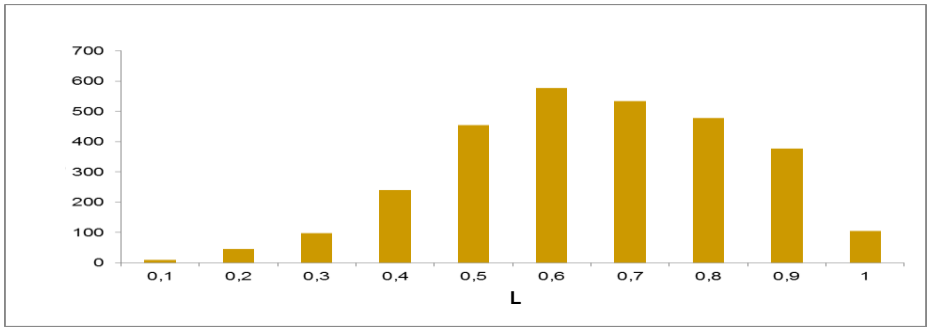

Fig. 8. Histogram of the factor L of automatic base before borders correction.

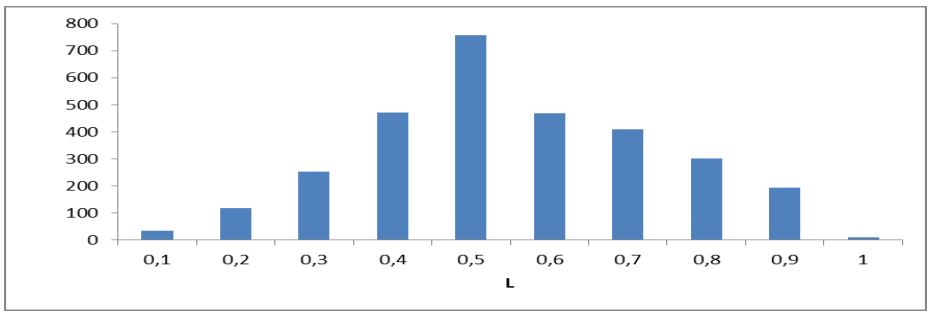

Fig. 9. Histogram of the factor L of automatic base after borders correction.

6 Summary

Research shows, that presented correction algorithms of borders of allophones appointed in automatic way improve the quality obtained base. Feedback from subjective researches also confirms it. It turns out that analytic base correctness ratio is a useful tool which, in numerical way, defines the quality of allophonic bases. Drawback of this approach and simultaneously drawback of this ratio, is the fact that it is possible to apply it only to voiced units of speech signal. However theoretically it is possible to use it in order to describe the quality of bases composed of other acoustic units than allophones.

7 Bibliography

1. Almpanidis G., Kotropoulos C., "*Automatic Phonemic Segmentation Using The Bayesian Information Criterion With Generalised Gamma Priors"*, Proceedings of EUSIPCO 2007
2. Szklanny K., Oliver D., *"Creation and analysis of a Polish speech database for use in unit selection speech synthesis*", Genova, LREC Conference, 2006
3. Dutoit T., "An Introduction to text-to-speech synthesis", Kluwer Academic Publishers 1997
4. Taylor P., *"Text-to-Speech Synthesis"*, Cambridge University Press 2009
5. Van Santen, Sproat R., Olive J., Hirshberg J., *"Progress in speech synthesis"*, Springer Verlag, New York 1997
6. Szpilewski E., Piórkowska B., Rafałko J., Lobanov B., Kiselov V., Tsirulnik L., *"Polish TTS in Multi-Voice Slavonic Languages Speech Synthesis System"*, SPECOM'2004 Proceedings, 9th International Conference Speech and Computer, Saint-Petersburg, Russia 2004, pp. 565 – 570
7. Jassem W., *„Podstawy fonetyki akustycznej"*, wyd. PWN, Warszawa 1973
8. Lobanov B., Piórkowska B., Rafałko J., Cyrulnik L.,*"Реализация межъязыковых различий интонации завиершённости и незавиершённости в синтезаторе русской и полской речи по тексту"*, Computational Linguistics and Intellectual Technologies, International Conference Dialogue'2005 Proceedings, Zvenigorod, Russia 2005, pp. 356–362
9. Matoušek J., *"Building a New Czech Text-to-Speech System Using Triphonebased Speech Units"*, Text, Speech and Dialog, Proceedings of the 3-rd international workshop TSD'2000, Brno, Czech Republic 2000, pp. 223–228
10. Rafalko J. *"The algorithms of automation of the process of creating acoustic units databases in the Polish speech synthesis"*, in Novel Developments in Uncertainty Representation and Processing, Springer 2015, pp. 373 – 383
11. Skrelin P., *"Allophone-based concatenative speech synthesis system for Russian"*, Text, Speech and Dialog, Proceedings of the 2-nd international workshop TSD'99, Pilsen, Czech Republic 1999, pp. 156–159