Algorithms for Automatic Selection of Allophones to the Acoustic Units Database

Janusz Rafalko

Algorithms for automatic selection of allophones to the acoustic units database

Janusz Rafałko

Faculty of Mathematics and Information Science, Warsaw University of Technology, Poland

j.rafalko[at]mini.pw.edu.pl

Keywords: TTS, speech synthesis, phoneme, allophone, selection of acoustic units

Abstract. The paper presents algorithms and coefficients developed in order to select specific acoustic units to the base used in concatenative speech synthesis. The approach is based on the assumption that the database is created automatically. In the natural speech signal, which is a sample of the voice of a particular person, the acoustic unit must be marked and then cut out. This generates often very large, redundant collection of units from which the best units should be selected to the final base. Described coefficients refer to allophones databases in TTS synthesis.

1 Introduction

In concatenative speech synthesis one of the very important elements is a database of speech units containing natural acoustic unit of speech. The units may be allophones, diphones, syllables or others. Nowadays in current synthesis systems, such bases are created manually or by using algorithms that automate part of the process. When using automated algorithms, the first step is to obtain a set of units, which is a redundant set. It contains many of the same elements, but from different words. The boundaries of such allophones can be marked in various ways and could be marked incorrectly. In the concatenative approach only one copy of each unit is needed in the final database. Therefore, an appropriate selection should be done.

The article presents developed coefficients which allow us to make a units selection and select the appropriate allophones into the database. The units, which are evaluated, are the acoustic allophones. This database is intended

adfa, p. 1, 2011.

to be used directly in the TTS synthesizer based precisely on these units. It is possible, however, to extend this to databases created with other units.

2 The set of acoustic units

Different approaches to speech synthesis on the basis of the text are described quite detailed in [1, 2]. In the synthesizer, signal compiled from natural speech segments is subjected to a modification in which the prosodic signal parameters are changed. In [3] there are basic assumptions of concatenated TTS system for Polish language, based on allophones in the context of the synthesis of Slavic languages. This paper refers to the acoustic units, which include several context groups of a tested phoneme, which can be identified with acoustic allophone presented in [4] by W. Jassem.

The advantages of the choice of allophones as basic units [5, 6] are based on the fact that speech units retain the synergistic effects between sounds. The difficulty with this approach is the need for precise marking of allophones' borders in the segmentation of a natural speech signal.

The set of acoustic units may be created manually, or by using suitable algorithms. In this case, it is used modified DTW (Dynamic Time Wrapping) algorithm to create this set , which automatically marked the boundaries of all allophones in natural speech signal [7]. As the result, created collection contains many of the same units. It happens so, because the corpus of the speech on the basis of which the set is created is redundant, it contains the whole words, sentences and texts in which the unit occurs multiple times.

3 The acoustic units selection

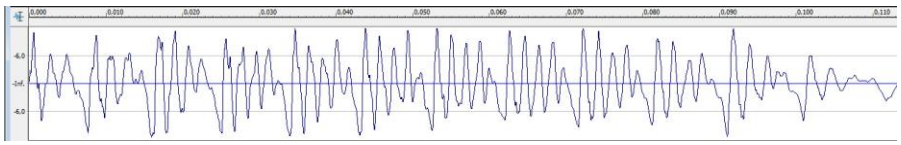
In order to create the base of acoustic units it is necessary to analyse in detail all the received units. First it is necessary to remove the units which are not suitable to the final allophone base. After rejecting the worst, however, we are still left with redundant units from which we should select the best ones. Selection is performed using a reference base realized manually. This is the allophone base units of professional voice actor working as a radio presenter.

3.1 Rejection of the worst

The main reasons for the use of this operation is the phenomena of reduction and simplification of phonemes in natural speech leading to almost complete disappearance of phonemes, with the result that the phonetic content of the synthesized speech does not coincide with natural speech, e.g. wiśniewski → viɕɲɛsci (alphabet IPA – International Phonetic Alphabet). Another reason is the inaccurate markings of the natural signal in the process of boundaries setting, with the result that a segment whose acoustic content does not correspond to the phonetic content is cut out from the natural speech.

Fig. 1 shows an allophone o0022 cut out from natural signal in the automatic segmentation. This presents the situation of inadequate marking of the borders in the unit cut out automatically (b), in comparison with that of the reference (a). As it can be seen, the phonetic contents of unit is different from the pattern. Similarly, the duration differs significantly from the duration of reference one. The time scale on both figures is the same. The duration of the reference unit is about 0.11 s, while the tested unit is 0.08 s.

a)



b)

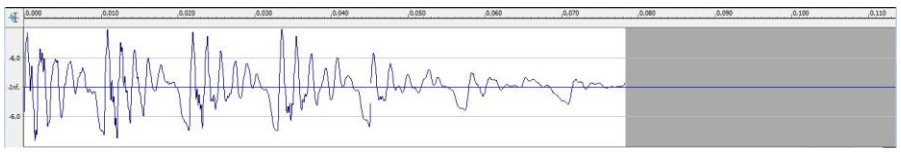


Fig. 1. Inaccuracy in the marking of the allophone border.

Rejection operation is achieved by testing the time parameters, i.e. the duration of the allophone, and parameters of acoustically - phonetical, i.e. the cost of matching in segmentation algorithm with units of the reference base. The duration of the test units T_T obtained in the segmentation process is compared with the duration of reference units T_R used in the synthesis mod-

ule. Duration of units from reference base of professional voice actor is in the range from 35 ms to 0.52 s. Length of automatically cut units is in the range from about 0.05 ms to 1.6 seconds.

Figures 2 and 3 show histograms of units duration, the first of reference base and the second of units cut automatically. The reference base contains about 2,000 units, while a collection of units cut automatically contains more than 11,000 items.

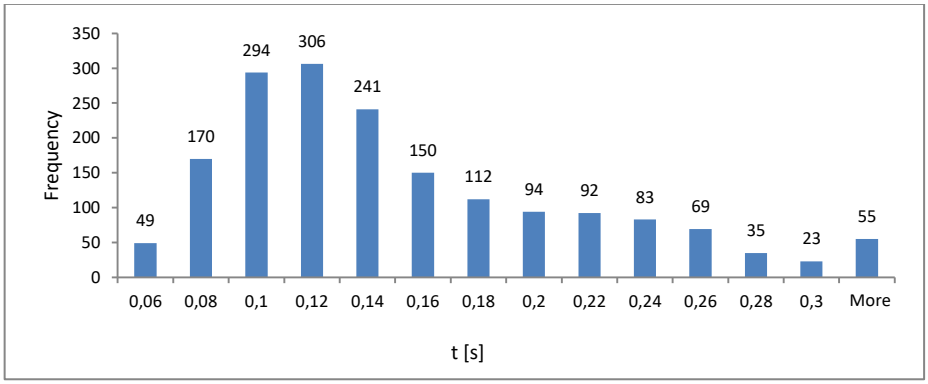


Fig. 2. Histogram of duration of allophones in the reference base.

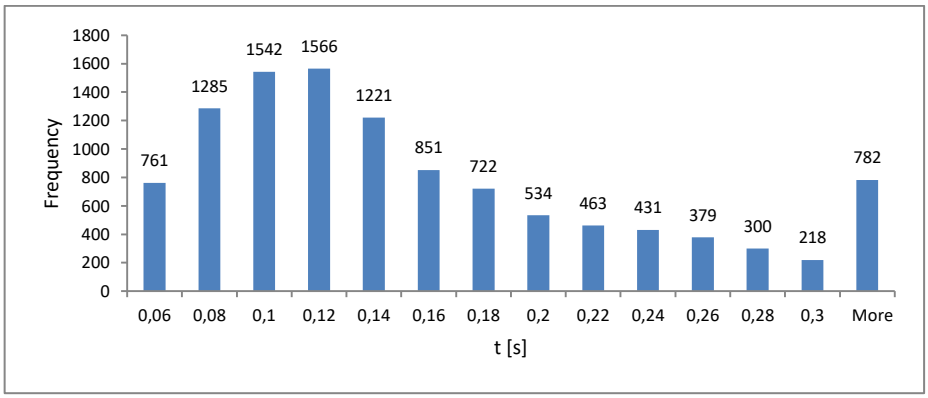


Fig. 3. Histogram of duration of allophones cut automatically.

Such a large number of received units is related precisely to the redundancy of the acoustic corpus, where one element is presented in more than a hundred copies. As it can be seen on the graphs, units durations of the reference base and the set cut automatically have similar distributions. There are about

40 units which are too short, of duration less than 20 ms in this collection and about 70 units which are too long with the duration of more than 0.7 second. This is less than 1% of all units cut automatically. Of course, this does not mean that only those units should be rejected. There may be more units cut automatically which are too short or too long in relation to its pattern. In this criterion, the difference in durations of the reference unit and units obtained automatically is important. Because the allophones units have different lengths, they cannot be taken as a measure of absolute error but a relative error, according to the formula (1).

$$\delta_t = \frac{|T_T - T_R|}{T_R} > \alpha \tag{1}$$

If the relative error of the units duration is greater than the threshold, the element is rejected. Figure 4 shows the relative error of the units duration of exemplary set. About 80% of cut allophones have a duration error less than 50%. Experimentally, for different collections it has been determined that if the error exceeds 80%, then such allophone is not suitable for the final base. There are about 1,000 of such allophones in the sample set, which is less than 9% of their total number. This means that the parameter α should be 0.8.

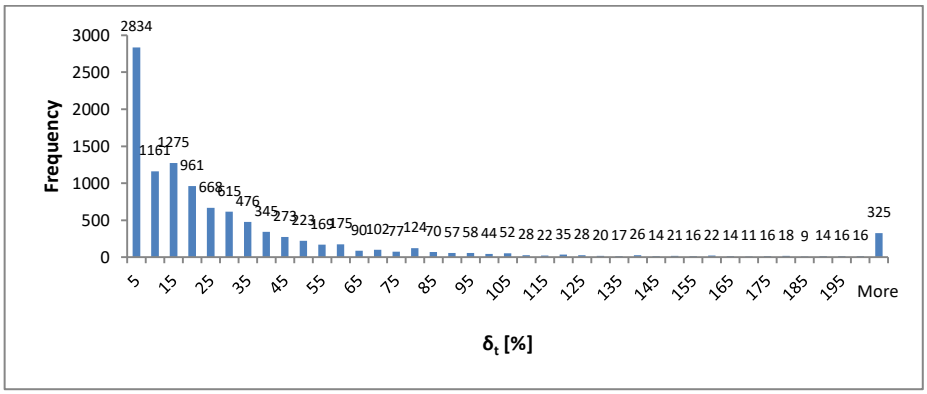


Fig. 4. Histogram of duration relative error of allophones.

The second criterion in this operation is the cost of matching both units, tested and referenced - in this case, the cost of matching in DTW algorithm. This cost is the sum of local distances within the alignment path determined

for those units. Because the local distance determines the degree of similarity of these units in the frequency domain, this cost can be a measure of the phonetic accuracy of tested allophone unit. This cost, similarly as a unit duration, may differ for particular units. If we take the cost of matching two identical units in the DTW algorithm, it would be zero. In contrast, if there is any difference it will be greater than zero. In tested collections it was in the range from 0.05 to about 50. Figure 5 shows a histogram of the matching cost of units in the tested set. 90% of units has a matching cost less than 10.

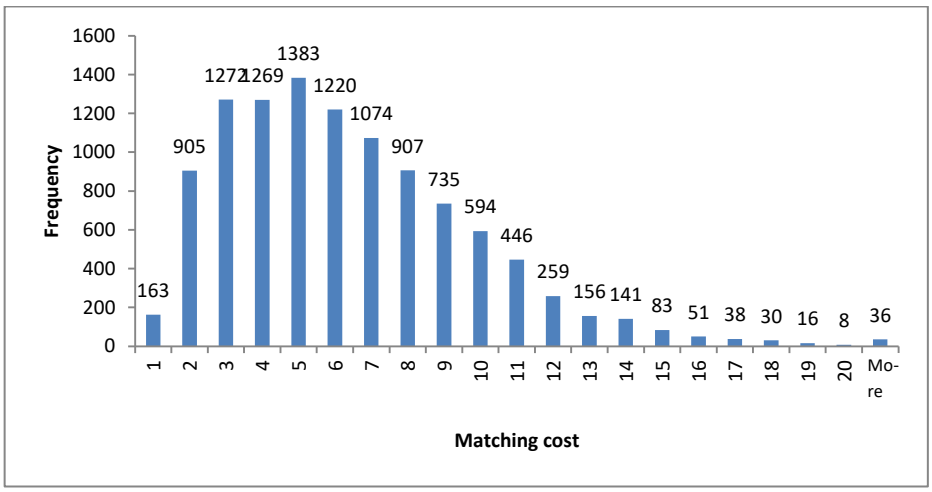


Fig. 5. Histogram of allophones matching cost.

In case of this parameter, similarly to the previous case, it largely depends on the specific unit. That is why we had to develop a relative factor like in previous case. However, the cost of matching the reference unit cannot be taken into account, because it is equal to zero. That is why, the average value of matching cost for all instances of a given entity was taken into account. Formula (2) shows the developed ratio:

$$\delta_C = \frac{|C_P - C_{Pave}|}{C_{Pave}} > \beta \tag{2}$$

where:

C_P – matching cost of allophone unit

C_{Pave} – the average matching cost of all instances of the unit

We reject units where matching cost error exceeds the threshold. On the following histogram in figure 6, we can conclude that the 90% of cut out allophones have a matching cost error less than 35%. In this case, it was experimentally determined for varied sets that an error greater than 50% disqualifies such cut unit from the use in output base. This is about 3.5% of the total number of units.

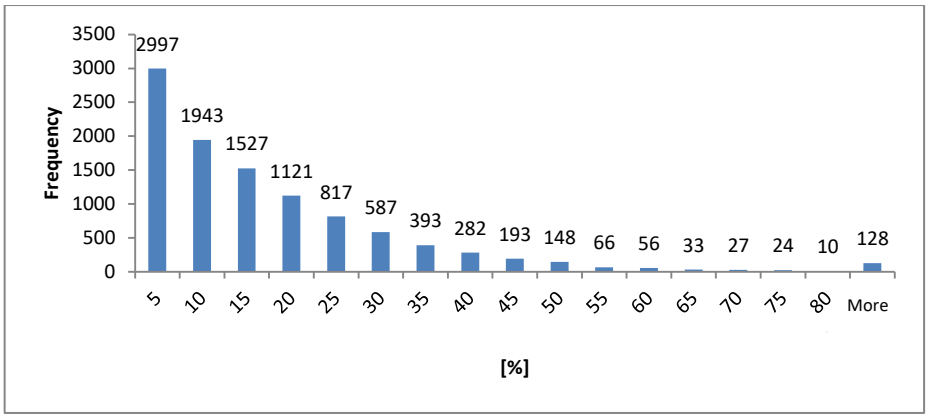


Fig. 6. Histogram of relative matching cost error of allophones.

As a result of the rejection process all the copies for which $\delta_t > \alpha$ or $\delta_C > \beta$ are excluded from further processing. As the experiments show, the best results of this operations are obtained at $\alpha = 0.8$ and $\beta = 0.5$.

3.2 The selection of acoustic units

Number of copies of the allophone units which have undergone rejection operation, depends on the number of such units in the acoustic corpus, the quality of the voice actor's speech and the accuracy of the marking of unit boundaries. For the pieces that remain in the set after rejection the selection operation is applied, which produces one and the best representative of any unit.

The acoustically - phonetical characteristics of each copy of allophone which past rejection should already be good enough to be put it in the created base. Taking into account that the resulting elements will be modified in the process of prosodic modification in speech synthesis, the most typical item

by the value of prosodic characteristics should be selected: the basic tone frequency F_0, amplitude A and duration T. As such an item in the operation selection is selected, it has been chosen the entity which has the characteristics closest to the average values of these parameters. If after the rejection operation number of copies of the unit is n, then for each such unit the duration T_i, the average amplitude A_i^{ave}, and the average value of frequency of the basic tone F_{0i}^{ave} is evaluated. Then we calculate the average values for the entire set of units of one type: T^{AVE}, A^{AVE} and F_0^{AVE}. Normalized in scale [0...1] similarity coefficient of unit prosodic characteristics, which could be called the selection coefficient, is calculated as:

$$D_i = \frac{1}{3}\left(\frac{\left|T_i - T^{AVE}\right|}{\max\limits_{j=1}^{n}\left|T_j - T^{AVE}\right|} + \frac{\left|A_i^{ave} - A^{AVE}\right|}{\max\limits_{j=1}^{n}\left|A_j^{ave} - A^{AVE}\right|} + \frac{\left|F_{0_i}^{ave} - F_0^{AVE}\right|}{\max\limits_{j=1}^{n}\left|F_{0_j}^{ave} - F_0^{AVE}\right|} \right) \quad (3)$$

As a result of the selection operation we select the copy of allophone which performs the condition:

$$k = \arg \min_{i=1}^{n}(D_i) \quad (4)$$

The first part of the formula (3) refers to the duration of the unit. E.g. allophone A1001 in a set received automatically occurs in 117 copies. After the rejection operation, there are 96 copies. The duration of the reference unit in this case is 0.1359 s, and the average duration of the set of units is 0.1534 s. The selected unit is the one with a duration of 0.1592 s. This is shown in the figure 7.

A second part of selection coefficient refers to the amplitude of the allophones units therefore it is associated with the volume of the signal. Before the segmentation, the speech signal has been normalized, so it is possible to compare amplitude of the same units cut out from different words.

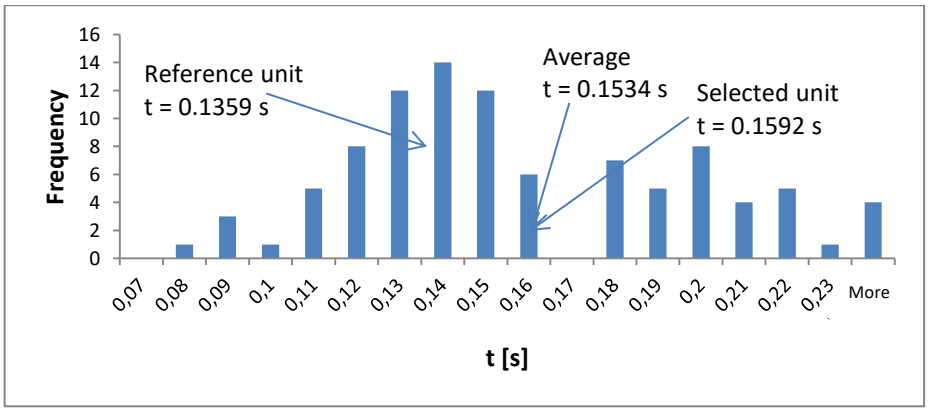


Fig. 7. Histogram of duration of a set A1001 units in selection operation.

For an exemplary set of A1001 units in the figure 8 we showed the amplitude histogram and selected unit which goes to the final base.

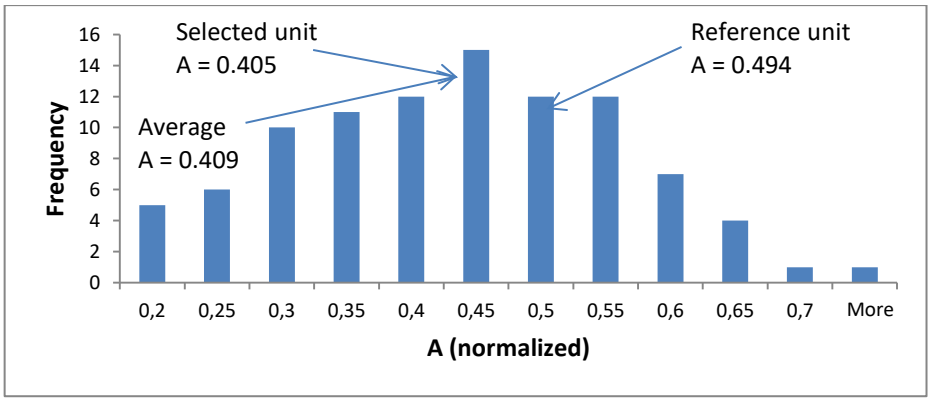


Fig. 8. Histogram of amplitude of a set A1001 units in selection operation.

The last part of the selection coefficient refers to the frequency of the basic tone. This value must also be selected as close to the average as possible. As the figure 9 shows, the average F_0 of this set differs from F_0 of reference unit. This is due to the fact that the reference base it is the voice of a different person than the voice of person who recorded the acoustic corpus. The unit chosen into the base has a higher frequency of the basic tone than average. This is due to the fact that, according to the formula (3), simultaneously the three parameters must be selected to be as close as possible to the average.

This is provided by the selection coefficient developed by the author and described above.

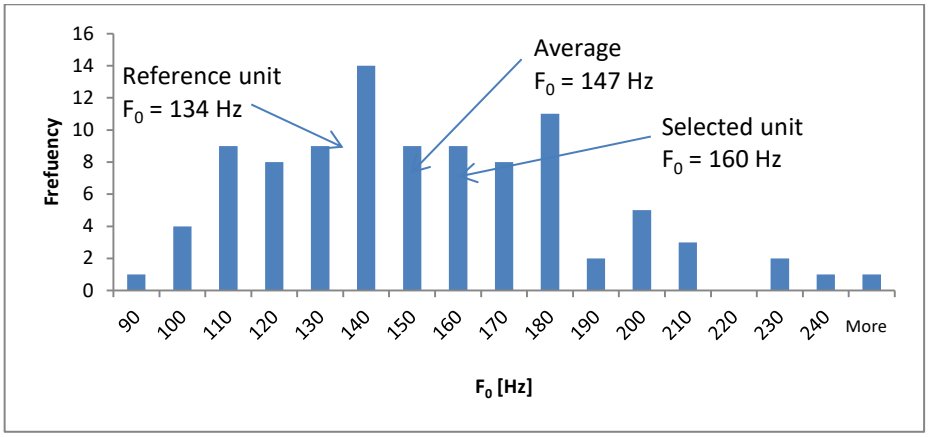


Fig. 9. Histogram of basic tone frequency of a set A1001 units in selection operation.

As a result of these algorithms we create a set of units which will be placed in an allophone base.

4 Conclusion

Creating the bases of acoustic units of different voices designed for speech synthesis is necessary to automate the process because of the time-consuming nature of the manual approach. Manual approach also requires extensive knowledge and experience. Automation saves time and allows to create synthesizers speaking practically in any person's voice providing, the appropriate voice sample was taken. One stage in this approach, is the choice of a particular natural sound units to a final base. The factors and algorithms that allow this choice, are developed for allophones, but can easily be generalized to other units. The selected units can, in subsequent stages, undergo further modifications in order to obtain a base of high quality and naturalness of synthesized speech.

5 Bibliography

1. Taylor P., *"Text-to-Speech Synthesis"*, Cambridge University Press 2009

2. Van Santen J, Sproat R., Olive J., Hirshberg J., *"Progress in speech synthesis"*, Springer Verlag, New York 1997

3. 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

4. Jassem W., *„Podstawy fonetyki akustycznej"*, wyd. PWN, Warszawa 1973

5. 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

6. 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

7. 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, Advances in Intelligent Systems and Computing, 26-28 October 2015, Cracow, Poland, Springer 2015, pp. 373 – 383