Post-print of: Rafałko J., Czyżewski A. (2020) Automatic Marking of Allophone Boundaries in Isolated English Spoken Words. In: Saeed K., Dvorský J. (eds) Computer Information Systems and Industrial Management. CISIM 2020. Lecture Notes in Computer Science, vol 12133. Springer, Cham. https://doi.org/10.1007/978-3-030-47679-3_5

# Automatic marking of allophone boundaries in isolated English spoken words

Janusz Rafałko[1][0000-0002-5369-5868] and Andrzej Czyżewski[2][0000-0001-9159-8658]

[1] Warsaw University of Technology,
Faculty of Mathematics and Information Science,
Warsaw, Poland
j.rafalko@mini.pw.edu.pl
[2] Gdańsk University of Technology,
Faculty of Electronics, Telecommunications, and Informatics,
Gdańsk, Poland

**Abstract.** The work presents a method that allows delimiting the borders of allophones in isolated English words. The described method is based on the DTW algorithm combining two signals, a reference signal and an analyzed one. As the reference signal, recordings from the MODALITY database were used, from which the words were extracted. This database was also used for tests, which were described. Test results show that the automatic determination of the allophone limits in English words is possible with good accuracy. Tests have been carried out to determine the error of particular allophones borders marking and to find out the cost of matching the given allophone to the reference one. Based on this cost, a coefficient has been introduced that allows for determining in percentage how much the automatically marked allophone is similar to the reference one. This coefficient can be used for an assessment of the correctness of the pronunciation of the allophone. The possibilities of further research and development of this method were also analyzed.

**Keywords:** Speech recognition, speech analysis, phoneme, allophone.

## 1 Introduction

This work is a part of a project devoted to the multimodal signal analysis of allophones, where an allophone can be defined as a variant of a particular phoneme [5], [8]. The project is based on combining audio and visual modalities [3]. The combination of these two modalities leads to improved accuracy of allophone transcription and recognition.

The applied algorithm is intended for the automatic marking of allophone boundaries in isolated English words based on the reference speech base obtained in the result of the preparation of the audio-video corpus [4], [17]. Although the corpus contains multimodal material, this work employs a single modality only, namely sound. This is the first part of the work that will answer the question if we can mark the limits of allophones in a continuous speech based on the sound signal only, with good

accuracy. Further research, on the other hand, will allow answering the question of whether adding a second modality will improve this accuracy.

The problem of marking the borders of allophones in continuous speech is a very important issue in speech technology. It is related to such subjects as, for example, speech recognition and transcription, speech synthesis, or learning a foreign language. Speech transcription and recognition are described extensively in the literature, as well as speech synthesis, which is also covered broadly. For example, the paper of Szpilewski et. al. [16] shows the approach to the concatenative TTS (Text-to-Speech) system based on allophones in the context of multilingual synthesis.

Another area where the appropriate marking of the allophone boundaries is an important task is the field of foreign language learning. Appropriate marking allophones in English has high development potential, because nowadays according to David Cristal, author of the "English as a Global Language" [2], non-native English speakers are three times as many as native speakers. That is why algorithms related to the correct delimitation of allophones can be very helpful in learning the correct pronunciation.

In Chapter 2, the speech recording process done in the project will be described, as well as the reference speech database used for this work to mark the allophone boundaries is described.

Chapter 3 is devoted to the description of the determination of the allophone boundaries algorithm in continuous speech. The algorithm is based on the DTW (Dynamic Time Warping) method. The modification of this method allows to marking of the boundaries of allophones. The parameters of the method are also presented.

In Chapter 4, a coefficient related to the correctness of the allophone boundaries and the correctness of the pronunciation of the allophone will be defined.

Chapter 5 contains an evaluation of the results. The test will be described that were carried out for various recordings, i.e., speech of different people.

Chapter 6 contains a summary and conclusions, as well as a discussion on planned further research.

## 2 Database

### 2.1 Phonetic material

Audio recordings from the Modality [4] corpus were used as a referenced corpus in this study as well as a data source to carry out the tests (www.modality-corpus.org). The Modality corpus material consists of spoken numbers, names of months and days, and a set of verbs and nouns mostly related to controlling computer devices. It was presented to speakers as a list containing a series of consecutive, isolated words, and sequences of continuous speech. The corpus includes recordings of 35 speakers. The gender composition is 26 male and 9 female speakers. The corpus is split between native and non-native English speakers. Approximately half of the isolated words represented some typical command-like sentences, while the rest was formed into isolated word sequences. Every speaker participated in 12 recording sessions. Half of the sessions were recorded in quiet conditions, and the second half includes noise.

Reference noise-only recording sessions were performed in order to enable a precise calculation of SNR (signal-noise ratio) for every sentence spoken by the speaker. The audio-visual material was collected in an acoustically adapted room. The video material was recorded using two Basler ace 2000-340kc cameras. The cameras were set up to capture video streams at 100 frames per second, in 1080x1920 resolution. The audio material was collected from an array of 8 B&K measurement microphones placed in different distances from the speaker. The audio data were recorded using 16-bit samples at 44.1 kSa/s sampling rate with PCM encoding. The setup was completed by loudspeakers placed in the corners of the room, serving as noise sources. The average SNR calculated in the 300 – 3500 Hz of frequency range was 36 dB for the quiet condition, and 17.2 dB for noisy conditions.

### 2.2 Reference database description

The reference base used for the presented system contains words from the Modality database stored in separate files. In addition, each word contains the boundaries of the allophones that it consists of. The marking of these boundaries was performed manually. Fig. 1 shows an example word “clever” of the male native speaker, IPA notation: ˈklevəʳ, from the reference database with the allophone boundaries marked manually by the author of this work. Recordings of twenty people, eight women, and eleven men were selected for the reference database, of which two female and six male voices belonged to native speakers. The speakers were between the ages of 22 and 58. One of these voices was always used as a reference voice in the system, and the others were used to mark allophones in them. Only recordings without additional noise have been selected from the Modality database, which is only with ambient noise.

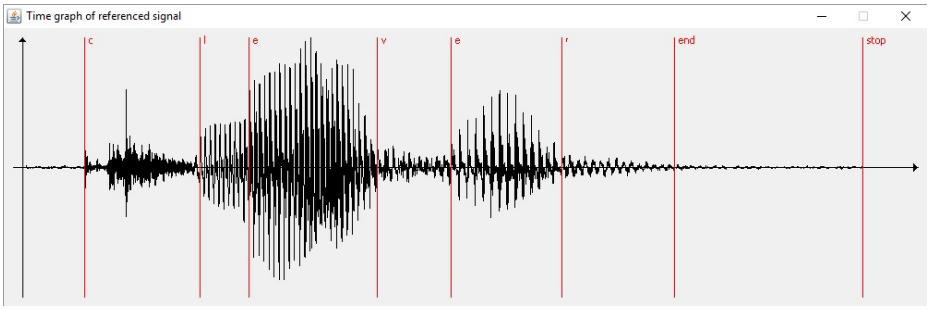


**Fig. 1.** Reference word “clever” with manually marked allophone borders

Because of the need to compare a homogenous material, there was not possible to mark the allophone boundaries automatically in any speech signal, but only in such recordings for which there is an equivalent in the reference database. Based on the marked allophones in the word from the reference database, it is possible to mark the boundaries of the allophones in the same word also from outside the base.

For the automatic system building, these boundaries are stored in a separate text file in the format: “allophone, sample number.”

# 3 Algorithm for automatic marking of allophones

The algorithm for automatic marking of allophone boundaries is based on the DTW method [7], [9], [15]. A system using a similar approach, but designed to create bases of acoustic units in speech synthesis, was presented in the author's earlier paper [12]. The approach presented in this paper differs from the previous one in several fundamental issues. First of all, it concerns the English language, instead of Polish. Secondly, in the previous system, the bases of acoustic units were created employing a much larger, redundant recording corpus, which allowed for averaging the results and selecting the best one unit. In the previous work, the databases were created for the concatenation speech synthesis, in which only one realization of a given allophone was needed, hence it was necessary to find the best one. Here we should approach each allophone individually because e.g. we want to determine whether the given allophone was pronounced correctly or not. In this case, the point is not to find one allophone in the whole corpus. In this case, we want to define limits for each allophone in the spoken word, to determine how correctly the given allophone has been pronounced. The consecutive steps leading to this aim are described in subsequent subchapters.

## 3.1 Combining reference and analyzed signals

The algorithm of automatic segmentation of the speech signal is based on the DTW method [1], [9], [10], [11] however, unlike the classic DTW, it does not rely on the signal in the time domain, but on the frequency domain representation. The referenced speech signal (spoken word) and the analyzed signal (the same word pronounced by another person) in the time domain are divided into frames that can overlap each other. In each frame, the Fast Fourier transform (FFT) is computed. Before calculating the FFT, the Hamming windows [6] are used for avoiding spectral blur.

Each transformed frame represents a vector of spectral features. Elements of the local distance matrix are counted using these vectors as:

$$c(n,m) = \|S(n), E(m)\| = \sum_{k=1}^{K} |S(n,k) - E(m,k)| \quad (1)$$

where:

*S (n)* - vector of spectral features of the referenced signal in the *n-th* frame
*E (m)* - vector of spectral features of the natural signal in the *m-th* frame
*K* - length of the spectral features vector

The reference signal frame is combined with the analyzed signal frame, and then the distance between these vectors is calculated using the signal spectrum in the frame. In formula 1, the distance is calculated according to the Manhattan metric, being used for the algorithm.

The global distance matrix calculated is shown in Fig. 2. It presents speech signals to which 256-sample frames were applied, employing the Hamming window and no overlapping. The reference signal is shown in the form of a spectrogram drawn vertically on the left side of the drawing. The analyzed signal is also presented in the form

of a spectrogram but at the bottom of the drawing. Both signals combined are identical, and in the considered case, there are of the voice of a male native speaker. The square area in the center of Fig. 2 shows the global distance matrix. The bright areas indicate small values of the distance between the signal frames that are the signals spectrally similar, while the dark areas indicate a large distance between frames, which are signals that differ in spectral features.

The warping path is also shown there, which is going through the areas of the lowest cost. Because in this case, both signals are identical, the warping path is going alongside the anti-diagonal of the global distance matrix.

Having the optimal warping path and the boundaries of the allophones in the reference signal, it is possible to determine the allophone boundaries in the analyzed signal, as is shown in Fig. 2. A vertical spectrogram on the left-hand side shows a reference signal in which allophone boundaries are known because they have been manually marked by a phonetician expert. In the figure, they are marked with horizontal lines, which additionally go into the area of the global distance matrix and end in the optimal warping path. The points of intersection of these lines and the warping path determine the boundaries of the allophones in the analyzed signal, as it is represented in the figure by vertical lines going down from the warping path. These lines pass to the spectrogram of the analyzed signal, as it is shown in Fig. 2, and they determine the limits of the allophones in this signal. In this example, both phrases are identical, therefore the warping path is a straight line.

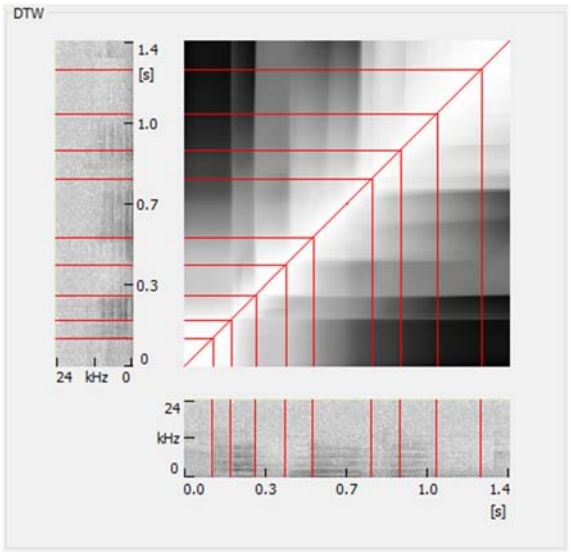


**Fig. 2.** Global distance matrix of "keep moving" phrase and determination of allophone boundaries in the analyzed signal

In case when we set boundaries in a word spoken by a different speaker than the one who produced the reference one, the matching path will no longer be a straight line, as is shown in Fig. 3. The reference signal belongs to a native speaker, whereas the ana-

lyzed signal to a non-native speaker. The parameters for determining the path are the same as before.

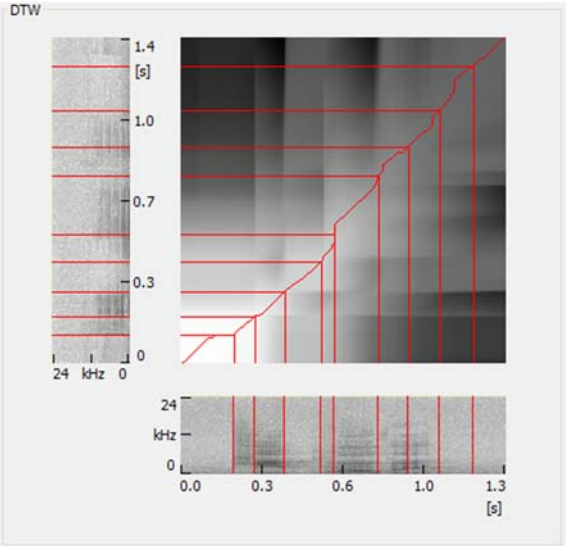


**Fig. 3.** The warping path and the allophone boundaries in the “keep moving” phrase of speaker different to the reference speaker

The allophone boundaries set in this way can be compared with manually marked boundaries, as is seen in Fig. 4. The same "keep moving" phrase is shown on both graphs. The upper graph shows the boundaries of allophones determined manually, while the lower graph - boundaries determined automatically. In Fig. 4 we can visually evaluate the location, and thus roughly the quality of automatically defined boundaries. As is seen, the differences, in this case, are minimal. In Section 5 the statistical assessment will be presented computed on the reference set discussed earlier.

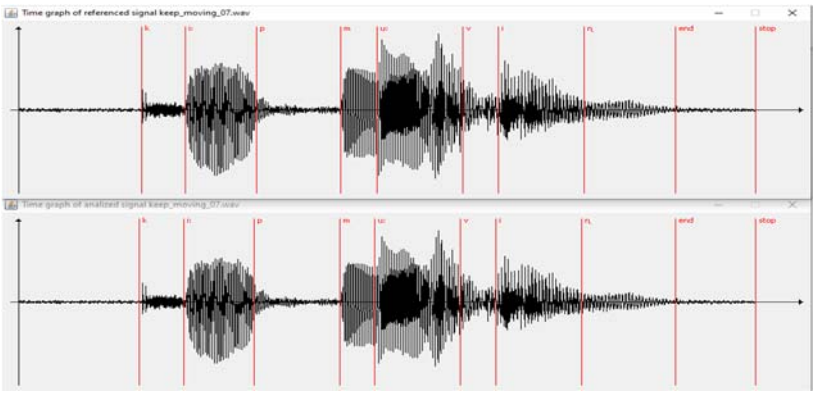

**Fig. 4.** Comparison of boundaries marked manually and automatically

The boundaries of allophones obtained in this way are not always correct. The reason may be incorrect pronunciation or errors in the automatic determination of the border. Errors resulting from algorithm inaccuracies can be corrected. A discussion on this topic is presented in previous papers [13], [14]. The work [13] refers to the correction of allophone boundaries in acoustic databases used in Polish speech synthesis, where only one allophone cut out from a set of many redundant allophones of the same type is subject to correction. In [14], an approach to the correction of all allophones in a word is presented.

## 4 The cost of allophone matching

While determining the boundaries of allophones in the analyzed word, we use a reference word spoken by another speaker. By juxtaposing both words, the warping path is determined. The values on the path are the cost of matching, and they are derived from the global distance matrix. These values determine the similarity of both signals. Having determined allophone limits manually in the reference signal, besides determining the allophone limits in the analyzed signal, we can also calculate the cost of matching of individual allophones. The result represents the cost of matching the allophone designated on the warping path in terms of a difference in the value of points on the path (formula 2):

$$C_M(allophone) = C_{end}(allophone) - C_{start}(allophone) \quad (2)$$

where:

$C_M(allophone)$ – the matching cost of allophone

$C_{end}(allophone)$ – the cost of matching of the end of the allophone in the point of the intersection of its border with the warping path

$C_{start}(allophone)$ – the cost of matching of the beginning of the allophone in the point of the intersection of its border with the warping path

The above cost (formula 2), determined when two identical signals are juxtaposed, will have a value of 0. Such a situation will take place, as is shown in the example presented in Fig. 2. In case of determining the allophone limits in a different signal than the reference one, the cost will be greater than zero. Table 1 shows the cost of matching of allophones in the word “clever” determined using the previously described method. The table also includes the limits of the beginning of allophones marked manually and automatically. The fourth column presents the absolute boundary determination error calculated in relation to the length of allophones marked manually. To obtain this error, allophones should be manually marked in the analyzed word, too.

The absolute error applies only to the location of the boundary, while the cost of matching determines the similarity of the allophones. For example, allophones 4 (v) and 5 (e): the error of defining the border for the allophone "v" is larger than that of the allophone "e." However the cost of matching is presented in the opposite way: for

the allophone "v" the cost is less than for "e," which means that "v" is more frequency-related to the reference allophone than "e" to the reference "e."

**Table 1.** Error and cost of allophone matching

| Allophone | Manual borders | Automatic borders | Error | Cost of matching | Correctness |
|---|---|---|---|---|---|
| c | 22960 | 21999 | 7,10% | 508.0 | 0,01 |
| l | 27568 | 27633 | 7,91% | 406.1 | 0,13 |
| e | 29611 | 29779 | 5,53% | 1440.8 | 0,04 |
| v | 33812 | 33536 | 7,13% | 222.8 | 0,16 |
| e | 36831 | 36755 | 1,33% | 549.2 | 0,17 |
| r | 42753 | 42389 | 6,22% | 131.7 | 0,10 |

The specified match cost can be used for calculating the coefficient determining the correctness of the automatically marked allophone. An absolute error cannot play the role of such a factor, because it requires correctly, manually set the limits of the allophones, thus it can be used only for testing purposes. In order to obtain such a factor, the average cost of determining the allophone at a given reference signal should be calculated. Furthermore, this average cost should be set for correctly pronounced allophones, which is for the voices of native speakers. With this average cost, we can associate the coefficient of similarity of the given allophone to the reference one as shown in formula 3. This coefficient can be used to determine the correctness of the pronunciation of the allophone.

$$S_{al} = \left|\frac{C_{al} - C_{avr}}{C_{avr}}\right| \tag{3}$$

where:

$S_{al}$ – allophone similarity to the referenced one coefficient
$C_{al}$ – the matching cost of the allophone
$C_{avr}$ – the average matching cost of the allophone

The smaller this ratio is, the more allophone is similar to the reference one. An example illustrating this procedure for the word “clever” is presented in Table 1.

## 5 Tests and evaluation of results

The tests were carried out using the referenced database described in Section 2. Table 2 presents the parameters of the example allophones marked automatically. The average error is presented related to the marking of the allophone together with the standard deviation of this error for the case of determining this allophone in the words of native speakers and non-native speakers. Similarly, the average cost of matching the allophone on the warping path of the global distance matrix of the DTW algorithm is given. Similarly, as in the case of a marking error, the average cost and its standard

deviation are calculated separately for native and non-native speakers. In order to achieve these results, 20 different words were used uttered by 20 speakers described earlier. The upper part of the table shows male voices, the lower part - female voices. In both cases, the reference signal was a native speaker's voice. The signal processing parameters are 256-samples frame, with the Hamming window overlapped by 50 %.

**Table 2.** The average error of the allophone marking and the average cost of matching

| Male | Non-native | | Native | | Non-native | | Native | |
|---|---|---|---|---|---|---|---|---|
| | Av. Error % | Err. St. Dev. | Av. Error % | Err. St. Dev. | Av. Cost | Cost St. Dev. | Av. Cost | Cost St. Dev. |
| s | 7,90 | 5,03 | 7,57 | 7,64 | 2221 | 591 | 2065 | 102 |
| e | 5,21 | 3,16 | 6,15 | 6,11 | 1918 | 1122 | 1443 | 203 |
| t | 28,86 | 19,07 | 7,08 | 8,93 | 1057 | 211 | 908 | 44 |
| ə | 32,57 | 23,32 | 10,00 | 5,14 | 570 | 235 | 379 | 30 |
| n | 11,09 | 12,91 | 8,23 | 5,74 | 697 | 217 | 633 | 119 |
| | | | | | | | | |
| **Female** | **Non-native** | | **Native** | | **Non-native** | | **Native** | |
| | **Av. Error %** | **Err. St. Dev.** | **Av. Error %** | **Err. St. Dev.** | **Av. Cost** | **Cost St. Dev.** | **Av. Cost** | **Cost St. Dev.** |
| s | 9,32 | 7,19 | 7,69 | 5,16 | 4117 | 496 | 3583 | 237 |
| e | 6,61 | 2,73 | 3,62 | 2,15 | 2544 | 898 | 2052 | 240 |
| t | 15,49 | 12,50 | 9,70 | 7,84 | 1899 | 362 | 1042 | 178 |
| ə | 25,24 | 11,77 | 17,46 | 10,02 | 2765 | 725 | 1287 | 52 |
| n | 20,36 | 17,89 | 2,58 | 1,61 | 1129 | 513 | 882 | 218 |

Depending on the allophone, average errors in their marking can range from about 2% to about 20% for native speakers, and from about 5% to even several dozen for non-native-speakers. However, in cases where the pronunciation is correct, allophones marking errors of non-native speakers are similar to native ones. Similarly, the standard deviation for non-native speakers is larger but does not differ significantly from native speakers.

When the marking error exceeds the value of about 20 %, it means that the allophone limit has been determined incorrectly and that it was significantly shifted in relation to the manually determined one. It should be remembered that this error is determined in relation to the correctly, manually marked border. This case is illustrated in Fig. 5, where for the non-native speaker, the beginning of allophone "c" has been incorrectly designated. The upper part of the drawing shows the boundaries marked manually, the lower part shows borders marked automatically. Such cases where the error exceeds 50 % and the allophone is correctly pronounced have not been included in the calculation of the average errors in Table 2.

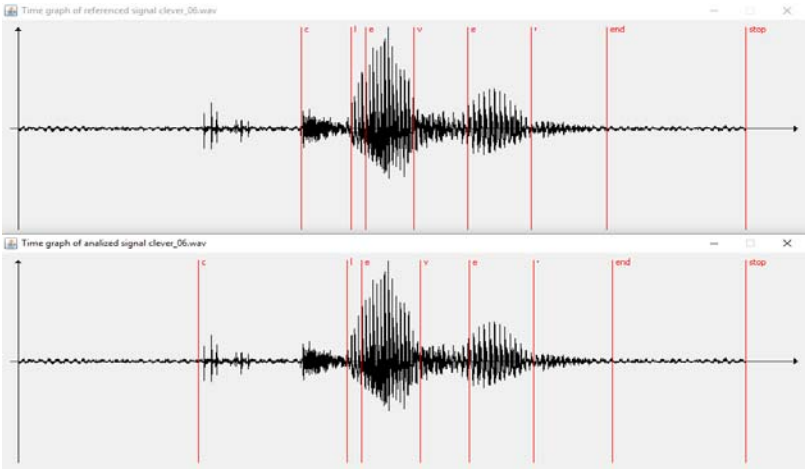


**Fig. 5.** Error in determining the border

The considerably high error in determining the border does not mean, however, that the allophone was pronounced wrongly. To determine the correctness of the spoken allophone, the cost of matching it to the reference one can be used. A slightly different situation is for this cost parameter, as Table 2 shows. The average cost for non-native speakers in the case of each allophone is higher, but also the standard deviation is noticeably higher. This average cost of matching can be used as a correctness coefficient, as is defined by the formula 3. Table 3 shows this coefficient with the error of the allophone limit marking for words belonging to different categories. The reference signal is a male native speaker. On its basis, the boundary for the voice of a male native speaker is marked, as well as for the male non- native speaker and for the female non-native voice.

**Table 3.** The correctness coefficient of the allophone

| Reference male | Native male | | Non-native male | | Non-native female | |
|---|---|---|---|---|---|---|
| | **Error** | **Correctness** | **Error** | **Correctness** | **Error** | **Correctness** |
| s | 2,23 | 0,22 | 7,30 | 0,22 | 23,75 | 0,04 |
| e | 4,16 | 0,17 | 4,20 | 0,03 | 6,55 | 0,55 |
| t | 0,15 | 0,03 | 36,23 | 0,54 | 16,51 | 0,98 |
| ə | 11,0 | 0,34 | 44,34 | 0,62 | 38,83 | 4,15 |
| n | 1,29 | 0,01 | 11,20 | 0,30 | 40,41 | 0,27 |

When the correctness factor has a value close to zero, it means that the given allophone is similar to the reference one, and that also means that it is correctly pronounced. For correctly pronounced words of voices of the same gender speakers, this coefficient is below 1. However, in the case of voices significantly differing in laryngeal frequency, very often, it can be as high as about 5, as is shown in Table 3, where on the basis of the male voice, the boundaries in the female voice are determined. For

the allophone “ə”, the correctness factor is 4.15, and also the error of determining the border is high, over 38%. The same situation occurs in the opposite case of determining the boundaries in the male voice on the basis of the female voice.

It can also be noticed another case here, that is a wrong set boundary does not mean that the allophone was pronounced wrongly. For the non-native male speaker, allophones “t” and “ə” are marked wrong in automatic mode, and errors exceed 30%, but similarity coefficients remain below 1, which means that these allophones were pronounced not quite wrongly.

## 6 Conclusions

The presented approach allows for determining the borders of allophones in isolated English words with satisfactory accuracy. Knowing determined average matching costs for every allophone, the correctness coefficient reflecting the similarity of the marked allophone to its reference counterpart is introduced. Using this factor, we can determine how much the spoken allophone is similar to the referenced allophone. It can be used e.g., for learning the correct pronunciation of a given allophone. In this way, it is also possible to determine compatibility with the correct pronunciation of words or phrases.

The results presented in this work allows for stating that allophone boundaries are determined correctly in the majority of cases. The correct determination of these boundaries in combination with the coefficient determining similarity to the referenced allophone can be used to improve the efficiency of recognizing specific allophones in speech, and thus to improve the quality of speech recognition.

The presented method can be further developed. In the presented work, all tests were carried out using the reference signal of one speaker, while the boundaries of allophones were determined for the words uttered by other speakers, with their voices produced employing different laryngeal tone. The presented method could be modified in such a way that the frequency of the laryngeal tone in the reference signal is changed so that for voiced allophones, it will be closer to the frequency of the analyzed signal. In addition, in the presented solution, the limits of allophones are determined only for words that have equivalents in the reference database. Meanwhile, the research and the solution can be extended towards determining boundaries in any words or phrases. It could be done, for example, by using a speech synthesizer, which would synthesize an appropriate signal from the reference database for which the boundaries in the analyzed voice will then be determined.

ACKNOWLEDGMENTS

Research sponsored by the Polish National Science Centre, Dec. No. 2015/17/B/ST6/01874.

## References


1. Bellman R., Kalaba R. (1959). On adaptive control processes, Automatic Control, *IRE Transactions*, vol. 4, no. 2, pp. 1– 9.
2. Crystal D. (2003). *English as a Global Language*, Cambridge University Press, 2 edition
3. Czyżewski A., Ciszewski T., Kostek B. (2017). Methodology and technology for the polymodal allophonic speech transcription, *The Journal of the Acoustical Society of America*, vvol. 139, issue 4, pp. 2017-2017
4. Czyżewski A., Kostek B., Bratoszewski P., Kotus J., Szykulski M. (2017). An audio-visual corpus for multimodal automatic speech recognition, *Journal of Intelligent Information Systems*, Volume 49, Issue 2, pp. 167–192.
5. Gafos A. (1999). *The articulatory basis of locality in phonology*, Routledge Taylor & Francis Group.
6. Harris F.J. (1978). On the Use of Windows for Harmonic Analysis with the Discrete Fourier Transform, *Proceedings of the IEEE*, vol. 66, no. 1, pp. 51 – 84.
7. Keogh E.J., Pazzani M.J. (2001). Derivative dynamic time warping, the 1st *SIAM International Conference on Data Mining*, Chicago, IL, USA.
8. Kiritani, S., Itoh, K., Hirose, H., Sawashima, M. (1977). Coordination of the consonant and vowel articulations — X-ray microbeam study on Japanese and English, *Annual Bulletin of the Research Institute of Logopedics and Phoniatry*.
9. Müller M. (2007). *Information Retrieval for Music and Motion*, Springer Berlin Heidelberg, part I, chapter 4, Dynamic Time Warping, pp. 69 – 74.
10. Myers C.S., Rabiner L.R. (1981). A comparative study of several dynamic time-warping algorithms for connected word recognition, *The Bell System Technical Journal*, Vol. 60, pp. 1389 – 1409.
11. Rabiner L.R., Rosenberg A., Levinson S. (1978). Considerations in dynamic time warping algorithms for discrete word recognition, *IEEE Transactions on Acoustics Speech Signal Process*, Vol. 26, pp. 575 – 582.
12. Rafałko J. (2015). The algorithms of automation of the process of creating acoustic units databases in the Polish speech synthesis, Novel Developments in Uncertainty Representation and Processing, series Advances in *Intelligent Systems and Computing*, vol. 401, Springer, pp. 373 – 383.
13. Rafalko J. Algorithm of Allophone Border Correction in Automatic Segmentation of Acoustic Units, in *Computer Information Systems and Industrial Management, series Lecture Notes in Computer Science* vol. 9842 (2016), Springer, 2016, pp. 462-469.
14. Rafalko J., Czyżewski A., Adjusting automatically marked voiced English allophone borders, Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA), Poznan, 18-20 Sept. 2019, doi: 10.23919/SPA.2019.8936805
15. Salvador S., Chan P. (2004). FastDTW: Toward Accurate Dynamic Time Warping in Linear Time and Space, *KDD Workshop on Mining Temporal and Sequential Data*, pp. 70 – 80.
16. Szpilewski E., Piórkowska B., Rafałko J., Lobanov B., Kiselov V., Tsirulnik L. (2004). Polish TTS in Multi-Voice Slavonic Languages Speech Synthesis System, *SPECOM'2004 Proceedings*, 9th International Conference Speech and Computer, Saint-Petersburg, Russia, pp. 565 – 570.
17. Modality Corpus. http://www.modality-corpus.org. [Accessed on 26.03.2019]