ALLOPHONE GROUP SELECTION FACTORS FOR POLISH SPEECH SYNTHESIS

Bożena Kozłowska, Janusz Rafałko and Mariusz Rybnik
Institute of Computer Sciences, University of Bialystok, Sosnowa str. 64, 15-887 Bialystok, Poland

Keywords: Speech synthesis, Allophone, Allophone group.

Abstract: The article concerns selection of allophone groups for Polish speech synthesis. It describes factors to be taken into consideration while dividing allophones into certain groups. Thus, the presentation includes classification suggested by the authors. Although the described factors regard Polish language, they may facilitate any study on similar division concerning any other language. Each language has determined specificity pronounces, therefore should choose suitable allophonic groups for the language. However precise description on what elements we should special attention give, where later problems can appear in pronunciation e.g. certainty will make easier work to persons making similar division in different languages.

1 INTRODUCTION

Issue of transforming text information into its voice equivalent constitutes a relevant part of the today's world. The need for speech synthesis is present in many fields of life. Therefore, it is important to develop this particular domain and create better and better speech synthesers that will allow us to gain as natural speech quality as possible.

At present, the most common method, and providing the best speech quality, is a method consisting in linking acoustic units that have been recorded and processed by an author beforehand. It is a so-called concatenation method which may link phones, diphones, triphones or syllables. A phoneme, from a phonological point of view, represents a speech sound i.e. the smallest distinguishable element of human speech. A phoneme is an abstract notion that defines a collection of articulatory features of a speech sound that allows one to differentiate it from other sounds. Thus a speech sound as a real creation is practical, audible realization of a phoneme. A phoneme may have several sound representations occurring in various contexts. They are so-called allophones. A diphone is transition between two phonemes. By a triphone one understands a sequence of three consecutive phones.

Choosing the kind of acoustic units and their number has a great impact upon the volume of acoustic database and the quality of speech synthesis. For instance, concatenation of phones requires the smallest number of acoustic units, but it produces the worst speech quality. However, in case of syllables, good quality speech imposes an enormous acoustic database.

Good quality speech may be also obtained using allophones, but it is crucial that the allophone groups are chosen in a way with possibly smallest number of acoustic units while maintaining the best quality speech.

From the most suitable allophone group point of view, several main factors are of utmost importance. First are articulatory properties of individual speech sounds.

Each language has a limited number of sounds which are uttered in a characteristic manner for a given language. What is more, the tone of an each speech sound is dependant on its environment; to be more precise, on the preceding and the following sound. Sound dependence on a language is essential, that is why each language may have, more or less, a different classification. Second important factor are acoustic parameters such as energy and duration of a speech sound. An important factor while selecting allophone group is also the level of difficulty with determining the boundary of duration or length of a sound.

Kozłowska B., Rafałko J. and Rybnik M..
ALLOPHONE GROUP SELECTION FACTORS FOR POLISH SPEECH SYNTHESIS.
DOI: 10.5220/0003889804510456
In *Proceedings of the 4th International Conference on Agents and Artificial Intelligence* (ICAART-2012), pages 451-456
ISBN: 978-989-8425-95-9

2 SOUND ARTICULATORY CLASSIFICATION IN POLISH LANGUAGE

Articulatory classification relates to an organ arrangement of articulatory apparatus, so-called articulators, characteristic of a particular speech sound. They comprise: palate, a tongue, lips, a uvula, and vocal ligaments. Depending on a kind of a sound classification criteria differ a bit.

In principle, sounds in Polish language may be divided into 3 groups: vowels, consonants and semivowels, on the understanding that "l" and "ł" are semivowels.

Common features of the vowels are sonority, openness and syllable creation. Apart from this, one may group them according to articulatory classification criterion.

Vowel classification

- with regard to horizontal movement of a tongue:
  - front – e, ę, i, y
  - central – a
  - back – ą, o, u
- with regard to vertical movement of a tongue:
  - high – i, u, y
  - mid – ą, e, ę, o
  - low – a
- with regard to central tongue's position:
  - soft – i
  - hard – a, ą, e, ę, u, y
- with regard to lips formation:
  - flat – e, ę, i, y
  - rounded – ą, o, u
  - neutral – a
- with regard to uvula position:
  - oral – a, e, i, o, u, y
  - nasal – ą, ę.

Consonant classification

- with regard to vocal ligaments activity:
  - voiced – b, b', d, d', g, g', z, z', ź, ż, dz, dź, dż, w, w', j, m, m', n, ń, l, l', ł, r, r'
  - voiceless – p, p', t, t', k, k', s, s', ś, sz, c, c', ć, cz, f, f', h, h'
- with regard to a center tongue's position:
  - soft – b', ć, s', ś, z', ź, ć, c', dź, ń, p', f', w', m', l', k', g', h', j
  - hard – b, c, s, z, dz, p, f, w, m, l, k, g, h, t, d, sz, ż, cz, dż, r, ł
- with regard to manner of articulation i.e. speech organs closure:
  - plosive – b, b', p, p', d, t, k, k', g, g'
  - affricate – c, dz ,dź, cz, dź, ć
  - fricative – f, f', w, w', z, s, z', ż, sz, ź, ś, s', h, h'
  - unobstructed – m, m', n, ń, r, l, l' (ł)
- with regard to place of articulation:
  - bilabial – p, b, m
  - labio-dental – f, w
  - apico-dental – t, d, c, dz, s, z
  - alveolar – l, r, sz, ż, cz, dż
  - palatal – ś, ż, ć, dź, ń
  - velar – k, g, h
- with regard to uvula position:
  - nasal – m, m', n, ń
  - oral – the others.

The above-presented articulatory classification will facilitate understanding of further proceedings during the selection of allophone groups. While uttering a word one does not voice each sound separately but "passes on" fluently from one speech sound to the other. In this connection the position of articulators before voicing the particular sound as well as the position they assume to voice another sound is very important.

3 INFLUENCE OF ADJACENT SOUNDS ON SOUND TONE

The basic parameter, on which a speech sound tone depends, is adjacent sounds influence. That influence is not equal in every case i.e. some sounds affect their environments to a greater extend than the others.

This very sound tone dependence on its environment determines selection of allophone group and what follows is the number and kind of allophones. Each allophone is described by its environment, left context i.e. the preceding sound, and right context i.e. the following sound. It is presented in figure 1.

Observation of adjacent sounds influence has indicated several important aspects:

1. Strength of influence of adjacent sounds is greater for vowels than consonants. Thus, there is need to create different allophone group classifications for vowels and consonants.
2. Intensity and the way the preceding and the following sounds affect voicing is not the same. The fact that a particular group of the preceding sounds have a similar impact on a speech sound does not mean that the same group, as the following sounds, will influence

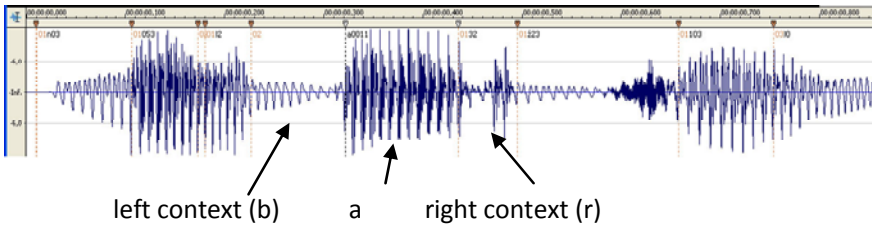

Figure 1: Left and right contexts of allophones.

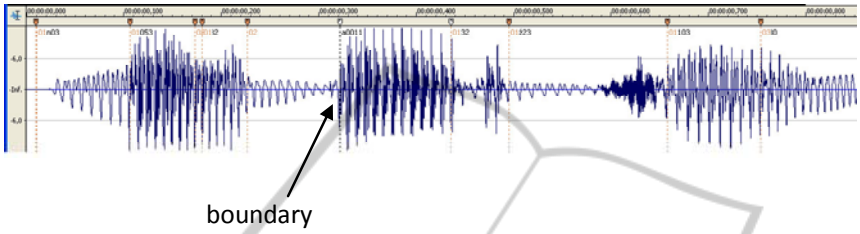

Figure 2: Precise setting the boundary between two allophones.

the sound similarly. Moreover, the strong influence of a given preceding sound does not also imply that it will have the same strength as the following sound. For this reason, left context groups and right context groups may differ.

3. The preceding sound has a lot more greater influence on vowels – left context, than the following sound – right context. Therefore, for vowels, the number of left context groups will be greater than the number of right context groups.
4. The manner one voices adjacent sounds greatly influences the tones of vowels. During the process, central tongue's position (softness, hardness), and the place of articulation (labiality, frontness, centralness and backness) are especially important.
5. In case of consonants, voicedness/voicelessness of adjacent sounds is important. The preceding voiced sounds affect a speech sound similarly, as is the case with the following voiced sounds. The preceding voiceless sounds affect a speech sound similarly, as is the case with the following voiceless sounds.

4 DIFFICULTY LEVEL OF SETTING ALLOPHONE BOUNDARIES

Another crucial factor affecting selection of an allophone group is precise determining allophone boundaries which is indispensable for separating an allophone from a natural speech signal. Incorrectly determined allophone boundaries will cause distortion in synthesized speech.

In most cases precise setting the boundaries is not a difficult task as the points where an allophone begins and ends can be seen clearly. The figure 2 presents an example of clear boundaries between allophones of "a" and "b" sounds.

There are cases, however, in which it is difficult to determine the boundary. It is connected with the influence of some left context sounds on the following vowel. Those sounds make the boundary between them and the vowel indistinct and it is problematic to define the point where one allophone ends and the other begins. As shown in the figure 3. First of all, this very problem concerns sonore sounds i.e. m, n, r, l, ł, j and vowels. Their influence upon the following speech sound causes the boundary, between one sound and the other, to fade away. Transition area appears. One can hear two sounds simultaneously. You cannot link the two sounds in a single group as they affect each other's tones in a unique manner. Taking it into consideration, each of the above-mentioned sounds comprises an individual left context group.

5 IMPORTANCE OF ACOUSTIC PARAMETERS

Acoustic parameters which have greatly affected the selection of allophone groups are energy and the duration of a sound. They are especially important in

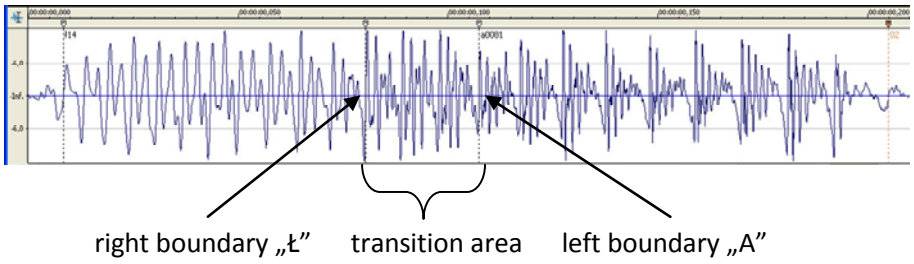

Figure 3: Transition area between two allophones.

case of speech sounds that may be stressed. Polish language vowels are the only sounds that may be stressed. The research performed has portrayed the position of a vowel in relation to a word stress to be essential for the sound energy as well as its duration. It is important whether the vowel is placed before, or after the stress, or if the stress falls exactly on it.

The diagram below presents the average values of vowel allophone durations with the division according to the relation to a word stress. It can be easily noticed that irrespectively of a lecturer the duration time of an each vowel allophone is enormously different. The shortest vowel allophones are located before the stress. The time of duration, of allophones on which the stress falls or the ones that are located in word after the stress, is similar.

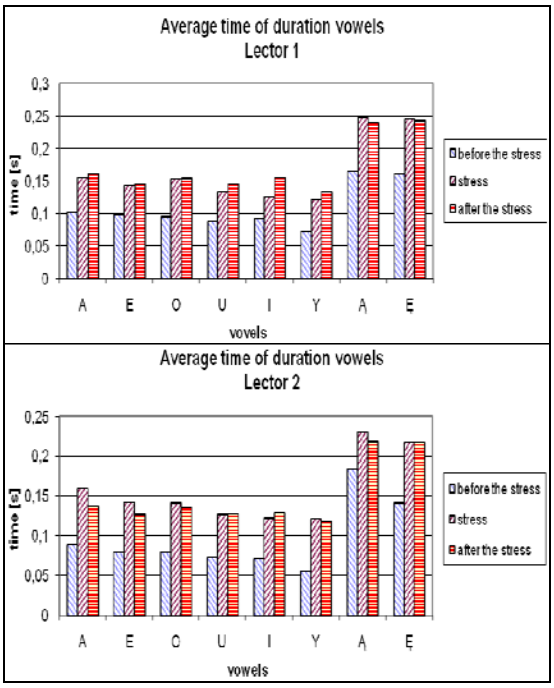

Figure 4: The average values of vowel allophone durations with the division according to the relation to a word stress.

In case of energy, it has been noticed that irrespectively of a lecturer, the highest value falls on vowel allophones located before the stress, and the lowest on allophones after it. The figure 5 also presents the relationships.

Examining the energy and the duration of individual allophones has shown that for the selection of vowel allophone groups it is essential to take into account the position of a vowel in relation to a stress. Thus, it is necessary to add, to the previously presented classification of allophone groups, further division of vowels into vowels located before and after the stress, as well as vowels that are stressed.

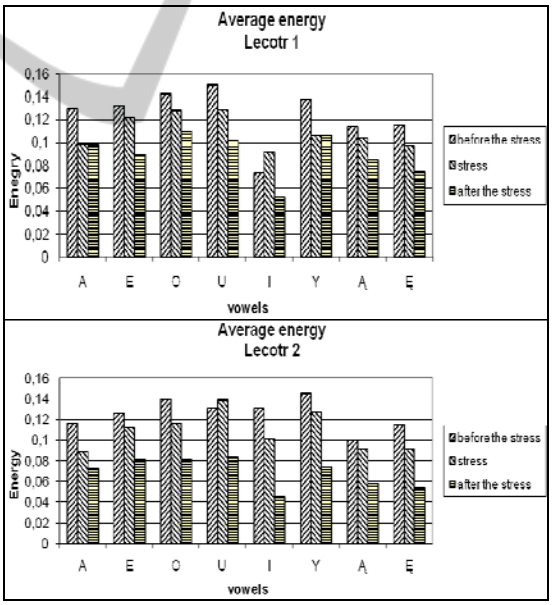

Figure 5: The average energy of vowel allophone with the division according to the relation to a word stress.

6 SUMMARY

In order to select proper allophone groups, it is necessary to take into consideration all the described factors and examine if it is sufficient for the speech synthesis to be distinct, understandable and as

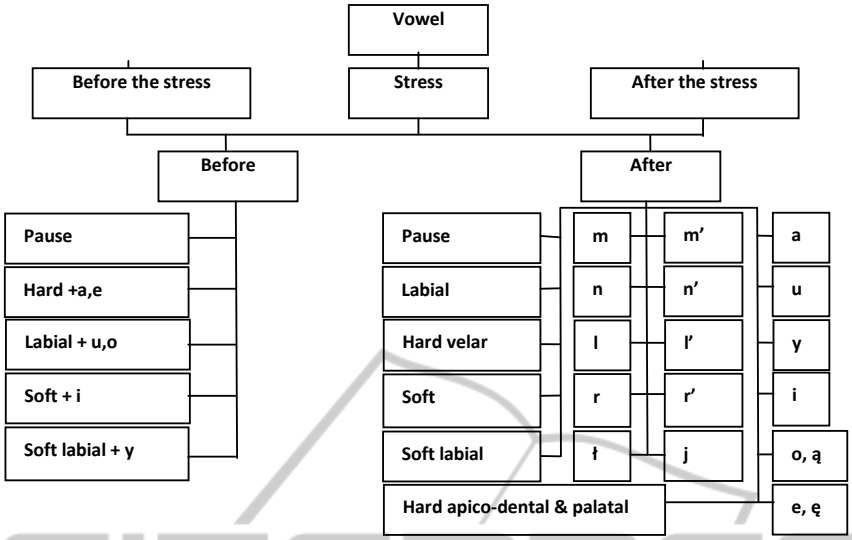

Figure 6: Allophone group classification for vowels.

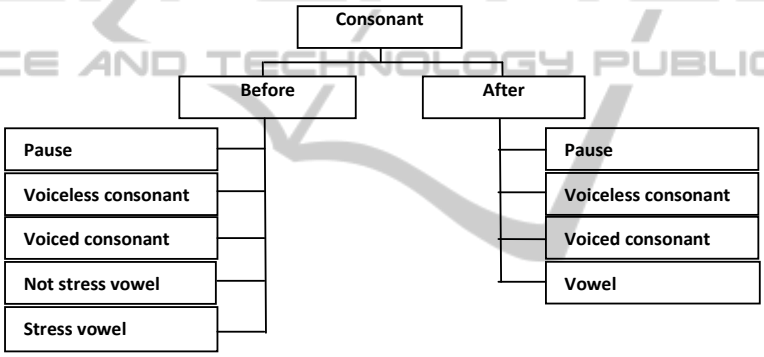

Figure 7: Allophone group classification for consonants.

natural as possible. The figures below, present allophone group classification for vowels (Fig. 6) and consonants (Fig. 7) suggested by the authors. "Before" index means that the allophone is before a given sound group while "after" index means the allophone is after a given sound group. "Pause" means that before/after the allophone there is no sound i.e. the allophone beginning or ending a word.

Such a division, theoretically, produces 3500 allophones. However, one will never achieve the number like this as Polish language specification does not allow for certain allophone clusters. For instance, an "i" allophone will be present only after soft consonants because it always palatalizes the preceding consonant. Moreover, there are allophones, theoretically existent, that in practice have no representation in any word. In search for such allophones approximately 4 million Polish words were examined. Ultimately the acoustic database includes about 2300 allophones for a speaker.

Speech synthesis working on an acoustic database with such classification brings good results. The speech is quite clear and understandable. Obviously the receiver of the certain text message has no doubts it is not natural speech as it is necessary to preserve proper rhythmics and intonation of sentences. However, hearing single words, in most events, it is difficult to assess whether a particular word is synthesized or natural.

REFERENCES

M. Dłuska, *„Fonetyka Polska"*, 1981.

T. Dutoit, "*An Introduction to text-to-speech synthesis"*, *Kluwer Academic Publishers* 1997, pp. 286.

J. Van Santen, R. Sproat, J. Olive, J. Hirshberg, "*Progress in speech synthesis"*, Springer Verlag, New York 1997, Chapter 4, "Concatenative Synthesis and Automated Segmentation", pp. 259–220.

X. Huang, A. Acero, H. Hon, *"Spoken Language Processing"*, Prentice Hall PTR, New Jersey 2001, Chapter 2 "Spoken Language Structure", pp. 19–69.

E. Szpilewski, B. Piórkowska, J. Rafałko, B. Lobanov, V. Kiselov, L. Tsirulnik, *"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.

Piórkowska B., Popowski K., Rafałko J., Szpilewski E., *„Synteza mowy polskiej na podstawie tekstu"*, XI Symposium AES, New Trends in Audio and Video, Conference Program, Abstracts and Proceedings, 20 – 22 september 2006, Bialystok, Poland, pp. 150 – 169.

Piórkowska B., Popowski K., Rafałko J., Szpilewski E., *„Polish Language Speech Synthesis Basis on Text Information"*, New Trends in Audio and Video, vol. I, *Bialystok Technical University*, Treatise Nr 134, 2006, pp. 507 – 526.

T. Taylor, *"Text-to-Speech Synthesis"*, Cambridge University Press 2009, Chapter 11 "Acoustic models of speech production", pp. 309 – 340.