Int J Speech Technol (2015) 18:187–193
DOI 10.1007/s10772-014-9259-z

Using bands of frequencies for vowel recognition for Polish language

Marcin Płonkowski

Received: 19 July 2014 / Accepted: 14 October 2014 / Published online: 26 October 2014

Abstract This paper presents a simple and useful method of vowel recognition of the Polish language. It relies on determining some characteristic bands of frequencies for each vowel. These bands are chosen so as to provide maximum separability of all vowels. Within each band we determine three parameters: average, a standard deviation and a maximum value. Comparing these values with the previously designated boundary values, we can classify a given vowel. As shown by the test, this method has a low percentage of an incorrectly recognized vowel. An additional advantage is its efficiency. It is four times faster than the method based on the formants.

Keywords Vowel recognition · Polish language · Bands of frequencies

1 Introduction

Currently, the most commonly used methods for automatic speech recognition systems are methods based on MFCC. This works very well when we are dealing with continuous speech. However, if we can qualify some parts of speech as vowels [e. g. by determining that a fragment of speech is voiced and based on the duration of the phoneme (Ziółko and Ziółko 2011a)], we can then use formants. This will allow to define the vowel with which we are dealing (Kodandaramaiah et al. 2010). The use of formants allows to obtain very good results with an accuracy of 95 % (Alotaibi and Hussain 2010).

M. Płonkowski (✉)
Institute of Mathematics and Computer Science, The John Paul II Catholic University of Lublin, Al. Racławickie 14, 20-950 Lublin, Poland
e-mail: plomyk@kul.pl

Formants are the commonly used tool in the analysis and recognition of vowels. This is due to the fact that, for each vowel, it is possible to determine a characteristic pattern, which helps differentiate it from other vowels. This classification can successfully be utilised for different speakers, a variable rate of speech or different emotional states. The most important advantage of the formants is their relative stability of using small amounts of information.

In the fifties, it was proposed the representation of the first two formants of English vowels on the plane (Peterson and Barney 1952), which is sufficient to simply distinguish vowels. This information is not sufficient to obtain certain quality parameters of the vowels (e.g. rounding) (Hayward 2000). Therefore, it is suggested to use the first three formants (Ladefoged 2006).

Despite many advantages of formants, there are practical problems with their calculation. Namely, there is not exact number of formants. Good results we obtain using the first three formants that are best seen (Prica and Ilić 2010). Quite well, we can also observe formants F4 and sometimes F5. However, higher formants such as F4 and F5 have lower amplitude in the spectrogram and may be difficult to distinguish (Zhou et al. 2008). Sometimes the distinction of higher formants is important. Espy-Wilson et al. have suggested that the higher formants may contain cues to tongue configuration and vocal tract dimensions (Espy-Wilson and Boyce 1999; Espy-Wilson 2004).

The second problem is due to the fact that distances between some of the formants are small, reaching only a few hundred Hz (Catford 1988). The third problem are typical errors that for F0 $<$ 300 reach 60 Hz. The errors are even greater for F0 at the level of 500–600 Hz (among some women and children) (Traunmüller and Eriksson 1997). In addition, if we consider the fact that the signal contains some

distortions, a precise determination of the formant frequency can be very difficult.

Another important aspect is the efficiency of the process of determining the formants, which is insufficient. For example, the program Praat (Boersma and Weenink 2009), to determine the formant uses the algorithm described in (Lee 1988). For audio samples lasting 5.7 s, it takes 0.5 s to calculate the three formants (on a PC with 4 × 2.5 GHz CPU). If we take into account other elements of the analysis, the time will be longer. Thus, the transfer of this algorithm for mobile devices (e.g., smartphone, which has worse parameters than the PC) makes that recognition not possible in the real time.

In this article the author proposes a new method for recognition of Polish vowels. It is not based on a determination of formant frequencies but it is based on the calculation of statistical parameters in a characteristic range of frequency bands. These parameters allow the recognition of vowels. An additional advantage will be a significant performance increase, characterized by a significant decrease in processing time for this algorithm. For audio samples lasting 5.7 s it takes 125 ms to calculate all parameters, (on a PC with 4 × 2.5 GHz CPU). Which is four times faster than the method based on the formants.

2 Analysis of the frequencies bands

The proposed algorithm, as well as the vast majority of vowel recognition methods, use the information contained in the frequency domain. Calculations were based on samples of 6 vowels of the Polish language (a, e, i, o, u, y). Vowels were recorded by 24 speakers (10 women and 14 men, aged 20–60 years) in a 16-bit quality. Test set consists of a utterances six people (three woman and three man). Recordings were made in the home environment, using a dynamic microphone.

The analysis was performed using the FFT algorithm. The window size was set to last 0.025625 second. This is a typical value used in programs for speech recognition. For example, the program Sphinx-3 defines this value (CMU Sphinx 2014). The sampling rate was set to 16,000 Hz. Thus, each frame contains 410 samples. The frame shift is 160 samples. The FFT size parameter will be set to 512. Thus, the frequency resolution of each spectral line is equal to 31.25 Hz.

Before performing the FFT algorithm, each frame is multiplied by a Hamming window. The result of the FFT algorithm are complex numbers, hence we need to calculate the magnitude of each of the numbers (the phase information was rejected).

Magnitude values are absolute values hence it is difficult to compare them directly. A process of scaling is, therefore, necessary. To do this, we first calculate the average value (arithmetic mean) of the amplitude of the whole band for each frame. The average value represents the average energy which is contained in a given frequency band. Then we scale each value as the difference compared to the average:

```
for (int j = 0; j < tab.Length; j++)
{
  tab[j] = (tab[j] - mean) / mean;
}
```

Listing 1. Scaling with respect to the arithmetic mean.

Taking into account the specificity of the FFT algorithm, the size of the array is 257 elements. Of course, we can reduce the scope of the analysis, because, in principle, above 5,000 Hz information from the viewpoint of vowel recognition is no longer important. Such a range is widely used (Ladefoged and Ferrari Disner 2012). Then, on the basis of utterances of 24 persons speaking each vowel several times, an average frame representing a given vowel was calculated.

Observing the graphs of scaled values (for the averaged vowel some important bands can be observed.

This and the following charts show only band to 2,500 Hz, where you can see the most important changes. The program analyzed the entire bandwidth up to 8,000 Hz.

From the above chart we can see, that the most characteristic band of the vowel *a* is the band between 1,000 and 1,600 Hz. It is clear that the vowel *a* stands out in this range. The analysis confirms that the most important range is between 1,125 and 1687.5 Hz.

In the first stage of recognition, we will test whether the frame corresponds to the vowel a. If the answer is yes, we finish checking. If the answer is no, then check to see what

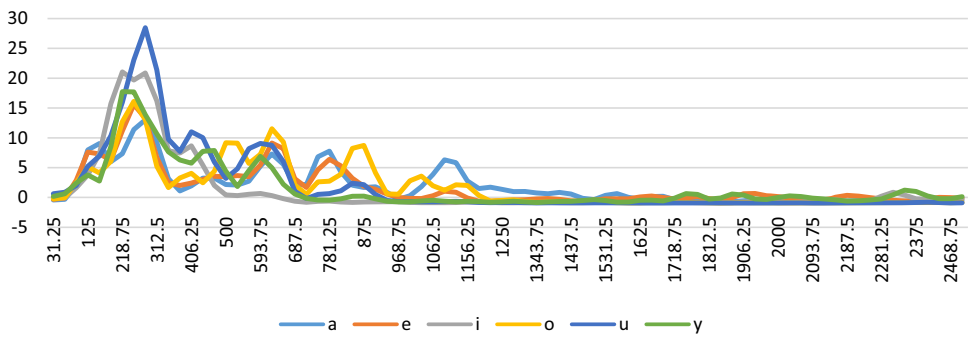

Fig. 1 The amplitudes of the frequency of vowels: a, e, i, o, u, y

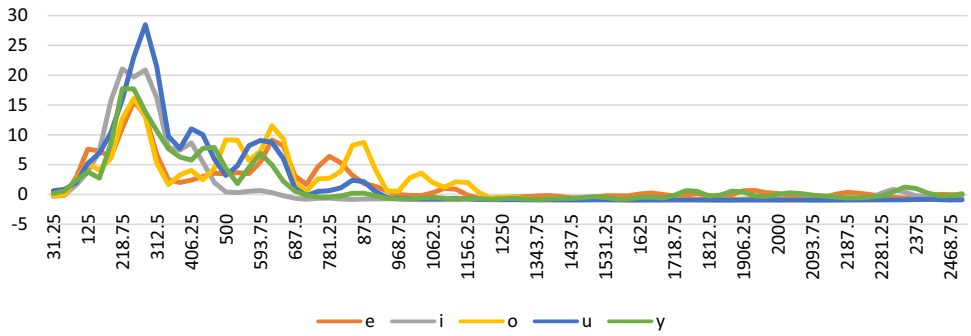

Fig. 2 The amplitudes of the frequency of vowels: e, i, o, u, y

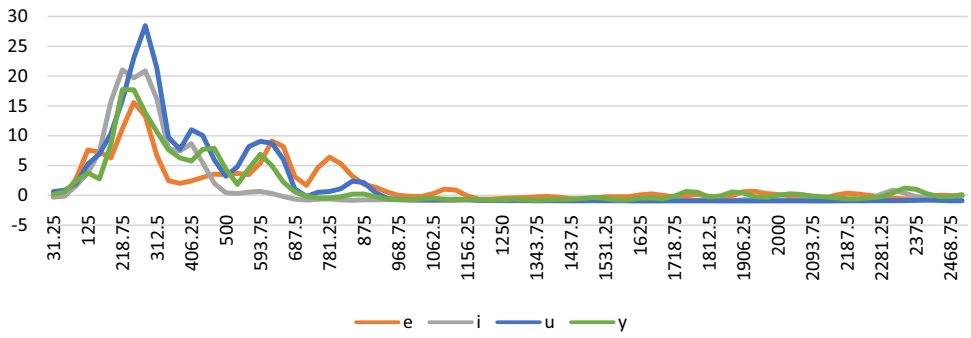

Fig. 3 The amplitudes of the frequency of vowels: e, i, u, y

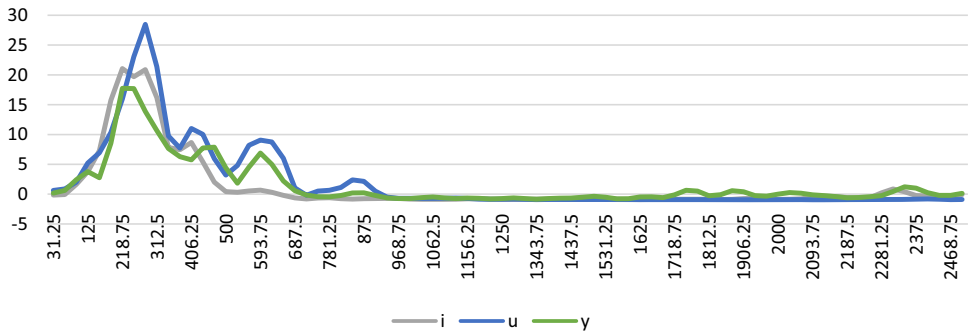

Fig. 4 The amplitudes of the frequency of vowels: i, u, y

other vowel, this frame can contain. Hence, we do not show the characteristics of the vowel a on the next graph. Makes it easier to observe the differences between the other vowels (Fig. 1).

The next chart shows the amplitude of frequency bands without the vowel *a*.

In Fig. 2, we can see that the vowel *o* stands out in the band 800–1,200 Hz. The analysis confirms that the most important range is the range from 843.75 to 1,375 Hz.

The next chart shows the amplitude of the frequency bands without the vowel *o*.

In Fig. 3, we can see that the vowel *e* stands out in the band 700–1,200 Hz.

The next chart shows the amplitude of the frequency bands without the vowel *e*.

In Fig. 4, we can see that the vowel *u* stands out in the band 500–900 Hz and the vowel *y* stands out in the band 1,700–2,100 Hz. During the detailed analysis, it has turned out that it is easier to separate the vowel *y*. Therefore, we choose a range of 1718.7–2093.75 Hz.

The next chart shows the amplitude of frequency bands for the vowels *i* and *u*.

In Fig. 5, we can see that the vowels *u* and *y* can be distinguished in several bands.

A detailed analysis will determine the best band in the 593.75–968.75 Hz.

3 Determination of the characteristic bands for vowels

We will want to confirm the detailed analyses on the basis of the above observations. Vowels are represented by the average value scaled as in listing 1. In addition, we will extend these values by the standard deviation. By subtracting the standard deviation we create the lower limit fluctuations, and by adding we create the upper limit fluctuations. Searching the best band separability, we will look for the biggest differences between the maximum and minimum values (in the selected range).

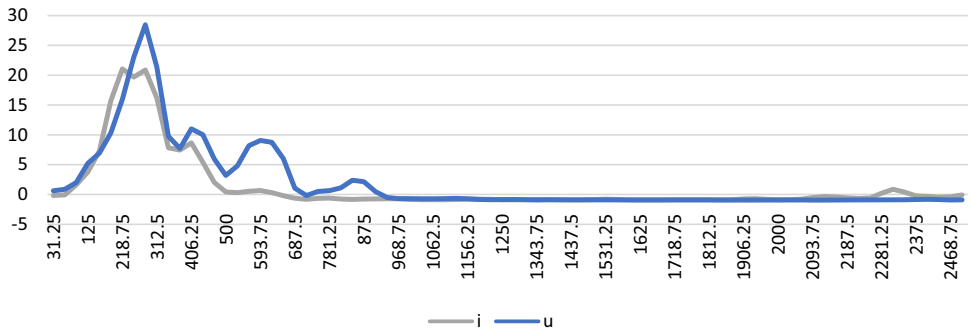

Fig. 5 The amplitudes of the frequency of vowels: i, u

Table 1 Distances between vowel a and extreme values of the other vowels

	Distance	Band (Hz)
Average	0.825	1125–1437.5
SD	0.792	1125–1687.5
Maximum value	0.953	1218.75–1687.5

Table 2 Distances between vowel o and extreme values of the other vowels

	Distance	Band (Hz)
Average	0.688	843.75–1156.25
SD	0.245	906.25–1375
Maximum value	–	–

Table 3 Distances between vowel e and extreme values of the other vowels

	Distance	Band (Hz)
Average	0.773	687.5–1937.5
SD	0.426	937.5–1187.5
Maximum value	0.392	1156.25–1593.75

Table 4 Distances between vowel y and extreme values of the other vowels

	Distance	Band (Hz)
Average	0.582	1750–2093.75
SD	0.748	1718.75–2062.5
Maximum value	0.787	1781.25–2093.75

Table 5 Distances between vowel u and extreme values of the other vowels

	Distance	Band (Hz)
Average	2.014	593.75–937.5
SD	1.083	625–968.75
Maximum value	1.296	656.25–937.5

These differences will be computed in three areas. The first parameter is the average energy in a given frequency band. The second parameter is the standard deviation. And the third parameter is the maximum value of a given band. Just as described above, we begin the analysis of the separate *a* vowels. The following table presents the distances between the vowel *a* and extreme values of other vowels (Table 1).

The next analysed vowel is *o*. Table 2 presents the distances between the vowel *o* and extreme values of other vowels.

The blank value in the last row indicates that there has been found no margin separability for this parameter.

The next analyzed vowel is e. Table 3 presents the distances between vowel e and extreme values of the other vowels.

The next analyzed vowel is y. Table 4 presents the distances between vowel y and extreme values of the other vowels.

The next analyzed vowel is u. Table 5 presents the distances between vowel u and extreme values of the other vowels.

4 Determination of the boundary values for vowels

Another important step is to find an appropriate boundary value at which we recognize a given vowel. If the value is too high, then, some correct instance of vowels will not be recognized. If the value is too low, other vowels will be inappropriately classified.

For example, in our analysis, the vowel a has the mean energy value (appropriately scaled in accordance with the Listing 1) equal to 0.328 (in the band of 1,125–1437.5 Hz). The remaining vowels do not exceed the value of −0.497 in this band. Thus, the difference is 0.825, as shown in Table 1.

Now the key is, which value in the range of −0.497 to 0.328 ensures the lowest percentage of errors. As you can see in the table below this value is −0.39. In the same way

Table 6 Boundary values for the individual parameters of vowel a

	Boundary value	Band (Hz)
Average	−0.39	1,125–1437.5
SD	1.01	1125–1687.5
Maximum value	3.12	1218.75–1687.5

Table 7 Boundary values for the individual parameters of vowel o

	Boundary value	Band (Hz)
Average	0.87	843.75–1156.25
SD	0.89	906.25–1,375
Maximum value	–	–

Table 8 Boundary values for the individual parameters of vowel e

	Boundary value	Band (Hz)
Average	0.22	687.5–1937.5
SD	0.29	937.5–1187.5
Maximum value	−0.31	1156.25–1593.75

Table 9 Boundary values for the individual parameters of vowel y

	Boundary value	Band (Hz)
Average	−0.79	1,750–2093.75
SD	0.09	1718.75–2062.5
Maximum value	−0.66	1781.25–2093.75

we have set boundary values for the other parameters and vowels.

We find the appropriate value using a statistical analysis, which minimizes the amount of erroneous recognition. This value will be characterized by the lowest percentage of errors.

Table 6 shows values that have been selected for individual parameters. If the values are higher for all three parameters, we consider a given vowel as recognized. Otherwise, go to look for another possible vowel.

Table 7 shows values that have been selected for individual parameters of the vowel *o*. The maximum value was not found for vowels *o*. Hence, this parameter is omitted. However, for successive vowels all three parameters must be met.

Table 8 shows the values that have been selected for the individual parameters of the vowel e.

Table 9 shows the values that have been selected for the individual parameters of the vowel y.

Table 10 shows the values that have been selected for the individual parameters of the vowel u.

Table 10 Boundary values for the individual parameters of vowel u

	Boundary value	Band (Hz)
Average	0.26	593.75– 937.5
SD	0.15	625–968.75
Maximum value	−0.19	656.25–937.5

Table 11 Number of incorrectly recognized vowels using the individual parameters

Vowels	Percentage of incorrectly recognized vowels (%)	Parameters taken into account (average, standard deviation, maximum value)
a	2.8148	Average, maximum value
o	4.8689	Average
e	3.3002	Average, standard deviation, maximum value
y	0.5824	Average, standard deviation
u	1.2775	Average
i	0.1896	Average

5 The correctness of the vowel recognition

The last part of the article is a summary showing the number of correctly and incorrectly recognized vowels (on the basis of test data).

The recognition process starts by checking whether the frame corresponds to the vowel a. If not, then we move on to check whether the same frame contains the vowel o (subsequently, in the case of a negative answer we check the vowel e, etc.). If the answer is yes, we finish the checking. In subsequent steps of the testing frame, we have to check less vowels. For example, checking whether the frame contains the vowel y, we do not check whether the frame contains the vowels previously tested. Therefore, the results differ quite substantially.

The following table shows the amounts of wrongly recognized vowels (in percentage terms).

In addition, the third column in Table 11 contains the parameters taken into account (based on the analysis of all possibilities).

For the vowel *a* it turned out that the standard deviation parameter adds nothing of substance to the incorrectly recognized vowels. Therefore, we analyse only the other two parameters (average and maximum value).

The same is also done for other vowels. Namely, we have analysed the contribution of all parameters to the quality of the vowel recognition. It turned out that sometimes we only need the same average for obtaining the best possible result

(vowels *o* or *u*). In contrast, it is sometimes necessary to use all three parameters (the vowel *e*).

6 Comparison with other research results

There are many publications, that address the problem of speech recognition for the Polish language (Tadeusiewicz 1988; Ziółko and Ziółko 2011b). Many of the papers make use of packages such as HTK (Ziółko et al. 2008; Pawlaczyk and Bosky 2009) and Sphinx (Janicki and Wawer 2011; Płonkowski and Urbanovich 2014).

Although, there are many similarities between Polish language and English language, the English is a very typical positional language and the Polish is highly inflective. The Polish language is one of the 25 most influential languages in the world (List25.com 2014). Therefore, we need studies that relate to the specifics of the Polish language (Ziółko 2009).

Our results are similar to results obtained by other researchers for the Polish language. Pietruch et al. (2009) obtained the recognition results of vowels at the level of 98 % (using formants and artificial neural networks). Similar results were obtained by researchers for other languages. Alotaibi (2012) obtained recognition results of vowels at the level of 92.13 % for Arabic (using lpc and artificial neural networks). Koulagudi et al. (2012) obtained recognition results of vowels at the level of 91.4 % for Hindi (using MFCC). Thakur et al. (2012) received the recognition results of vowels at the level of 93.2 % for English (using MFCC).

7 Summary

In this paper a new method of vowel recognition has been presented. It is based on the designation of frequency bands for each of the vowels, which best separate one from another. It has turned out that this simple method gives quite good results and allows us to distinguish vowels with a small margin of error (below 5.0 %).

An additional advantage of this method is its speed (four times faster than the method based on formants) which predisposes it to be used for mobile devices.

References

Alotaibi, Y. A. (2012). Comparing ANN to HMM in implementing limited Arabic vocabulary ASR systems. *International Journal of Speech Technology*, *15*(1), 25–32.

Alotaibi, Y. A., & Hussain, A. (2010). Comparative analysis of arabic vowels using formants and an automatic speech recognition system. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, *3*, 11–22.

Boersma, P., Weenink, D., (2009). Praat: Doing phonetics by computer. Computer Software.

Catford, J. C. (1988). *A practical tntroduction to phonetics*. Oxford: Oxford University Press.

CMU Sphinx (2014). The Carnegie Mellon Sphinx Project. Retrieved from http://cmusphinx.sourceforge.net/ June 2014.

Espy-Wilson, C. Y. (2004). *Articulatory strategies, speech acoustics and variability* (pp. B62–B76). Proceedings of Sound to Sense: Fifty+ Years of Discoveries in Speech Communication.

Espy-Wilson, C. Y., & Boyce, S. E. (1999). The relevance of F4 in distinguishing between different articulatory configurations of American English /r. *The Journal of the Acoustical Society of America*, *105*, 1400. doi:10.1121/1.426610.

Hayward, K. (2000). *Experimental phonetics*. Harlow, UK: Pearson.

Janicki, A., Wawer, D. (2011). Automatic speech recognition for Polish in a computer game interface. Proceedings of the federated conference on computer science and information systems (FedCSIS 2011), Szczecin, p. 711–716.

Kodandaramaiah, G. N., Giriprasad, M. N., Rao, M. M., (2010). Independent speaker recognition for native english vowels. *International Journal of Electronic Engineering Research*, vol. 2.

Koulagudi, S.G., Thakur, S.N., Barthwal, A., Singh, M.K., Rawat, R., Sreenivasa, R.K. (2012). Vowel recognition from telephonic speech using MFCCs and Gaussian mixture models. Communications in Computer and Information Science, 305 CCIS, pp. 170–177.

Ladefoged, P. (2006). *A course in phonetics* (5th ed.). Boston, MA: Thomson Wadsworth.

Ladefoged, P., & Ferrari Disner, S. (2012). *Vowels and consonants*. New York: Wiley.

Lee, C. H. (1988). On robust linear prediction of speech. *IEEE Transactions on ASSP*, *36*, 642–649.

List25.com. (2014). http://list25.com/the-25-most-influential-languages-in-the-world/, June 2014.

Pawlaczyk, L., & Bosky, P. (2009). Skrybot—A system for automatic speech recognition of Polish language. *ICMMI*, *2009*, 381–387.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in the study of the vowels. *The Journal of the Acoustical Society of America*, *24*, 175–184.

Pietruch, R., Grzanka, A., & Konopka, W. (2009). Vowels recognition using video and audio data with an application to laryngectomees' voice analysis. In 16th International Congress on Sound and Vibration. *Cracow*, 2009, 1–8.

Płonkowski, M., & Urbanovich, P. (2014). Tuning a CMU Sphinx-III speech recognition system for Polish language. *Przegląd Elektrotechniczny*, *4*(2014), 181–184.

Prica, B., & Ilić, S. (2010). Recognition of vowels in continuous speech by using formants. *Series: Electronics and Energetics*, *23*(3), 379–393.

Tadeusiewicz, R. (1988). *Sygnał Mowy*. Warszawa: Wydawnictwo Komunikacji I Łkaczności.

Thakur, S. N., Singh, M. K., Barthwal, A. (2012). Telephonic Vowel Recognition in the Case of English Vowels. IC3 2012, pp. 500–501.

Traunmüller, H., & Eriksson, A. (1997). A method of measuring formant frequencies at high fundamental frequencies. *Proceedings of EuroSpeech'97*, *1*, 477–480.

Zhou, X., Espy-Wilson, C. Y., Boyce, S., Tiede, M., Holland, C., & Choe, A. (2008). A magnetic resonance imaging-based articulatory and acoustic study of, retroflex" and, bunched" American English /r/. *The Journal of the Acoustical Society of America*, *123*, 4466–4481.

Ziółko, B. (2009). Speech Recognition of Highly Inflective Languages. PHD thesis. Department for Computer Science, University of York, p. 122.

Ziółko, B., & Ziółko, M. (2011). Time durations of phonemes in Polish language for speech and speaker recognition. Human language technology. Challenges for Computer Science and Linguistics. *Lecture Notes in Computer Science*, *6562*(2011), 105–114.

Ziółko, B., & Ziółko, M. (2011). *Przetwarzanie mowy*. Krakow: Wydawnictwa AGH.

Ziółko, B., Manandhar, S., Wilson, R. C., Ziółko, M., Gałka, J. (2008). Application of HTK to the Polish Language. Proceedings of IEEE International Conference on Audio, Language and Image Processing, Shanghai.