ARCHIVES OF ACOUSTICS
Vol. **44**, No. 4, pp. 693–707 (2019)
DOI: 10.24425/aoa.2019.129725

Comparison of Lithuanian and Polish Consonant Phonemes Based on Acoustic Analysis – Preliminary Results

Gražina KORVEL[(1)*], Olga KURASOVA[(1)], Božena KOSTEK[(2)]

[(1)] *Institute of Data Science and Digital Technologies, Vilnius University*
Lithuania
[*]Corresponding Author e-mail: grazina.korvel@mii.vu.lt

[(2)] *Audio Acoustics Laboratory*
Faculty of Electronics, Telecommunications and Informatics
Gdańsk University of Technology
Poland

(*received February 25, 2019; accepted August 23, 2019*)

The goal of this research is to find a set of acoustic parameters that are related to differences between Polish and Lithuanian language consonants. In order to identify these differences, an acoustic analysis is performed, and the phoneme sounds are described as the vectors of acoustic parameters. Parameters known from the speech domain as well as those from the music information retrieval area are employed. These parameters are time- and frequency-domain descriptors. English language as an auxiliary language is used in the experiments. In the first part of the experiments, an analysis of Lithuanian and Polish language samples is carried out, features are extracted, and the most discriminating ones are determined. In the second part of the experiments, automatic classification of Lithuanian/English, Polish/English, and Lithuanian/Polish phonemes is performed.

Keywords: acoustic analysis; consonant phonemes; acoustic parameters; machine learning methods.

1. Introduction

The state-of-the-art methods applied to speech technology are mostly based on the extraction of parameters and machine learning. Recently, also, deep learning is applied to automatic speech recognition (ASR) (Bourlard, 2018; Korvel *et al.*, 2018; Padmanabhan, Premkumar, 2015). The acoustic parameters of the speech signal are widely used for various tasks, such as speech or speaker recognition, emotion recognition, phoneme modeling, and speech analytics. The goal of this research is to determine a vector of acoustic parameters, that is related to the most distinctive differences between Polish and Lithuanian consonants and then compared with English as an auxiliary language.

In the literature, we can find a description of various parameterization techniques and various modification of standard techniques. The popular implementations of the Mel Frequency Cepstral Coefficients (MFCCs), the Linear Prediction Cepstral Coefficients (LPCCs) and perceptual linear prediction (PLP) parameters (Chia *et al.*, 2012; Upadhya *et al.*, 2018; Eringis, Tamulevicius, 2015). The attention of researches focused on fractal features, pitch, intensity, formants, autocorrelation, noise-to-harmonics ratio, the harmonics-to-noise ratio (Bageshree *et al.*, 2012; Noroozi *et al.*, 2017; Spangler *et al.*, 2017; Taylor *et al.*, 2017). Our previous experiments show that using standard speech parameters along with parameters from the music area gives better phoneme recognition accuracy (Korvel, Kostek, 2017a; Korvel *et al.*, 2019). Therefore, we use standard speech and music domain-derived parameters for speech parametrization in this research study. The analyzed speech signal parameters are time- and frequency-domain features. It should be noted that there are also approaches without performing parameter extraction. For example, this process is discarded (Badshah *et al.*, 2017; Deng *et al.*, 2010) for Deep Neural Networks (DNNs). However, in the context of inter-language research, a thorough analysis of individual spoken elements needs to be performed as there is basic knowledge still missing in this context.

In linguistics, the phoneme is defined as the minimum unit of sound (Gibbon *et al.*, 1997; Gut, 2014). According to Girdenis (2003), two or more sounds are considered as separate phonemes if, in substituting one for the other in at least one position, the meaning of the word changes. A phoneme may contain several phones, e.g., phoneme /p/ can be produced with aspiration or without aspiration in English. Another example could be the phoneme /l/, which can be stressed or not stressed in Lithuanian. These phones are called allophones. In this research, only the phonemes are however analyzed, even though allophonic analysis becomes of interest recently (Czyżewski *et al.*, 2017; Kostek *et al.*, 2017; Kozierski *et al.*, 2016; Mitterer *et al.*, 2018; Recasens, 2012). This is because we believe that the analysis of speech sounds which are acoustically similar to one another and analysis of those which are not acoustically similar are two different tasks. The uttered words will be transcribed into phonemes. For this purpose, the phonetic alphabet is used. One of the widely used phonetic alphabets is the International Phonetic Alphabet (IPA) (Decker, 1999). The IPA is designed to represent qualities of speech that are distinctive in spoken language: phonemes, intonation, and the separation of words and syllables. Due to the special IPA fonts, a machine-readable version of this alphabet has been created. This alphabet, called SAMPA (Gut, 2014; Howard, Murphy, 2007), contains only the symbols that are available on a computer keyboard. Therefore, the symbols of the SAMPA alphabet are used in this research.

The objective of this research is the consonant phoneme signal analysis and in particular a comparison of acoustic resemblance and highlighting the acoustic differences between these chosen languages based on acoustics parameters and two classifiers (k-Nearest Neighbors (kNN) and Support Vector Machine (SVM)).

Generally, the character of vowel phonemes is periodic. Meanwhile, some of the consonant phonemes can be considered as quasi-periodic signals in noise, and others are aperiodic signals. Also, we can divide consonant phonemes into two sets: voiced and voiceless sounds (Domagała, 1994; Krynicki, 2006). The difference between these sets lies in the action of the vocal folds. For voiced sounds, the vocal folds vibrate while saying these sounds, for voiceless they are apart. In general case, the character of consonants is varying, and the consonant phoneme signals are more difficult for processing as those of vowel. This fact is the main reason why broad-spectrum acoustic features are used in this research.

The literature review reveals that little attention (if any) has been paid to differences between Polish and Lithuanian speech acoustical properties even though there are bilingual Lithuanian and Polish speakers having to learn both languages in early childhood (either Lithuanian or Polish being the mother tongue, in some case both languages may be treated as mother tongue). The goal of the study by Labarre (2011) was to show differences between Polish and American English phonology. The study was carried out at the University of Washington by the author having Polish ancestry. In the study of Krynicki (2006), some contrasting aspects of Polish and English phonetics were shown, and adequate examples of such were recalled. Prior to that study, the phonology of Polish was described in many sources (e.g., Gussmann, 2007; Jassem, 2003; Oliver, Szklanny, 2006). It should also be noted that much effort was performed by several Polish and Lithuanian research centers aiming at speech recognition, a few examples of which are given in here: (Kłosowski *et al.*, 2014), analysis of acoustics speech properties (Izydorczyk, Kłosowski, 2001), adaptation of foreign language speech recognition engines for Lithuanian speech recognition (Rudzionis *et al.*, 2009; Kasparaitis, 2008), development of phonemic language corpus for Polish (Kłosowski, 2017) by employing automatic grapheme-to-phoneme conversion of the source orthographic language corpus, obtained from the National Corpus of Polish (NCP) (Przepiórkowski *et al.*, 2012), creating Polish phoneme statistics (Ziółko *et al.*, 2009; 2014), etc.

In this research, it is believed that the acoustic analysis of not closely-related languages let us identify the most prominent features which can be used to distinguish differences between languages. Moreover, the optimized feature vector will serve us as a multidimensional quality assessment applied to the synthesized phonemes. Some preliminary work was already performed towards this direction (Korvel, Kostek, 2017b; Korvel *et al.*, 2019). Discovering acoustic differences in speech is justified by its numerous possible uses. The following can be named: speech synthesis, speech and speaker recognition, transcription of sounds, helping with pronunciation and learning foreign languages, studies in linguistic, medical field.

2. Review of Lithuanian and Polish phonemes

This section discusses the relationship between graphemes and phonemes of languages chosen for our research. The basic units of text are graphemes. Lithuanian language consists of 32 graphemes: a, ą, b, c, č, d, e, ę, ė, f, g, h, i, į, y, j, k, l, m, n, o, p, r, s, š, t, u, ū, ų, v, z, ž and covers 20 consonants. The Polish language is also based on the set of 32 graphemes: a, ą, b, c, ć, d, e, ę, f, g, h, i, j, k, l, ł, m, n, ń, o, ó, p, r, s, ś, t, u, w, y, z, ź, ż, but includes 23 consonants. Some researchers used graphemes in speech recognition systems (Lileikytė *et al.*, 2016; Gales *et al.*, 2015). However, in most studies, grapheme to phoneme conversion is performed, especially in the text-to-speech task. The conversion is made because of the fact the uttered signal is represented by phonemes. Typically, lexicons are utilized to map graphemes to phonemes. Different researchers propose to employ different sets of phonemes for the same language. It should be noted that the size of the phoneme set depends on the task to be solved. For example, a set of phonemes used for speech synthesis is bigger than the ones used for speech recognition. Lithuanian language phonemes have been studied by Girdenis (2003). Lithuanian is described by the author as having 43 consonant phonemes. All these phonemes are unstressed. A set of phonemes appended by stressed phonemes and compound diphthongs is given in Kasparaitis' work (Kasparaitis, 2005). This set was also used in Liepa – Lithuanian speech corpus (Laurinciukaite *et al.*, 2018). Both mentioned authors assumed that a phoneme becomes two new phonemes over time through palatalization. Lithuanian phoneme sets of different size are given and tested by Greibus *et al.* (2017) in the context of speech recognition. The experiment results show that the Baseline phoneme set (set without palatalization and stress) outperformed other sets. The consonant phonemes of this set appended with the examples of their usage by the authors of this paper are used in this research. These phonemes are given in Table 1.

Table 1. Lithuanian consonant phonemes.

SAMPA symbol	Example	Transcription
b	būdas	bu:das
ts	caras	tsaras
tS	čarškalas	tSarSkalas
x	choras	xoras
d	darbas	darbas
dz	Dzukija	dzukija
dZ	džaulis	dZaulis
f	forma	forma
g	gamta	gamta
G	herbas	Gerbas
j	jūra	ju:ra
k	katinas	katinas
l	lapas	lapas
m	maras	maras
n	namas	namas
p	pažymys	paZi:mi:s
r	ratas	ratas
s	statiniai	statiniai
S	šaka	Saka
t	tapyba	tapi:ba
v	vasara	vasara
z	zuikis	zuikis
Z	žodynas	Zodi:nas

In terms of Polish consonants, Labarre (2011) distinguished 36 contrastive consonant phonemes. The author only distinguished bilabial palatalized consonants, disregarding the palatalization of non-labial consonants. The phonetic alphabet described by Demenko and her collegaues (Demenko *et al.*, 2003) is commonly used by Polish researchers (Ziółko *et al.*, 2009; Igras *et al.*, 2013). This alphabet is also used in this paper (see Table 2).

Table 2. Polish consonant phonemes (Demenko *et al.*, 2003).

SAMPA symbol	Example	Transcription
p	pik	pik
b	byt	byt
t	test	test
d	dym	dym
k	kat	kat
g	gen	gen
c	kiedy	cjedy
J	giełda	Jjewda
f	fan	fan
v	wilk	vilk
s	syk	syk
z	zbir	zbir
S	szyk	Syk
Z	żyto	Zyto
s'	świt	s'fit
z'	źle	z'le
x	hymn	xymn
t^s	cyk	t^syk
d^z	dzwon	d^zvon
t^S	czyn	t^Syn
d^Z	dżem	d^Zem
t^s'	ćma	t^s'ma
d^z'	dźwig	d^z'vik
m	mysz	myS
n	nasz	naS
n'	koń	kon'
N	pęk	peNk
l	luk	luk
r	ryk	ryk
w	łyk	wyk
j	jak	jak

As we see from Tables 1 and 2, Lithuanian and Polish languages share many of the same consonants and spell them similarly. Despite this, the shared phonemes may have different articulation. A comparison of acoustic resemblance and highlighting the acoustic differences between these languages is the goal of this research.

For phoneme encoding, the SAMPA symbols were used (Tables 1 and 2). The application of SAMPA is extended to 24 languages. Polish and English languages are also part of them [SAMPA En, SAMPA Pl]. For Lithuanian speech, SAMPA recommendations proposed by Raškinis *et al.* (2003) were used.

3. Parameters extraction

In order to extract inter-language differences, it is important to find a suitable parametric description of the speech signal. We investigate an extensive set of parameters included time- and frequency-domain features. These parameters are descriptors from the speech as well as music domains. Before parameter extraction, signal pre-processing is carried out. Let $\mathbf{x} = (x_1, x_2, ..., x_N)$ be equidistant samples of the speech signal. These samples are normalized according to the formula:

$$y_n = \frac{x_n}{|\max(x_1, x_2, ..., x_N)|}, \tag{1}$$

where $n = 1, ..., N$.

The speech signal is divided into overlapping frames, length of which – M samples (M is the power of 2). Let $\mathbf{y} = (y_1, y_2, ..., y_M)$ be elements of such an interval. An overlap between successive windows is equal to 50%.

The time-domain parameters are extracted directly from the samples of the audio signal. As mentioned before, the character of the consonant signals is varying. In order to measure the differences between two languages, Root Mean Square (RMS) energy is calculated. This parameter gives a lower value for the unvoiced segment than that for the voiced segment and can be expressed as follows:

$$\text{RMS} = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(y_i)^2}. \tag{2}$$

Equation (2) provides the RMS energy of the signal. We will use this value within the extraction process of most of the temporal parameters.

Next two temporal parameters that we use in our research are Temporal Centroid (TC) and Zero Crossing Rate (ZC). First of them (TC) is time average over signal energy envelope and is given by the following expression:

$$\text{TC} = \frac{\sum\limits_{i=1}^{M} i(y_i)^2}{\sum\limits_{i=1}^{M}(y_i)^2}. \tag{3}$$

The second parameter (ZC) is the number of the signal crossing the time axis. The formula of this parameter is as follows:

$$\text{ZC} = \frac{\sum\limits_{i=2}^{M}|s_i - s_{i-1}|}{M-1}, \tag{4}$$

where

$$s_i = \begin{cases} 1, & \text{if} \quad y_i > 0, \\ 0, & \text{if} \quad y_i \le 0. \end{cases} \tag{5}$$

Also, the so-called 'dedicated' parameters proposed by Kostek and her co-workers (Kostek *et al.*, 2011) are calculated. The dedicated parameters are based on the analysis of the distribution of sound sample values in relation to RMS. The following sets of parameters are calculated:

- k_1, k_2, k_3 – the number of samples exceeding levels RMS, $2 \times$ RMS, $3 \times$ RMS. Parameters contained in this group are the values resulting from the entire segment analysis.
- Peak to RMS – calculated as the mean value of the ratio calculated in 10 sub-frames.
- p_1, p_2, p_3, p_4 – the mean value of the signal crossings in relation to zero, RMS, $2 \times$ RMS, $3 \times$ RMS averaged for 10 sub-frames.
- q_1, q_2, q_3, q_4 – the variance of the signal crossings in relation to zero, RMS, $2 \times$ RMS, $3 \times$ RMS averaged for 10 sub-frames.

A graphical representation of levels RMS, $2 \times$ RMS, $3 \times$ RMS for the /k/ and /g/ phoneme entire segments is shown in Figs 1 and 2, respectively.

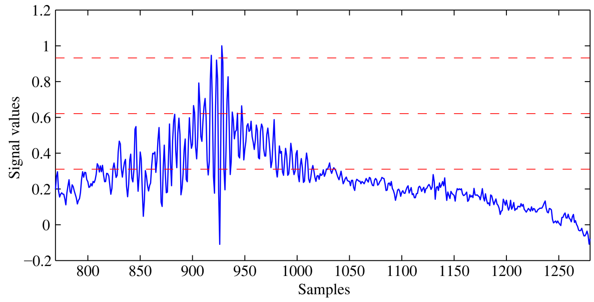

Fig. 1. An example of the phoneme /k/ segment. RMS of this segment is 0.3107.

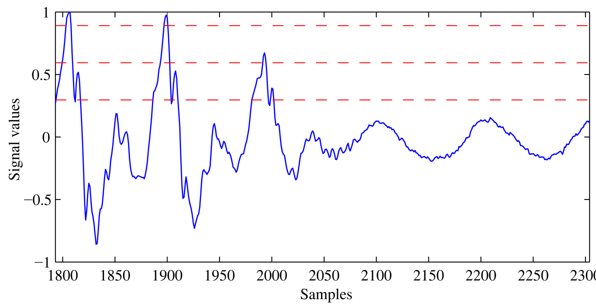

Fig. 2. An example of the phoneme /g/ segment. RMS of this segment is 0.2970.

In order to obtain parameters of the spectrum, we compute the Discrete Fourier transform of each segment:

$$\text{FT}(k) = \sum_{m=1}^{M} y_m w_m e^{(-2\pi i)(m-1)\frac{(k-1)}{M}}, \tag{6}$$

where $\mathrm{FT}(k)$ ($k = 1, ..., M_{\mathrm{FT}}$) are Fourier transform coefficients, M_{FT} denotes the number of Fourier transform coefficients ($M_{\mathrm{FT}} \geq M$, M_{FT} is an integer power of 2), w_m is the window function.

The power spectrum is given by the following formula:

$$\mathrm{PS}(k) = \frac{1}{M_{\mathrm{FT}}}\sqrt{(\mathrm{FT}(k))^2_{re} + (\mathrm{FT}(k))^2_{im}}, \tag{7}$$

where $k = 1, ..., M_{\mathrm{FT}}$, re means a real part, and im – an imaginary part.

The first group of spectrum descriptors is spectral shape parameters. We extracted spectral shape parameters based on the MPEG-7 audio content description standard (Kim *et al.*, 2005). By this standard, the parameters are defined on the log-frequency power spectrum, and these measures are based on an octave frequency scale centered at 1 kHz. Our previous research showed that applying standard speech parameters along with descriptors coming from music information retrieval (MIR) to the phoneme analysis gives better results (Korvel *et al.*, 2019). In this research, the following spectral shape parameters are extracted:

- Audio Spectral Centroid (ASC) – describes the center of gravity of the log-frequency power spectrum;
- Audio Spectral Spread (ASSp) – shows the concentration of spectrum around the centroid;
- Audio Spectral Skewness (ASSk) – defines the spectral symmetry;
- Audio Spectral Kurtosis (ASK) – defines the flatness of spectrum;
- Spectral Entropy – gives a measure of spectrum irregularity (Wei *et al.*, 2018);
- Spectral RollOff – makes it possible to distinguish voiced and unvoiced speech;
- Spectral Brightness – gives a measure of sound timbre.

The parameter ASC is calculated as the first order central moment and is defined by the formula:

$$\mathrm{ASC} = \frac{\sum\limits_{i=1}^{M_{\mathrm{FT}}/2} \log_2\left(\frac{f(i)}{1000}\right)\mathrm{PS}(i)}{\sum\limits_{i=1}^{M_{\mathrm{FT}}/2} \mathrm{PS}(i)}, \tag{8}$$

where $f(i)$ is the frequency corresponding to bin i, $\mathrm{PS}(i)$ is the power spectrum given by Eq. (7), and M_{FT} – the number of the Fourier transform coefficients.

The parameter ASSp corresponds to the root square of the second order central moment of the spectrum, ASSk is the third order and ASK is the fourth order central moments. A more thorough description of these parameters as well as their formulas is given in (Korvel *et al.*, 2019).

Spectral Entropy can be expressed by the following formula:

$$\mathrm{Entropy} = -\frac{\sum\limits_{i=1}^{M_{\mathrm{FT}}/2} w_i \log_2 w_i}{\log_2 M_{\mathrm{FT}}/2}, \tag{9}$$

where

$$w_i = \frac{\mathrm{PS}(i)}{\sum\limits_{i=1}^{M_{\mathrm{FT}}/2} \mathrm{PS}(i)}. \tag{10}$$

The spectral RollOff is calculated as a frequency below which 85% of the magnitude distribution is concentrated. The formula of Spectral Brightness is given below:

$$\mathrm{Brightness} = \frac{\sum\limits_{i=f_c}^{M_{\mathrm{FT}}/2} \mathrm{PS}(i)}{\sum\limits_{i=1}^{M_{\mathrm{FT}}/2} \mathrm{PS}(i)}, \tag{11}$$

where f_c is cut-off frequency. This frequency was set to 1500 Hz in the experimental part of this research study.

In order to estimate the spectrum representation, Audio Spectrum Envelope (ASE) is calculated. For that, the frequency range is divided into sub-frames. The bands are logarithmically distributed, corresponding to a specific octave frequency (Kim *et al.*, 2005; Korvel *et al.*, 2019). ASE parameters are calculated by the following formula:

$$\mathrm{ASE}(k) = \begin{cases} \sum\limits_{i=0}^{P_1} \mathrm{PS}(i), & k = 1, \\ \sum\limits_{i=P_{k-1}}^{P_k} \mathrm{PS}(i), & 2 \leq k < K+1, \\ \sum\limits_{i=P_{k+1}}^{f_s/2} \mathrm{PS}(i), & k = K+2, \end{cases} \tag{12}$$

where $\mathrm{PS}(i)$ is the power spectral density of the segments of the phoneme, k is the frequency band number ($1 \leq k < K+1$). In this research, the frequency range is divided into 30 sub-frames, which consequently gives 29 AES parameters.

Due to the fact that formants play major role in most speech applications, the first four formants (F1–F4) are also included in our parameter set. Unlike the frequency parameters described above, formants are not based on the Fourier spectrum. They are calculated as the roots of the LPC polynomial.

The last group of analyzed parameters is Mel-Frequency Cepstral Coefficients (MFCCs). The MFCC feature extraction begins with calculating the power spectrum of the speech segment (see Eq. (7)). Then we triangle bandpass filters are constructed over the frequency range. The scale of the first 13 filters is linear;

for the rest of filters, the scale becomes logarithmic. The width of the linear filter is 66.67 Hz. The MFCCs are obtained by the following formula:

$$c_j = \sum_{i=0}^{L-1} m_i \cos\left(\frac{\pi j(i-1/2)}{L}\right), \tag{13}$$

where m_i are filterbank amplitudes, L – number of filters, $j = 0, ..., K$ (K – the number of cepstral coefficients).

We use 20 first coefficients of MFCC in this research.

Overall, we have 75 extracted parameters for each segment. In order to extract the parameters for the whole speech signal, statistical properties are computed based on these parameters obtained from all short-term segments. The used statistics are mean and variance.

Consequently, the resulting is the 150-dimensional feature vector.

4. Optimizing feature vector

Our goal is to determine acoustic speech parameters that let us distinguish interlanguage differences. For this purpose, the three-step algorithm is proposed:

Step 1. Phoneme parameter extraction.

Step 2. Rejection of parameters high-correlated with each other.

Step 3. Set the optimal number of features.

Step 4. Rejection of parameters which have the smallest differences of the averaged values between features of different languages.

The first step of the algorithm is parameterization of all audio samples. For this purpose, the features given in Sec. 2 are extracted. Then the features vectors are normalized. The normalization to the interval [0 1] is used. After parameter extraction, the rejection of high-correlated parameters is performed. For this purpose, the matrix of correlation coefficients is calculated. The parameters, for which correlation coefficients are larger than 0.75, are rejected. The rest of the parameters are used for the separability analysis in the interlanguage differences recognition process. For this purpose, the distances between the features of Lithuanian and Polish phonemes are calculated for all features separately. This process can be described by the following formula:

$$\begin{aligned}\mathrm{Dist(i)} &= \mathrm{Lithuanian\_feature(i)} \\ &\quad - \mathrm{Polish\_feature(i)}.\end{aligned} \tag{14}$$

In order to set the optimal number of parameters, the cross-validation check is performed. This process starts with creating a machine learning model based on one parameter. Then parameter one by one is added to the model. Parameters with highest distances (Eq. (14)) are used first. The model accuracy is calculated after adding each feature. This process is repeating until the accuracy starts to decrease.

For examining the extracted features, the two widely used classifiers, namely k-Nearest Neighbors (kNN) and Support Vector Machine (SVM) (Duda, 2000) are employed in this research.

5. Experiment results

The experiment consists of two parts. In the first part of the experiment, a comparative analysis of Lithuanian and Polish phonemes is performed. For the analysis, the consonant phonemes extracted from the recordings of Polish and Lithuanian speakers were used. These recordings consist of utterances of eight speakers (four females and four males) for each language. These utterances were recorded to the .wav file of the audio format with the following parameters: 48 kHz; 32 bit; mono. The recording scenario included only read sentences. These sentences have been segmented at phoneme units. The annotation was conducted manually using PRAAT program. The list of these phonemes used for the analysis is given in Fig. 3.

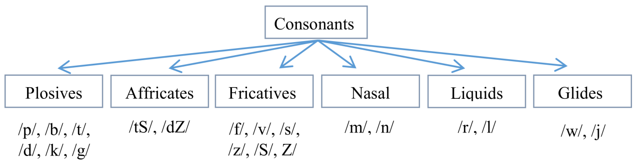

Fig. 3. Consonant phoneme used in the experiment.

The goal of this part of the experiment is to find vectors of acoustic parameters, that are related to differences between Polish and Lithuanian consonants. For that, we extracted parameters for all phonemes given in Fig. 3. Before the parameter extraction, the signal pre-processing is carried out. The frame length is equal to 512 samples; an overlap constitutes 50% of the segment length. To determine inter-language differences, we analyze the relation between particular vectors of parameters and speech phonemes. The analysis was performed for each phoneme separately. Evaluation of parameter suitability is based on calculation distances between parameters. As an example, the distances for parameters of phoneme /k/ arranged in descending numerical order are shown in Tables 3 and 4.

From the results shown in Tables 3 and 4, we see that some of the parameters are distinctly different. For example, the mean values of Audio Spectral Centroid (ASC) and Spectral Entropy have the biggest distances (see Table 3). An example of a graphical representation of separation based on these parameters is given in Fig. 4.

Table 3. Differences between parameters of Lithuanian and Polish phoneme /k/ based on the mean value (μ).

μ(ASC) 0.2829	μ(p2) 0.2829	μ(p3) 0.2829	μ(p4) 0.2829	μ(q1) 0.2829	μ(q2) 0.2829	μ(q3) 0.2829	μ(q4) 0.2829
μ(Entropy) 0.2766	μ(ASE3) 0.2438	μ(RollOff) 0.1886	μ(MFCC12) 0.1475	μ(ASSp) 0.1465	μ(k3) 0.1348	μ(ASE5) 0.125	μ(ASSk) 0.1215
μ(RMS) 0.0997	μ(ASE4) 0.0963	μ(Peak to RMS) 0.0934	μ(k1) 0.082	μ(MFCC13) 0.0802	μ(p1) 0.0798	μ(MFCC5) 0.0733	μ(Brightness) 0.0698
μ(MFCC11) 0.0696	μ(ASK) 0.0655	μ(MFCC16) 0.0596	μ(MFCC14) 0.059	μ(ASE10) 0.0568	μ(ASE11) 0.0552	μ(MFCC9) 0.0543	μ(TC) 0.0523
μ(ASE2) 0.0497	μ(MFCC18) 0.0469	μ(MFCC15) 0.0461	μ(MFCC17) 0.045	μ(MFCC10) 0.0446	μ(k2) 0.0414	μ(ASE27) 0.0407	μ(MFCC7) 0.0398
μ(ASE23) 0.0384	μ(ASE18) 0.0314	μ(MFCC1) 0.0289	μ(ASE22) 0.0285	μ(ASE21) 0.0263	μ(ASE24) 0.0262	μ(ASE12) 0.0239	μ(MFCC20) 0.0231
μ(ASE16) 0.0229	μ(ASE28) 0.0229	μ(ASE26) 0.0228	μ(MFCC8) 0.0199	μ(MFCC6) 0.0189	μ(F1) 0.0178	μ(ASE17) 0.0177	μ(ASE25) 0.0176
μ(MFCC4) 0.0175	μ(ZC) 0.0174	μ(ASE1) 0.0173	μ(ASE20) 0.0162	μ(F2) 0.0158	μ(ASE19) 0.0139	μ(MFCC19) 0.0118	μ(MFCC2) 0.0093
μ(MFCC3) 0.0079	μ(ASE14) 0.0059	μ(ASE29) 0.0051	μ(ASE13) 0.0045	μ(F4) 0.0029	μ(F3) 0.0003	μ(ASE6) 0.0002	μ(ASE7) 0.0001
μ(ASE8) 0.0001	μ(ASE9) 0.0001	μ(ASE15) 0					

Table 4. Differences between parameters of Lithuanian and Polish phoneme /k/ based on the variance (σ^2).

$\boldsymbol{\sigma}^2$(MFCC20) 0.2353	$\boldsymbol{\sigma}^2$(ASE3) 0.2305	$\boldsymbol{\sigma}^2$(MFCC18) 0.2148	$\boldsymbol{\sigma}^2$(MFCC19) 0.2038	$\boldsymbol{\sigma}^2$(MFCC15) 0.1723	$\boldsymbol{\sigma}^2$(MFCC14) 0.165	$\boldsymbol{\sigma}^2$(MFCC17) 0.1411	$\boldsymbol{\sigma}^2$(MFCC16) 0.1353
$\boldsymbol{\sigma}^2$(MFCC11) 0.1296	$\boldsymbol{\sigma}^2$(MFCC6) 0.1159	$\boldsymbol{\sigma}^2$(MFCC7) 0.1117	$\boldsymbol{\sigma}^2$(MFCC10) 0.1039	$\boldsymbol{\sigma}^2$(MFCC9) 0.0949	$\boldsymbol{\sigma}^2$(MFCC13) 0.0934	$\boldsymbol{\sigma}^2$(RollOff) 0.0822	$\boldsymbol{\sigma}^2$(Entropy) 0.082
$\boldsymbol{\sigma}^2$(ASC) 0.0772	$\boldsymbol{\sigma}^2$(ASE8) 0.0752	$\boldsymbol{\sigma}^2$(MFCC2) 0.0687	$\boldsymbol{\sigma}^2$(MFCC4) 0.0599	$\boldsymbol{\sigma}^2$(ASK) 0.0584	$\boldsymbol{\sigma}^2$(MFCC1) 0.0536	$\boldsymbol{\sigma}^2$(ASSp) 0.0533	$\boldsymbol{\sigma}^2$(ASE9) 0.0526
$\boldsymbol{\sigma}^2$(Peak to RMS) 0.0515	$\boldsymbol{\sigma}^2$(MFCC8) 0.0512	$\boldsymbol{\sigma}^2$(ASE10) 0.0492	$\boldsymbol{\sigma}^2$(q1) 0.0481	$\boldsymbol{\sigma}^2$(ASE6) 0.0358	$\boldsymbol{\sigma}^2$(Brightness) 0.0356	$\boldsymbol{\sigma}^2$(ASE11) 0.0326	$\boldsymbol{\sigma}^2$(MFCC5) 0.0308
$\boldsymbol{\sigma}^2$(ASE5) 0.0305	$\boldsymbol{\sigma}^2$(ASSk) 0.0254	$\boldsymbol{\sigma}^2$(ASE18) 0.0245	$\boldsymbol{\sigma}^2$(q2) 0.0209	$\boldsymbol{\sigma}^2$(ASE12) 0.0197	$\boldsymbol{\sigma}^2$(ASE25) 0.0188	$\boldsymbol{\sigma}^2$(ASE23) 0.0184	$\boldsymbol{\sigma}^2$(ASE27) 0.0178
$\boldsymbol{\sigma}^2$(ASE22) 0.0171	$\boldsymbol{\sigma}^2$(F2) 0.0152	$\boldsymbol{\sigma}^2$(ASE21) 0.0148	$\boldsymbol{\sigma}^2$(ASE26) 0.0137	$\boldsymbol{\sigma}^2$(ASE4) 0.0135	$\boldsymbol{\sigma}^2$(ASE20) 0.0133	$\boldsymbol{\sigma}^2$(ASE2) 0.0127	$\boldsymbol{\sigma}^2$(ASE24) 0.0127
$\boldsymbol{\sigma}^2$(ASE15) 0.0124	$\boldsymbol{\sigma}^2$(ASE14) 0.0116	$\boldsymbol{\sigma}^2$(ASE7) 0.0115	$\boldsymbol{\sigma}^2$(ASE1) 0.0113	$\boldsymbol{\sigma}^2$(ASE19) 0.0109	$\boldsymbol{\sigma}^2$(MFCC3) 0.0107	$\boldsymbol{\sigma}^2$(F1) 0.0099	$\boldsymbol{\sigma}^2$(ASE17) 0.0089
$\boldsymbol{\sigma}^2$(ASE16) 0.0088	$\boldsymbol{\sigma}^2$(ASE13) 0.0086	$\boldsymbol{\sigma}^2$(F3) 0.0071	$\boldsymbol{\sigma}^2$(F4) 0.0062	$\boldsymbol{\sigma}^2$(RMS) 0.0061	$\boldsymbol{\sigma}^2$(k1) 0.0059	$\boldsymbol{\sigma}^2$(ASE28) 0.0032	$\boldsymbol{\sigma}^2$(ZC) 0.0016
$\boldsymbol{\sigma}^2$(TC) 0.0009	$\boldsymbol{\sigma}^2$(q3) 0.0009	$\boldsymbol{\sigma}^2$(p1) 0.0008	$\boldsymbol{\sigma}^2$(p3) 0.0008	$\boldsymbol{\sigma}^2$(MFCC12) 0.0004	$\boldsymbol{\sigma}^2$(ASE29) 0.0003	$\boldsymbol{\sigma}^2$(k2) 0.0001	$\boldsymbol{\sigma}^2$(k3) 0
$\boldsymbol{\sigma}^2$(p2) 0	$\boldsymbol{\sigma}^2$(p4) 0	$\boldsymbol{\sigma}^2$(q4) 0					

The set of most suitable parameters in terms of phoneme separation is obtained by performing the cross-validation check. The machine learning model based on subsets of the initial feature set is tested. The model accuracy is the average accuracy of kNN and SVM methods. The obtained results are given in Table 5. As seen from Table 5 most of the listed parameters appear for the specific phoneme in various con-

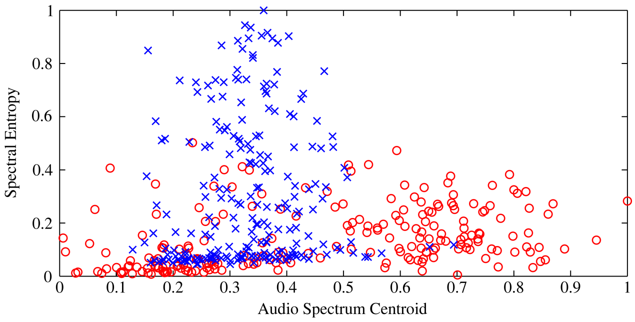

Fig. 4. Separation of Lithuanian and Polish phoneme /k/ (the circle denotes the Lithuanian phoneme; × – mark is used for the Polish phoneme).

figurations, but interestingly Audio Spectral Spread (ASSp), which shows the concentration of spectrum around the centroid is rarely visible. Parameters that occur in half or more phonemes are highlighted in bold font (see Table 5). Parameters, on the basis of which it is possible to separate all Lithuanian and Polish phonemes, are μ(MFCC5) and μ(MFCC2). Parameters that are useful for separation all phonemes except /l/ are σ^2(MFCC20), σ^2(MFCC10). The most common parameters also include μ(ASC), σ^2(MFCC11), μ(Entropy).

Table 5. The most suitable parameters for showing interlanguage differences.

/p/	**μ(Entropy)**, **μ(MFCC5)**, **μ(ASE3)**, **μ(ASC)**, **μ(MFCC4)**, σ^2(ASE3), **μ(q4)**, **μ(MFCC9)**, **μ(MFCC13)**, **μ(MFCC2)**, **σ^2(MFCC20)**, **μ(MFCC3)**, μ(MFCC16), **μ(ASSk)**, μ(MFCC10), μ(MFCC11), **σ^2(MFCC6)**, μ(ASE2) σ^2(RollOff), μ(ASE6), **σ^2(MFCC10)**, **μ(RollOff)**, μ(ASE4) μ(MFCC12), **μ(MFCC14)**, σ^2(k2)
/t/	**μ(Entropy)**, **μ(ASC)**, **μ(ASE3)**, **μ(MFCC5)**, **μ(MFCC4)**, **μ(MFCC9)**, **μ(RollOff)**, **μ(q4)**, μ(MFCC11), **μ(MFCC2)**, **μ(MFCC14)**, σ^2(k1), **μ(MFCC13)**, σ^2(ASE3), **σ^2(MFCC10)**, **μ(ASSk)**, μ(MFCC18), μ(ASE4) σ^2(Entropy) **μ(MFCC3)**, μ(MFCC17), **σ^2(MFCC18)**, μ(MFCC15), σ^2(ASC) σ^2(RollOff) μ(ASE5) μ(MFCC16), **σ^2(MFCC20)**, σ^2(Peak to RMS) μ(ASE2) μ(MFCC10), μ(MFCC12), **σ^2(MFCC12)**, σ^2(ASE2)
/d/	**μ(Entropy)**, **μ(MFCC5)**, **μ(ASE3)**, **μ(MFCC4)**, **μ(ASC)**, **μ(RollOff)**, **μ(MFCC9)**, **μ(MFCC2)**, **μ(q4)**, μ(MFCC11), **σ^2(MFCC10)**, σ^2(k1), **μ(MFCC14)**, **μ(MFCC13)**, **σ^2(MFCC20)**, **μ(MFCC3)**, σ^2(ASE3), σ^2(Entropy), σ^2(Peak to RMS), μ(ASE5), σ^2(ASC), **σ^2(MFCC12)**, **σ^2(MFCC18)**, σ^2(MFCC17), **μ(ASSk)**, μ(ASE4) μ(MFCC18), σ^2(RollOff), σ^2(MFCC8), μ(MFCC15), σ^2(MFCC9), μ(MFCC10), **σ^2(MFCC16)**, **σ^2(MFCC11)**, μ(ASE6)
/k/	**μ(ASC)**, **μ(MFCC5)**, **μ(Entropy)**, **μ(ASE3)**, **μ(RollOff)**, **μ(MFCC4)**, **σ^2(MFCC20)**, σ^2(ASE3), **μ(MFCC13)**, μ(MFCC16), **σ^2(MFCC10)**, **μ(q4)**, **μ(ASSk)**, μ(MFCC17), **μ(MFCC14)**, μ(MFCC18), **μ(MFCC9)**, **μ(MFCC2)**, **σ^2(MFCC18)**, **σ^2(MFCC14)**, μ(MFCC11), μ(MFCC15), μ(ASK), μ(MFCC10), σ^2(MFCC17), μ(MFCC12), **σ^2(MFCC16)**, μ(k3), **σ^2(MFCC11)**, μ(ASE5), **σ^2(MFCC15)**, μ(MFCC1), **σ^2(MFCC6)**, **σ^2(MFCC12)**
/g/	**μ(MFCC5)**, **μ(Entropy)**, **μ(ASE3)**, **μ(ASC)**, σ^2(ASE3), **μ(MFCC4)**, **σ^2(MFCC20)**, **σ^2(MFCC10)**, **μ(MFCC9)**, **μ(MFCC13)**, **μ(MFCC2)**, **σ^2(MFCC18)**, μ(MFCC16), **μ(q4)**
/tS/	**μ(ASC)**, **μ(MFCC5)**, σ^2(ASE3), **μ(MFCC4)**, **μ(MFCC9)**, **μ(ASE3)**, **μ(Entropy)**, **μ(q4)**, **σ^2(MFCC20)**, **μ(ASSk)**, **μ(MFCC13)**, **μ(MFCC2)**, **μ(RollOff)**, μ(ASK), **σ^2(MFCC10)**, μ(MFCC16), **μ(MFCC14)**, **σ^2(MFCC14)**, μ(MFCC17), μ(MFCC18), μ(MFCC15), μ(MFCC12), **σ^2(MFCC18)**, **σ^2(MFCC6)**, σ^2(MFCC17), μ(MFCC10), **σ^2(MFCC11)**
/f/	**μ(Entropy)**, **μ(MFCC5)**, **μ(ASE3)**, **μ(ASC)**, μ(ASE4) **μ(MFCC4)**, μ(ASE5), **σ^2(MFCC10)**, **μ(MFCC13)**, **σ^2(MFCC20)**, **μ(MFCC9)**, **μ(MFCC14)**, σ^2(ASE3), μ(ASE2) μ(MFCC10), μ(MFCC16), σ^2(MFCC17), μ(MFCC17), **σ^2(MFCC18)**, **μ(MFCC2)**, **μ(MFCC3)**, **σ^2(MFCC14)**, **σ^2(MFCC13)**, μ(MFCC11), σ^2(k2)
/v/	**μ(MFCC5)**, **μ(Entropy)**, **μ(ASE3)**, **σ^2(MFCC20)**, **σ^2(MFCC18)**, **σ^2(MFCC10)**, **μ(MFCC4)**, σ^2(ASE3), **μ(MFCC13)**, **σ^2(MFCC11)**, **μ(MFCC2)**, **σ^2(MFCC15)**, **μ(ASC)**, σ^2(MFCC17), **σ^2(MFCC13)**, **σ^2(MFCC16)**, **σ^2(MFCC14)**, σ^2(MFCC7)
/s/	μ(ZC), μ(p1), μ(p3) μ(q1), μ(q3), **μ(ASC)**, μ(ASK), **μ(MFCC9)**, **μ(MFCC2)**, **σ^2(MFCC10)**, **μ(RollOff)**, **σ^2(MFCC11)**, **μ(ASSk)**, **μ(MFCC4)**, μ(MFCC18), **σ^2(MFCC20)**, **μ(MFCC5)**, **σ^2(MFCC15)**, **σ^2(MFCC18)**, **σ^2(MFCC16)**, **μ(MFCC13)**, **μ(q4)**, μ(MFCC15), σ^2(RollOff), σ^2(MFCC9), μ(MFCC10), **σ^2(MFCC14)**, μ(MFCC12), **σ^2(MFCC6)**, **μ(MFCC14)**, **σ^2(MFCC13)**, **σ^2(MFCC12)**, σ^2(MFCC8)
/z/	μ(ZC), μ(p1), μ(p3), μ(q1), μ(q3), **μ(ASC)**, **μ(MFCC9)**, μ(ASK), **μ(MFCC2)**, **σ^2(MFCC10)**, **μ(RollOff)**, **μ(ASSk)**, **σ^2(MFCC11)**, **σ^2(MFCC15)**, **σ^2(MFCC20)**, **μ(MFCC4)**, μ(MFCC18), **μ(MFCC5)**, **μ(q4)**, **μ(MFCC13)**, **σ^2(MFCC6)**, **σ^2(MFCC16)**, **σ^2(MFCC18)**, μ(MFCC15), **σ^2(MFCC14)**, σ^2(MFCC9), μ(MFCC10), **σ^2(MFCC13)**, σ^2(MFCC8), μ(MFCC12), **μ(MFCC14)**, σ^2(RollOff), **σ^2(MFCC12)**
/S/	**μ(ASC)**, **σ^2(MFCC15)**, **μ(q4)**, **μ(MFCC5)**, μ(TC), **σ^2(MFCC10)**, μ(p1), μ(p3), μ(q1), μ(q3), **μ(MFCC2)**, μ(MFCC15), **μ(MFCC13)**, **μ(MFCC9)**, **σ^2(MFCC11)**, μ(MFCC18), **σ^2(MFCC16)**, **μ(MFCC4)**, **σ^2(MFCC13)**, **σ^2(MFCC14)**, μ(MFCC11), **σ^2(MFCC20)**, **σ^2(MFCC12)**, μ(ASK), **μ(ASSk)**, σ^2(MFCC9), σ^2(MFCC8), **σ^2(MFCC6)**, **σ^2(MFCC18)**, σ^2(ASC), σ^2(MFCC5), μ(MFCC1), μ(Peak to RMS)

Table 5. [Cont.]

/Z/	**μ(ASC)**, **σ^2(MFCC15)**, **μ(q4)**, **σ^2(MFCC10)**, **μ(MFCC5)**, **μ(MFCC9)**, μ(TC), **μ(MFCC2)**, μ(p1) μ(p3) μ(q1), μ(q3), **σ^2(MFCC11)**, **μ(MFCC4)**, **μ(MFCC13)**, μ(MFCC15), μ(MFCC18), μ(MFCC11), **σ^2(MFCC16)**, **σ^2(MFCC20)**, **σ^2(MFCC13)**, **σ^2(MFCC14)**, **σ^2(MFCC6)**, **σ^2(MFCC12)**, **μ(ASSk)**, σ^2(MFCC9), σ^2(MFCC8), μ(MFCC1), μ(ASK)
/m/	**μ(Entropy)**, **μ(MFCC5)**, **σ^2(MFCC6)**, **μ(RollOff)**, **σ^2(MFCC10)**, μ(MFCC1), **μ(MFCC2)**, **μ(MFCC4)**, **μ(ASE3)**, σ^2(MFCC8), **σ^2(MFCC11)**, **σ^2(MFCC12)**, μ(ASE5) **μ(MFCC3)**, μ(Brightness), **σ^2(MFCC14)**, σ^2(MFCC7), **σ^2(MFCC15)**, μ(ASE8), **σ^2(MFCC18)**, μ(ASE7), σ^2(MFCC9), **σ^2(MFCC20)**, μ(ASE8) **μ(MFCC9)**, **μ(MFCC14)**, **σ^2(MFCC13)**
/n/	**μ(MFCC5)**, **μ(ASC)**, **μ(MFCC3)**, **μ(Entropy)**, **σ^2(MFCC12)**, μ(ASK), **μ(MFCC2)**, **σ^2(MFCC15)**, μ(ASE7), **σ^2(MFCC13)**, μ(F4) μ(Brightness), **σ^2(MFCC10)**, **σ^2(MFCC11)**, μ(ASE5), σ^2(ASE7), σ^2(MFCC9), μ(MFCC1)
/r/	μ(k3), μ(Peak to RMS), σ^2(ZC) μ(Brightness), **μ(Entropy)**, **μ(MFCC5)**, μ(ASE1) **σ^2(MFCC6)**, **μ(RollOff)**, σ^2(MFCC17), **μ(MFCC4)**, σ^2(k2), σ^2(MFCC9), **μ(MFCC2)**, μ(ASE7) μ(k1), **σ^2(MFCC15)**, **σ^2(MFCC12)**, **μ(MFCC3)**, **σ^2(MFCC10)**, **σ^2(MFCC13)**, μ(MFCC20), μ(ASE2), σ^2(MFCC4), σ^2(MFCC2), **μ(ASE3)**, **σ^2(MFCC14)**, **σ^2(MFCC11)**, σ^2(ASSK), μ(ASE8), **σ^2(MFCC20)**, **σ^2(MFCC16)**, μ(p2)
/l/	μ(ASE1) μ(Brightness), **μ(MFCC5)**, **μ(Entropy)**, μ(ASE7) **μ(RollOff)**, σ^2(MFCC7), μ(Peak to RMS), μ(F4) **μ(ASC)**, **σ^2(MFCC6)**, **μ(MFCC2)**, **σ^2(MFCC13)**, **σ^2(MFCC20)**
/j/	μ(Brightness), μ(ASE1) **μ(MFCC5)**, μ(ASE7) **μ(Entropy)**, μ(Peak to RMS), μ(F4) **μ(RollOff)**, **σ^2(MFCC15)**, σ^2(MFCC7), **σ^2(MFCC6)**, μ(ASE14), **σ^2(MFCC13)**, **σ^2(MFCC12)**, **σ^2(MFCC10)**, σ^2(k2), **μ(MFCC2)**, μ(k3), μ(ASE8), **σ^2(MFCC14)**, μ(ASE13), **σ^2(MFCC20)**, σ^2(ASE1), **σ^2(MFCC16)**, **σ^2(MFCC11)**, μ(ASE2) **μ(ASC)**, **μ(MFCC4)**, σ^2(MFCC4), σ^2(MFCC8), σ^2(ZC), μ(MFCC20), **μ(MFCC3)**, μ(ASE5)

In the second part of the experiment, we test the effectiveness of the selected features in the context of automatic phoneme recognition. In addition, the English language, as an auxiliary language, is used. The recordings of Lithuanian and Polish speakers used are the same as in the first part of the experiment. For the English language, the well-known TIMIT Acoustic-Phonetic Continuous Speech Corpus is used (Garofolo *et al.*, 1993). This corpus contains recordings of 630 speakers of 8 major dialects of American English. In our research study recordings of a dialect named New York City were used. Recordings from 16 speakers (eight females and eight males) were used.

We extracted parameters for all the phonemes. The classification based on the most suitable parameters (given in Table 5) was performed. The feature set of each phoneme is divided into two parts: features for which the class labels are known (training dataset) and features for which class labels need to be determined (testing dataset). For the class determination SVM and kNN classifiers are used. The classifiers were used without parameter tuning. The obtained results are compared with the correct class labels of the data. In order to evaluate the classifier performance, the confusion matrix CM is calculated. Based on this matrix overall accuracy and three class-specific measures, i.e., class recall, class precision and $F1$-*measure*, are determined. The obtained results (averaged for all speakers, males and females separately) are given in Table 6, where A refers to the samples of Lithuanian and

Table 6. The results of classification based on the optimized feature vector (given in Table 5).

Samples		kNN				SVM			
		Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
Phoneme /p/ (54)									
A	All	0.978	0.985	0.970	0.977	0.948	0.941	0.955	0.948
	Female	0.975	0.952	1.000	0.976	0.975	1.000	0.950	0.974
	Male	0.946	1.000	0.893	0.943	1.000	1.000	1.000	1.000
B	All	0.925	0.890	0.970	0.929	0.993	0.985	1.000	0.993
	Female	0.863	0.809	0.950	0.874	0.988	0.976	1.000	0.988
	Male	0.893	0.824	1.000	0.903	0.661	0.596	1.000	0.747
C	All	0.985	0.971	1.000	0.985	1.000	1.000	1.000	1.000
	Female	0.963	0.951	0.975	0.963	0.975	0.975	0.975	0.975
	Male	0.946	0.903	1.000	0.949	0.893	0.824	1.000	0.903

Table 6. [Cont.]

Samples		kNN				SVM			
		Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
Phoneme /t/ (92)									
A	All	0.963	0.970	0.955	0.962	0.881	0.823	0.970	0.890
	Female	0.938	0.973	0.900	0.935	0.913	0.971	0.850	0.907
	Male	0.964	1.000	0.929	0.963	0.964	1.000	0.929	0.963
B	All	0.896	0.884	0.910	0.897	0.866	0.845	0.896	0.870
	Female	0.838	0.829	0.850	0.840	0.888	0.970	0.800	0.877
	Male	0.875	0.800	1.000	0.889	0.946	0.903	1.000	0.949
C	All	0.955	0.918	1.000	0.957	0.970	0.944	1.000	0.971
	Female	0.950	0.929	0.975	0.951	0.963	0.930	1.000	0.964
	Male	0.946	0.903	1.000	0.949	0.964	0.933	1.000	0.966
Phoneme /d/ (21)									
A	All	0.955	0.942	0.970	0.956	0.910	0.867	0.970	0.916
	Female	0.975	0.975	0.975	0.975	0.938	1.000	0.875	0.933
	Male	1.000	1.000	1.000	1.000	0.929	0.962	0.893	0.926
B	All	0.813	0.763	0.910	0.830	0.858	0.843	0.881	0.861
	Female	0.675	0.675	0.675	0.675	0.875	0.895	0.850	0.872
	Male	0.768	0.703	0.929	0.800	0.786	0.735	0.893	0.807
C	All	0.970	0.957	0.985	0.971	0.985	0.971	1.000	0.985
	Female	0.950	0.950	0.950	0.950	0.963	0.930	1.000	0.964
	Male	0.946	0.903	1.000	0.949	0.964	0.933	1.000	0.966
Phoneme /k/ (122)									
A	All	0.955	0.930	0.985	0.957	0.925	0.880	0.985	0.930
	Female	0.750	0.667	1.000	0.800	0.738	0.732	0.750	0.741
	Male	0.929	0.962	0.893	0.926	0.946	1.000	0.893	0.943
B	All	0.761	0.706	0.896	0.790	0.806	0.781	0.851	0.814
	Female	0.675	0.652	0.750	0.698	0.713	0.774	0.600	0.676
	Male	0.768	0.703	0.929	0.800	0.821	0.750	0.964	0.844
C	All	0.963	0.956	0.970	0.963	0.993	0.985	1.000	0.993
	Female	0.825	0.964	0.675	0.794	0.838	0.846	0.825	0.835
	Male	0.911	0.849	1.000	0.918	0.929	0.962	0.893	0.926
Phoneme /g/ (28)									
A	All	0.955	0.930	0.985	0.957	0.888	0.833	0.970	0.897
	Female	0.950	0.909	1.000	0.952	0.850	0.967	0.725	0.829
	Male	0.946	1.000	0.893	0.943	0.893	0.923	0.857	0.889
B	All	0.769	0.700	0.940	0.803	0.881	0.870	0.896	0.882
	Female	0.738	0.686	0.875	0.769	0.900	0.900	0.900	0.900
	Male	0.768	0.683	1.000	0.812	0.982	0.966	1.000	0.983
C	All	0.955	0.930	0.985	0.957	0.970	0.957	0.985	0.971
	Female	0.975	0.975	0.975	0.975	0.963	0.951	0.975	0.963
	Male	0.893	0.824	1.000	0.903	0.946	0.903	1.000	0.949
Phoneme /tS/ (25)									
A	All	0.940	0.904	0.985	0.943	0.866	0.803	0.970	0.878
	Female	0.838	0.755	1.000	0.860	0.725	0.846	0.550	0.667
	Male	0.893	0.958	0.821	0.885	0.911	1.000	0.821	0.902
B	All	0.761	0.706	0.896	0.790	0.791	0.747	0.881	0.808
	Female	0.613	0.592	0.725	0.652	0.738	0.788	0.650	0.712
	Male	0.679	0.609	1.000	0.757	0.750	0.667	1.000	0.800
C	All	0.948	0.941	0.955	0.948	0.985	0.971	1.000	0.985
	Female	0.888	0.897	0.875	0.886	0.875	0.826	0.950	0.884
	Male	0.768	0.683	1.000	0.812	0.893	0.893	0.893	0.893

Table 6. [Cont.]

Samples		kNN				SVM			
		Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
Phoneme /f/ (52)									
A	All	0.896	0.863	0.940	0.900	0.918	0.889	0.955	0.921
	Female	0.863	0.796	0.975	0.876	0.888	0.861	0.925	0.892
	Male	0.893	1.000	0.786	0.880	0.786	0.808	0.750	0.778
B	All	0.784	0.726	0.910	0.808	0.851	0.805	0.925	0.861
	Female	0.825	0.771	0.925	0.841	0.750	0.738	0.775	0.756
	Male	0.821	0.781	0.893	0.833	0.679	0.625	0.893	0.735
C	All	0.940	0.893	1.000	0.944	0.948	0.929	0.970	0.949
	Female	0.825	0.861	0.775	0.816	0.875	0.895	0.850	0.872
	Male	0.786	0.700	1.000	0.824	0.625	0.578	0.929	0.712
Phoneme /v/ (49)									
A	All	0.940	0.928	0.955	0.941	0.918	0.924	0.910	0.917
	Female	0.988	1.000	0.975	0.987	0.950	0.974	0.925	0.949
	Male	0.911	1.000	0.821	0.902	0.946	0.963	0.929	0.946
B	All	0.776	0.718	0.910	0.803	0.813	0.792	0.851	0.820
	Female	0.750	0.794	0.675	0.730	0.800	0.875	0.700	0.778
	Male	0.875	0.862	0.893	0.877	0.875	0.862	0.893	0.877
C	All	0.940	0.904	0.985	0.943	0.963	0.931	1.000	0.964
	Female	0.925	0.905	0.950	0.927	0.950	0.909	1.000	0.952
	Male	0.875	0.800	1.000	0.889	0.911	0.849	1.000	0.918
Phoneme /s/ (86)									
A	All	0.970	0.970	0.970	0.970	0.925	0.952	0.896	0.923
	Female	0.750	0.672	0.975	0.796	0.700	0.667	0.800	0.727
	Male	0.875	0.957	0.786	0.863	0.750	0.733	0.786	0.759
B	All	0.910	0.887	0.940	0.913	0.873	0.891	0.851	0.870
	Female	0.738	0.732	0.750	0.741	0.838	0.909	0.750	0.822
	Male	0.696	1.000	0.393	0.564	0.893	0.923	0.857	0.889
C	All	0.985	0.971	1.000	0.985	0.963	0.931	1.000	0.964
	Female	0.850	1.000	0.700	0.824	0.888	1.000	0.775	0.873
	Male	0.875	1.000	0.750	0.857	0.875	1.000	0.750	0.857
Phoneme /z/ (17)									
A	All	0.970	0.985	0.955	0.970	0.918	0.878	0.970	0.922
	Female	0.775	0.704	0.950	0.809	0.763	0.733	0.825	0.777
	Male	0.911	1.000	0.821	0.902	0.696	0.790	0.536	0.638
B	All	0.896	0.853	0.955	0.901	0.851	0.822	0.896	0.857
	Female	0.688	0.683	0.700	0.691	0.713	0.718	0.700	0.709
	Male	0.679	1.000	0.357	0.526	0.679	0.708	0.607	0.654
C	All	0.963	0.931	1.000	0.964	0.955	0.930	0.985	0.957
	Female	0.850	0.889	0.800	0.842	0.900	1.000	0.800	0.889
	Male	0.839	1.000	0.679	0.809	0.750	0.889	0.571	0.696
Phoneme /S/ (83)									
A	All	0.978	0.985	0.970	0.977	0.985	0.985	0.985	0.985
	Female	0.750	0.672	0.975	0.796	0.900	0.864	0.950	0.905
	Male	0.893	1.000	0.786	0.880	0.804	0.743	0.929	0.825
B	All	0.896	0.863	0.940	0.900	0.866	0.866	0.866	0.866
	Female	0.713	0.667	0.850	0.747	0.700	0.750	0.600	0.667
	Male	0.804	1.000	0.607	0.756	0.607	1.000	0.214	0.353
C	All	0.955	0.918	1.000	0.957	0.970	0.944	1.000	0.971
	Female	0.913	1.000	0.825	0.904	0.963	0.974	0.950	0.962
	Male	0.964	1.000	0.929	0.963	0.857	0.857	0.857	0.857

Table 6. [Cont.]

Samples		kNN Accuracy	kNN Precision	kNN Recall	kNN F1	SVM Accuracy	SVM Precision	SVM Recall	SVM F1
Phoneme /Z/ (17)									
S	All	0.978	0.985	0.970	0.977	0.970	0.985	0.955	0.970
	Female	0.763	0.691	0.950	0.800	0.913	0.884	0.950	0.916
	Male	0.893	1.000	0.786	0.880	0.786	0.722	0.929	0.813
B	All	0.903	0.875	0.940	0.907	0.866	0.866	0.866	0.866
	Female	0.713	0.667	0.850	0.747	0.763	0.889	0.600	0.716
	Male	0.804	1.000	0.607	0.756	0.589	1.000	0.179	0.303
C	All	0.978	0.957	1.000	0.978	0.978	0.957	1.000	0.978
	Female	0.900	1.000	0.800	0.889	0.975	1.000	0.950	0.974
	Male	0.964	1.000	0.929	0.963	0.893	0.893	0.893	0.893
Phoneme /m/ (78)									
A	All	0.978	0.957	1.000	0.978	0.903	0.846	0.985	0.910
	Female	0.975	0.952	1.000	0.976	0.988	1.000	0.975	0.987
	Male	0.946	1.000	0.893	0.943	0.964	1.000	0.929	0.963
B	All	0.866	0.877	0.851	0.864	0.851	0.873	0.821	0.846
	Female	0.838	0.846	0.825	0.835	0.863	0.968	0.750	0.845
	Male	0.839	0.788	0.929	0.853	0.911	0.871	0.964	0.915
C	All	0.985	0.971	1.000	0.985	0.955	0.984	0.925	0.954
	Female	0.963	0.974	0.950	0.962	0.925	0.870	1.000	0.930
	Male	0.929	0.875	1.000	0.933	0.964	0.933	1.000	0.966
Phoneme /n/ (223)									
A	All	0.978	0.957	1.000	0.978	1.000	1.000	1.000	1.000
	Female	0.950	0.909	1.000	0.952	0.975	0.952	1.000	0.976
	Male	0.982	0.966	1.000	0.983	1.000	1.000	1.000	1.000
B	All	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Female	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Male	0.982	1.000	0.964	0.982	1.000	1.000	1.000	1.000
C	All	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Female	0.988	1.000	0.975	0.987	1.000	1.000	1.000	1.000
	Male	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Phoneme /r/ (203)									
A	All	0.970	0.985	0.955	0.970	0.978	0.971	0.985	0.978
	Female	0.938	0.927	0.950	0.938	0.975	1.000	0.950	0.974
	Male	0.982	1.000	0.964	0.982	1.000	1.000	1.000	1.000
B	All	0.910	0.887	0.940	0.913	0.881	0.849	0.925	0.886
	Female	0.825	0.771	0.925	0.841	0.788	0.926	0.625	0.746
	Male	1.000	1.000	1.000	1.000	0.982	1.000	0.964	0.982
C	All	0.970	0.944	1.000	0.971	0.985	0.971	1.000	0.985
	Female	0.950	0.909	1.000	0.952	0.950	0.909	1.000	0.952
	Male	0.982	0.966	1.000	0.983	1.000	1.000	1.000	1.000
Phoneme /l/ (143)									
A	All	0.970	0.957	0.985	0.971	0.948	0.917	0.985	0.950
	Female	0.929	0.900	0.964	0.931	0.982	0.966	1.000	0.983
	Male	0.881	0.881	0.881	0.881	0.993	0.985	1.000	0.993
B	All	0.850	0.769	1.000	0.870	0.950	0.909	1.000	0.952
	Female	0.661	0.596	1.000	0.747	0.536	0.519	1.000	0.683
	Male	0.970	0.944	1.000	0.971	1.000	1.000	1.000	1.000
C	All	0.950	0.909	1.000	0.952	0.988	0.976	1.000	0.988
	Female	0.911	0.871	0.964	0.915	0.946	0.903	1.000	0.949
	Male	0.982	0.966	1.000	0.983	1.000	1.000	1.000	1.000

Table 6. [Cont.]

Samples		kNN				SVM			
		Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
Phoneme /j/ (48)									
A	All	0.963	0.984	0.940	0.962	0.970	0.957	0.985	0.971
	Female	0.988	0.976	1.000	0.988	0.975	1.000	0.950	0.974
	Male	0.946	1.000	0.893	0.943	0.982	0.966	1.000	0.983
B	All	0.881	0.881	0.881	0.881	0.963	0.943	0.985	0.964
	Female	0.875	0.826	0.950	0.884	0.950	0.950	0.950	0.950
	Male	0.821	0.737	1.000	0.849	0.661	0.596	1.000	0.747
C	All	0.978	0.957	1.000	0.978	1.000	1.000	1.000	1.000
	Female	0.950	0.909	1.000	0.952	0.975	0.952	1.000	0.976
	Male	0.964	0.933	1.000	0.966	0.929	0.875	1.000	0.933

Polish languages, while B – samples of Lithuanian and English languages, and C – samples of Polish and English languages. The numbers in brackets (see Table 6) show how many phoneme samples were used in the experiment for each language.

The are several conclusions that may be derived from the results obtained. First of all, the accuracies obtained for "*All*" are very high regardless of the language. In most cases, kNN returns higher accuracy than the SVM classifier, but differences are not always statistically significant. It should be remembered that $F1$-measure may be more useful in this analysis as there is an uneven class distribution, but we can see that overall, it also gets very high values. When looking at the pairs of languages, it may be observed that in all but for the phoneme /*n*/ statistical measures obtained for the B case (Lithuanian-English phonemes) are lower than for A and C cases. This may be caused by the fact that feature vectors were parametrized for Lithuanian and Polish and not for the English language, but C case disproves such a hypothesis. Contrarily, we cannot say that Polish and English phonemes are better separated than Lithuanian-English, for such a conclusion bigger corpora should be utilized. Moreover, when analyzing Table 4, we cannot say that higher values of measures are more often obtained in the case of female- or male- pronounced phonemes regardless of the language used.

6. Conclusions

A comparison analysis based on acoustic parameters between Lithuanian and Polish language consonants has been performed. A set of acoustic parameters, optimized by the separability analysis, related to differences between Polish and Lithuanian language consonants has been obtained for each consonant. In order to evaluate the classification accuracy, two methods, namely kNN and SVM, were used. The analyses were performed for the whole group of speakers, and male and female speakers separately. High classification accuracies show that the proposed and optimized parameters are useful in the process of determination of inter-language differences.

An interesting observation may be made when comparing the pairs of languages: Lithuanian-Polish, Lithuanian-English, and Polish-English, namely, it is clearly seen that Lithuanian-English phonemes are more difficult to separate. In the future experiments, a bigger corpus will be used to observe whether this trend remained true. Moreover, a larger set of acoustics features will be chosen and optimized for these three languages, as well as other machine learning algorithms will be employed.

Finally, we worked on the optimization of the feature vector to utilize it in the multidimensional quality assessment of the synthesized phonemes.

Acknowledgment

This research is funded by the European Social Fund under the No 09.3.3-LMT-K-712 "Development of Competences of Scientists, other Researchers and Students through Practical Research Activities" measure.

References

1. Badshah A.M. *et al.* (2019), *Deep features-based speech emotion recognition for smart affective services*, Multimedia Tools and Applications, **78**, 5, 5571–5589, doi: 10.1007/s11042-017-5292-7.
2. Bourlard H. (2018), *Evolution of Neural Network Architectures for speech recognition*, Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2–6 September 2018, p. 1767.
3. Chia Ai, Hariharan M., Yaacob S., Sin L. Chee (2012), *Classification of speech dysfluencies with MFCC and LPCC features*, Expert Systems with Ap-

plications, **39**, 2, 2157–2165, doi: 10.1016/j.eswa.2011.07.065.

4. Czyżewski A., Piotrowska M., Kostek B. (2017), *Analysis of allophones based on audio signal recordings and parameterization*, Journal of the Acoustical Society of America, **141**, 5, 3521–3521, doi: 10.1121/1.4987415.

5. Decker D.M. (1999), *Handbook of the international phonetic association: a guide to the use of the international phonetic alphabet*, Cambridge University Press.

6. Demenko G., Wypych M., Baranowska E. (2003), *Implementation of grapheme-to-phoneme rules and extended SAMPA alphabet in Polish text-to-speech synthesis*, Speech and Language Technology, **7**, 17, 79–97.

7. Deng L., Seltzer M.L., Yu D., Acero A., Mohamed A.-R., Hinton G.E. (2010), *Binary coding of speech spectrograms using a deep auto-encoder*, Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010, pp. 1692–1695.

8. Duda R.O., Hart P.E., Stork D.G. (2000), *Pattern classification*, 2nd ed., New York: Wiley.

9. Eringis D., Tamulevicius G. (2015), *Modified filterbank analysis features for speech recognition*, Baltic Journal of Modern Computing, **3**, 1, 29–42, https://www.bjmc.lu.lv/fileadmin/user_upload/lu_portal/projekti/bjmc/Contents/3_1_3_Eringis.pdf.

10. Gales M.J.F., Knill K.M., Ragni A. (2015), *Unicode-based graphemic systems for limited resource languages*, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 5186–5190, doi: 10.1109/ICASSP.2015.7178960.

11. Gibbon D., Moore R., Winski R. (1997), *Handbook of standards and resources for spoken language systems*, Berlin; New York: Mouton de Gruyter.

12. Girdenis A.S. (2003), *Theoretical bases of Lithuanian phonology* [in Lithuanian: *Teoriniai lietuvių fonologijos pagrindai*], Vilnius: Mokslo ir enciklopedijų leidybos institutas.

13. Greibus M., Ringelienė Ž., Telksnys L. (2017), *The phoneme set influence for Lithuanian speech commands recognition accuracy*, Open Conference of Electrical, Electronic and Information Sciences (eStream), 27–27 April 2017, Vilnius, Lithuania, pp. 82–85, doi: 10.1109/eStream.2017.7950321.

14. Gut U. (2014), *Introduction to English phonetics and phonology volume*, Bern: Peter Lang.

15. Gussmann E. (2007), *The Phonology of Polish*, New York: Oxford University Press.

16. Howard D.M., Murphy D.T. (2007), *Voice science, acoustics, and recording*, San Diego, CA: Plural Publishing.

17. Garofolo J.S., Lamel L.F., Fisher W.M., Fiscus J.G., Pallett D.S., Dahlgren N.L. (1993), *TIMIT acoustic-phonetic continuous speech corpus*, LDC93S1. Web Download. Philadelphia: Linguistic Data Consortium.

18. Igras M., Ziółko B., Jadczyk T. (2013), *Audiovisual database of Polish speech recordings*, Studia Informatica, **33**, 2B, 163–172, doi: 10.21936/si2012_v33.n2B.182.

19. Izydorczyk J., Kłosowski P. (2001), *Base acoustic properties of Polish speech*, International Conference Programable Devices and Systems PDS2001 IFAC Workshop, Gliwice, November 22–23, pp. 61–66.

20. Jassem W. (2003), *Polish*, Journal of the International Phonetic Association, **33**, 1, 103–107, doi: 10.1017/S0025100303001191.

21. Kasparaitis P. (2005), *Diphone databases for Lithuanian text-to-speech synthesis*, Informatica, **2**, 16, 193–202.

22. Kasparaitis P. (2008), *Lithuanian speech recognition using the English recognizer*, Informatica, **19**, 4, 505–516.

23. Kim H.-G., Moreau N., Sikora T. (2005), *MPEG-7 audio and beyond: audio content indexing and retrieval*, New York: Wiley & Sons.

24. Kłosowski P., Dustor A., Izydorczyk J., Kotas J., Slimok J. (2014), *Speech recognition based on open source speech processing software*, [In:] *Computer Networks*, CN. Vol. 431 of Communications in Computer and Information Science, ed. by A. Kwiecień, P. Gaj, and P. Stera, 21st International Science Conference on Computer Networks (CN), Poland, June 23–27 (Springer-Verlag Berlin, 2014), pp. 308–317.

25. Kłosowski P. (2017), *Statistical analysis of orthographic and phonemic language corpus for word-based and phoneme-based Polish language modelling*, EURASIP Journal on Audio, Speech, and Music Processing, **2017**, 5, doi: 10.1186/s13636-017-0102-8.

26. Korvel G., Kostek B. (2017a), *Examining feature vector for phoneme recognition*, 2017 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), Bilbao, 2017, pp. 394–398, doi: 10.1109/ISSPIT.2017.8388675.

27. Korvel G., Kostek B. (2017b), *Voiceless Stop Consonant Modelling and Synthesis Framework Based on MISO Dynamic System*, Archives of Acoustics, **42**, 3, 375–383, doi: 10.1515/aoa-2017-0039.

28. Korvel G., Kurowski A., Kostek B., Czyzewski A. (2019), *Speech analytics based on machine learning*, [in:] Tsihrintzis G., Sotiropoulos D., Jain L. [Eds], *Machine Learning Paradigms. Intelligent Systems Reference Library*, Vol. 149, pp. 129–157, Springer: Cham, doi: 10.1007/978-3-319-94030-4.

29. Korvel G., Treigys P., Tamulevičius G., Bernatavičienė J., Kostek B. (2018), *Analysis of 2d feature spaces for deep learning-based speech recognition*, Journal of the Audio Engineering Society, **66**, 12, 1072–1081, doi: 10.17743/jaes.2018.0066.

30. Kostek B. *et al.* (2011), *Report of the ISMIS 2011 Contest: Music Information Retrieval*, [in:] Kryszkiewicz M., Rybinski H., Skowron A., Raś Z.W. [Eds], *Foundations of Intelligent Systems. ISMIS 2011.*

Lecture Notes in Computer Science, Vol. 6804, pp. 715–724, Springer: Berlin, Heidelberg, doi: 10.1007/978-3-642-21916-0_75.

31. Kostek B., Piotrowska M., Czyżewski A. (2017), *Comparative study of self-organizing maps vs. subjective evaluation of quality of allophone pronunciation for nonnative English speakers*, 143rd Audio Engineering Society Convention, preprint 9847, New York.

32. Kozierski P., Sadalla T., Drgas S., Dąbrowski A. (2016), *Allophones in automatic whispery speech recognition*, 2016 21st International Conference on Methods and Models in Automation and Robotics (MMAR), Miedzyzdroje, 2016, pp. 811–815, doi: 10.1109/MMAR.2016.7575241.

33. Labarre T. (2011), *LING550: CLMS project on Polish*, https://www.academia.edu/5332895/LING550_CLMS_Project_on_Polish.

34. Laurinciukaite S., Telksnys L., Kasparaitis P., Kliukiene R., Paukstyte V. (2018), *Lithuanian Speech Corpus Liepa for development of human-computer interfaces working in voice recognition and synthesis mode*, Informatica, **29**, 3, 487–498, doi: 10.15388/informatica.2018.177.

35. Lileikytė R., Gorin A., Lamel L., Gauvain J., Fraga-Silva T. (2016), *Lithuanian broadcast speech transcription using semi-supervised acoustic model training*, Procedia Computer Science, **81**, 107–113, doi: 10.1016/j.procs.2016.04.037.

36. Mitterer H., Reinisch E., Mcqueen J.M. (2018), *Allophones, not phonemes in spoken-word recognition*, Journal of Memory and Language, **98**, 77–92, doi: 10.1016/j.jml.2017.09.005.

37. Noroozi F., Kamińska D., Sapinski T., Anbarjafari G. (2017), *Supervised Vocal-Based Emotion Recognition Using Multiclass Support Vector Machine, Random Forests, and AdaBoost*, Journal of the Audio Engineering Society, **65**, 7/8, 562–572, doi: 10.17743/jaes.2017.0022.

38. Oliver D., Szklanny K. (2006), *Creation and analysis of a Polish speech database for use in unit selection synthesis*, http://syntezamowy.pjwstk.edu.pl/publikacje/lrec2006.pdf (accessed Jan. 2019).

39. Padmanabhan J., Premkumar M.J.J. (2015), *Machine Learning in Automatic Speech Recognition: A Survey*. IETE Technical Review, **32**, 1–12, doi: 10.1080/02564602.2015.1010611.

40. Przepiórkowski A., Bańko M., Górski R.L., Lewandowska-Tomaszczyk B. (2012), *The National Corpus of Polish* [in Polish: *Narodowy korpus języka polskiego*], Wydawnictwo Naukowe PWN, Warszawa.

41. Raškinis A., Raškinis G., Kazlauskienė A. (2003), *SAMPA (speech assessment methods phonetic alphabet) for encoding transcriptions of Lithuanian speech corpora*, Information Technology and Control, **29**, 4, 50–56, https://hdl.handle.net/20.500.12259/55530.

42. Recasens D. (2012), *A cross-language acoustic study of initial and final allophones of /l/*, Speech Communication, **54**, 3, 368–383, doi: 10.1016/j.specom.2011.10.001.

43. Rudzionis V., Maskeliunas R., Rudzionis A., Ratkevicius K. (2009), *On the adaptation of foreign language speech recognition engines for Lithuanian speech recognition*, [in:] Abramowicz W., Flejter D. [Eds], *Business Information Systems Workshops. BIS 2009. Lecture Notes in Business Information Processing*, Vol. 37, pp. 113–118, Springer, Berlin, Heidelberg, doi: 10.1007/978-3-642-03424-4_13.

44. SAMPA En, https://www.phon.ucl.ac.uk/home/sampa/english.htm.

45. SAMPA Pl, https://www.phon.ucl.ac.uk/home/sampa/polish.htm.

46. Sathe-Pathak B.V., Panat A.R. (2012), *Extraction of pitch and formants and its analysis to identify 3 different emotional states of a person*, International Journal of Computer Science Issues, Vol. 9, Issue 4, No 1, http://www.ijcsi.org/papers/IJCSI-9-4-1-296-299.pdf.

47. Spangler T., Vinodchandran N.V., Samal A., Green J.R. (2017), *Fractal features for automatic detection of dysarthria*, 2017 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), pp. 437–440, doi: 10.1109/BHI.2017.7897299.

48. Upadhya S.S., Cheeran A.N., Nirmal J.H. (2018), *Thomson Multitaper MFCC and PLP voice features for early detection of Parkinson disease*, Biomedical Signal Processing and Control, **46**, 293–301, doi: 10.1016/j.bspc.2018.07.019.

49. Wei Y., Zeng Y., Li C. (2018), *Single-Channel Speech Enhancement Based on Sub-Band Spectral Entropy*, J. Audio Eng. Soc., **66**, 3, 100–113, doi: 10.17743/jaes.2018.000.

50. Ziółko B., Gałka J., Ziółko M. (2009), *Polish phoneme statistics obtained on large set of written texts*, Computer Science, **10**, 3, 97–106, doi: 10.7494/csci.2009.10.3.97.

51. Ziółko B., Żelasko P., Skurzok D. (2014), *Statistics of diphones and triphones presence on the word boundaries in the Polish language. Applications to ASR*, XXII Annual Pacific Voice Conference (PVC), Krakow, 2014, pp. 1–6, doi: 10.1109/PVC.2014.6845418.