sciendo

Int. J. Appl. Math. Comput. Sci., 2019, Vol. 29, No. 2, 393–405
DOI: 10.2478/amcs-2019-0029

amcs

MACHINE LEARNING–BASED ANALYSIS OF ENGLISH LATERAL ALLOPHONES

MAGDALENA PIOTROWSKA [a], GRAŽINA KORVEL [b], BOŻENA KOSTEK [c,*], TOMASZ CISZEWSKI [a], ANDRZEJ CZYŻEWSKI [a]

[a] Multimedia Systems Department
Gdansk University of Technology, G. Narutowicza 11/12, 80-233 Gdansk, Poland
e-mail: {mplewa,andcz}@multimed.org, angtc@ug.edu.pl

[b] Institute of Data Science and Digital Technologies
Vilnius University, Akademijos 4, LT-04812 Vilnius, Lithuania
e-mail: grazina.korvel@mii.vu.lt

[c] Laboratory of Audio Acoustics
Gdansk University of Technology, G. Narutowicza 11/12, 80-233 Gdansk, Poland
e-mail: bokostek@audioacoustics.org

Automatic classification methods, such as artificial neural networks (ANNs), the k-nearest neighbor (kNN) and self-organizing maps (SOMs), are applied to allophone analysis based on recorded speech. A list of 650 words was created for that purpose, containing positionally and/or contextually conditioned allophones. For each word, a group of 16 native and non-native speakers were audio-video recorded, from which seven native speakers' and phonology experts' speech was selected for analyses. For the purpose of the present study, a sub-list of 103 words containing the English alveolar lateral phoneme /l/ was compiled. The list includes 'dark' (velarized) allophonic realizations (which occur before a consonant or at the end of the word before silence) and 52 'clear' allophonic realizations (which occur before a vowel), as well as voicing variants. The recorded signals were segmented into allophones and parametrized using a set of descriptors, originating from the MPEG 7 standard, plus dedicated time-based parameters as well as modified MFCC features proposed by the authors. Classification methods such as ANNs, the kNN and the SOM were employed to automatically detect the two types of allophones. Various sets of features were tested to achieve the best performance of the automatic methods. In the final experiment, a selected set of features was used for automatic evaluation of the pronunciation of dark /l/ by non-native speakers.

Keywords: allophones, audio features, artificial neural networks (ANNs), k-nearest neighbor (kNN), self-organizing map (SOM).

1. Introduction

The aim of the research presented in this paper is to find efficient feature vectors that will enable automatic assignment of allophones extracted from English speech to an appropriate phonological group. The study performed is to be presented in the context of variations the allophones of /l/ to develop a methodology for other consonantal allophones. The achieved sets of parameters describe model pronunciation of lateral allophones. The goal of this selection is not only to differentiate between allophones, but to determine on the basis of objective measures whether the target phenomena were pronounced correctly. At the final stage of the experiment, dark /l/ performances of non-native English speakers were analyzed using a chosen set of features. Such an approach may be used in the future in an application checking the speaker's pronunciation in the process of language learning and providing him/her with an automatic feedback on speech correctness.

*Corresponding author

© 2019 M. Piotrowska et al.
This is an open access article distributed under
the Creative Commons Attribution-NonCommercial-NoDerivs license
(http://creativecommons.org/licenses/by-nc-nd/3.0/).

Lateral allophones were selected for the analysis as these linguistic phenomena are particularly difficult for Polish speakers. For a few decades, one has observed continuous improvement in the technology of speech-recognition systems. This is due to several factors. One of them is building audio-visual speech corpora in which fast camera recordings and video analysis (visemes) support audio recognition (Almajai *et al*., 2016; Cooke *et al*., 2006; Dalka *et al*., 2014). This concerns both English and national databases (Czyzewski *et al*., 2017b; Kunka *et al*., 2013; Benezeth *et al*., 2011; Trojanová *et al*., 2008; Żelasko *et al*., 2016; Kłosowski, 2017). For the audio part, the utterances may be spoken at a slow and normal speech pace; they may also contain prosodic features to improve the learning process of the audio-visual speech recognition system. Recently, more efficient machine learning methods have also appeared for speech analysis and recognition (Almajai *et al*., 2016; Jadczyk and Ziółko, 2015; Marasek and Gubrynowicz, 2005; Brocki and Marasek, 2015; Aubanel and Nguyen, 2010), such as deep learning (Mroueh *et al*., 2015; Noda *et al*., 2015). Moreover, a renewed interest in phonemic-level-based analyses has appeared (Biswas *et al*., 2015; Czyzewski *et al*., 2013; Kupryjanow and Czyzewski, 2013; Ziółko and Ziółko, 2009; Cooke *et al*., 2006) with applications to various areas, such as, e.g., biometry (Czyzewski *et al*., 2017a).

A unique feature of the audio-visual database (Fox *et al*., 2005) created at the Multimedia Systems Department, Gdansk University of Technology, is the recording of a list of specially selected words containing variations in allophones, thus enabling a more in-depth analysis of speech sounds. For the purpose of the experiments carried out, a series of audio-video recordings was performed to gather the required information. The framework of the experiments is briefly described in Section 2. The current study focuses on the audio analysis. Audio files were edited and segmented into allophones, and then parametrized using a set of descriptors. This is presented in Section 2. From the entire list a sub-list containing words with various allophones of /l/ was chosen for a detailed analysis.

As mentioned before, the rationale behind this study is to provide a methodology for allophone analysis and automatic assignment of selected allophones to an appropriate phonological group. The analysis involved native speakers (Standard Southern British or SSBrE—Standard Southern British English; both male and female) and, at the final stage, recordings of non-native speakers. The accent demonstrated by the phonology experts is also near native SSBrE (Standard Southern British English), and the articulation and distribution of dark /l/ in South African English is the same as in SSBrE. The discussion concerning the phonological features of the audio material assessed is presented in Section 3. The allophones are parametrized and the created feature vectors are checked for redundancy. To this end, the PCA (principal component analysis) is applied. This is shown in Section 3.

As mentioned before, the focus of the present study is on automatic allophone recognition employing optimized feature vectors and the machine learning approach. Application of artificial neural networks (ANNs), the k-nearest neighbor (kNN) and self-organizing maps SOMs is justified by a possibility to compare the approach presented in our study with other research results. Such a comparison cannot, however, be performed in a straightforward way, as other works in this area concentrate around speech recognition and not on elements of speech, such as allophones (Mitterer *et al*., 2018; Ali *et al*., 1999; Kozierski *et al*., 2016; Baghdasaryan and Beex, 2011). Kozierski and his collaborators came to the conclusion that an approach based on allophones cannot directly be used in automatic speech recognition without further research and modification of the employed methods (Kozierski *et al*., 2016). Nevertheless, to show that it is applicable to employ such algorithms to automatic allophone recognition, in Section 4, optimized feature vectors are fed into the ANNs, kNN and SOM algorithms to assign them to an appropriate phonological group, taking into account the speaker-individual features, the gender, as well as the native/non-native context. Finally, conclusions on the application of ANNs, the kNN and SOMs in automatic allophone assignment are drawn.

1.1. Allophonic material. It was observed that, at the level of conscious awareness, listeners are characteristically attuned only to the distinctions between phonemes. Making speakers aware of allophonic variation requires that their attention be carefully directed to the distinction (Giegerich, 1992). An example of such phonological problem in English is the allophonic variants of the phoneme /l/. The group includes velarized (dark) [ɬ], which is articulated with the back of the tongue raised towards the soft palate and occurs word-finally or before another consonant (e.g., ball, fool, all, etc.), dental [l] instead of alveolar when a $\langle th \rangle$ consonant follows (e.g., wealth, health, stealth, etc.), fully devoiced /l/ when the preceding consonant is voiceless (e.g., slight, flight, plow, etc.) and partially devoiced word-initially, whereby the (clear) [l] onset is voiceless and voicing starts at the end of the /l/ articulation (e.g., listen, lose, allow, etc.), and partially devoiced word-final (dark) [ɬ], whereby the allophone is devoiced only towards the end of its articulation. In the production of the dark /l/, the front constriction of the tongue tip is accompanied by a post-dorsal or pharyngeal gesture, which results in observable lowering of the formant frequencies level $L2$

(second formant level) and rising of $L1$ (first formant level) (Giles and Moll, 1975). The variants of /l/ depend on the position of the tongue in the mouth. These differences result in very different formant frequencies.

It was shown by Recasens (2012) that a relatively high F_2, about 1500–2000 Hz for clear /l/, and a lower F_2, about 800–1200 Hz for dark /l/, are characteristic for these allophones. It is also mentioned that F_1 is typically higher for dark /l/ than for clear /l/ (Recasens, 2012). The acoustic analysis of dark and clear /l/ is based on an investigation of the first (F_1) and the second (F_2) formant frequencies in our study.

Table 1. List of speakers included in the analysis (numbered as in the database).

No.	Gender	English spoken	Accent
A	Female	Near native	British
B	Male	Native	British
C	Male	Native	British (Eustary)
D	Male	Native	South African
E	Male	Native	British
F	Female	Near-Native	British
G	Female	Native	British

Table 2. List of words used in the experiment (containing dark variants of the allophone /l/).

Dark /l/	Words
fully voiced	album, already, bulldog, elbow, field, shield, tool, will rise
partially voiced	all, also, bell, bolt, bottle, cattle, comfortable, crawl, example, fail, feel, file, foil, hall, hell, help, jail, level, little, melt, mobile, pill, poll, pub meal, school, scowl, scrawl, shall, shell, sick girl, special, spill, steal, still, stole, style, welcome, zeal
partially voiced, syllabic	handle, middle, saddle
dentalized, partially voiced	although, health

Table 3. List of words used in the experiment (containing clear variants of the allophone /l/).

Clear /l/	Words
fully voiced	align, black cat, blackbird, deadly, family, lilly, nonetheless, will you
partially voiced	lack, lady, lair, lamp, large, lark, late, leak, Lear, learn, leave, leg, leisure, lend, let, level, light, lilly, lit, little, loaf, log cabin, long, longer, look, loot, lord, lot, loud, luck, lunch, lure
voiceless	chocolate, class, clever, close, cloud, complete, play, replace, splash, spleen, split)

1.2. Recordings. A special system consisting of video and audio modalities was prepared, although in this study the authors focus on audio recordings only. The audio files were recorded with a 48 kHz/16 bit resolution with three microphones (an LAV microphone and two condenser microphones situated 50 and 100 cm away from the speaker). Speech samples of 16 speakers (non-native English speakers, as well as English native speakers and a phonology expert) were recorded. For the purpose of this study, seven speakers were selected (Table 1). Four of them are native English speakers with a British accent (Standard Southern British or SBrE), Speaker D is native with a South African accent, and Speakers A and F are near-native phonology experts. As mentioned before, the accent demonstrated by the phonology experts is also near native SSBrE, and the articulation and distribution of dark /l/ in South African English is the same as in SSBrE.

From the recorded dataset of over 600 words, a list of 103 words containing the lateral phoneme was compiled. The list includes 51 'dark' (velarized) allophonic realizations before a consonant or word-finally and 52 'clear' allophonic realizations before a vowel, including different places of articulation and voicing variants (Tables 2 and 3). A total number of 721 samples was collected.

2. Parameterization

Before the parameterization phase took place, the first task was to locate the individual allophones of /l/ in all selected words, edit them, and then annotate them phonologically. Therefore, the /l/ allophones were extracted from all selected words. Indexation was performed manually by a sound engineer experienced in dialogue editing and familiar with English allophony. However, since the accuracy of indexation may have influence on the parameterization and final results, all measurements and indexations were performed with meticulous care and double checked. An automated segmentation system could not be used (Makowski and Hossa, 2014) since there was a need for allophone-focused editing. The duration of extracted /l/ allophones ranges from 30 up to over 200 milliseconds. Such a very detailed analysis of a speech phenomenon may also be useful in automatic detection of voice pathologies (Panek *et al.*, 2015), requiring the editing process to be manual. The goal of the presented study was to find a vector of features that are related to differences between dark and clear variations in the lateral phoneme /l/. Therefore, an extensive set of over 200

parameters including mel-frequency cepstral coefficients (MFCCs), as well as time and energy-based features was calculated. A full list of calculated features is contained in Appendix (Table A1).

Various sub-vectors were tested using automatic classification methods (ANNs, the kNN and SOMs). The analyzed parameters were chosen based on the approach typical for audio analysis (Mermelstein, 1976; Kim *et al*., 2006), also taking into consideration previous studies performed by the authors (Kostek *et al*., 2011; Plewa and Kostek, 2015). Features included in the course of the presented study were chosen to describe spectral and time characteristics which differentiate between allophones. The variants of /l/ are characterized by different formant values and their distribution in time, therefore both domains should be represented. More details about the specific features and their interpretation are included in the summary below. Since they are presented in Appendix along with their description, only those which need an additional explanation are described more thoroughly below.

Parameters 1–14 refer to the time domain and are commonly used in sound analysis (Kostek *et al*., 2011; Plewa and Kostek, 2015; Song *et al*., 2009). Parameters 5–14 are obtained through the analysis of the signal samples distribution in relation to the RMS. Three reference levels were defined: r_1, r_2, r_3 (equal to RMS, 2×RMS and 3×RMS in the analyzed signal frame). Parameters 5, 6 and 7 correspond to the number of samples exceeding levels: r_1, r_2, r_3, and are defined by the formula

$$p_n = \frac{\text{count samples exceeding } r_n}{\text{length}(x(k))}, \tag{1}$$

where $n = 1, 2, 3$ and $x(k)$ represents the analyzed signal fragment. In order to obtain a more thorough description of the RMS (root mean square) energy changes in the analyzed frame, Kostek *et al*. (2011) devised and introduced another approach, where each frame is divided into 10 smaller segments. In each of these, parameters p_n (Eqn. (1)) are calculated. As a result, the P_n sequence

$$P_n = (p_n^1, p_n^2, p_n^3, \ldots, p_n^{10}) \tag{2}$$

is obtained, where $n = 1, 2, 3$.

In this way, six new features (parameters 8–13) were defined on the basis of P_n sequences. The features are denoted as the mean (q_n) and variance (v_n) of P_n and are calculated as follows:

$$q_n = \frac{1}{10}\sum_{k=1}^{10} p_n^k, \tag{3}$$

$$v_n = \frac{1}{9}\sum_{k=1}^{10} (p_k^n - q_n), \tag{4}$$

where $n = 1, 2, 3$. Index n is related to different reference values of r_1, r_2 and r_3. Parameter 14 is defined as the 'peak to the RMS' ratio, calculated as the ratio in the entire frame. The additional parameters (15–16) are the mean and variance values of the 'peak to RMS' ratio calculated in 10 sub-frames.

Parameters 17–28 are based on the observation of the zero crossing rate and the threshold crossing rate (TCR). These values (similarly to other previously presented parameters) are defined in three different ways: the value for the entire frame and as the mean and variance of the TCR calculated for 10 sub-frames.

Parameters 29–36 are related to the frequency domain, which often refers to audible characteristics of the sound (Misra *et al*., 2004).

Parameters 37–56: mel-frequency cepstral coefficients (MFCCs). They were introduced by Mermelstein (1976) as a tool for speech recognition and are among the most widely used acoustic features in speech and audio processing (Kłosowski, 2017; Kupryjanow and Czyzewski, 2013). They are described as a low-dimensional representation of the spectrum divided according to the mel-scale, which reflects the nonlinear frequency sensitivity of the human auditory system.

Parameters 57–172 represent audio spectral features. They are similar to parameters 29–36, but the calculation method is different. Since features extraction is to be performed on an allophone signal divided into short-time segments, in this study, the input signal is segmented into frames of 1024 samples, and then, for each frame, the Hamming windows are applied with an overlap of 50%.

Parameters 173–212 represent MFCC mean values and variances. MFCC mean values given as averaged MFCC were obtained from each segment. As in the case of audio spectral features, the speech signal is divided into short-time segments.

Parameters 213–232 are modified MFCCs. The MFCCs are derived from discrete Fourier transform (DFT) spectra. The spectrum is a combination of the source and a filtered representation, and contains both speaker-dependent and phoneme-dependent information. In order to achieve a higher accuracy of phoneme separation, we may have to eliminate information dependent on the speaker's gender. For this purpose, we consider the fundamental frequency as the main discriminating factor between male and female voices. The fundamental frequency determines the distribution of harmonics, which consequently yields a different distribution of energy through the frequency values. Based on this principle, the boundary points of the filter bank are constructed with regard to the extension coefficient, which depends on the ratio of female and male fundamental frequencies and is different for the

female and male voices. The process of creating the modified MFCC extraction is shown in Fig. 1, in which the phoneme signal is denoted as $s(n)$, where $n = 1, \ldots, N$ (N is the number of samples).

In order to determine the extension coefficient, we use values known from the literature, i.e., a typical male adult has a fundamental frequency (F_0) between 85 Hz to 180 Hz and an adult female has a fundamental frequency in the range of 165 Hz to 225 Hz (Baken and Orlikoff, 2000). For further analysis, we choose mean values, i.e., $F_0 = 132$ for the male and $F_0 = 195$ for female. The extension coefficient (denoted by the symbol Ex) is equal to the ratio of the female and male fundamental frequencies, i.e., $Ex = 1.45$. Moreover, the analyses performed show that the pitch is linear in low frequencies. Therefore, we may assume that the frequency scale is linear with a logarithmic spacing. In the literature, the upper boundary of the frequency range above which the scale becomes logarithmic is often set to 1000 Hz (Wang and Van Hamme, 2011). In this paper, the fixed length of the linear band $F_0 = 66.67$ Hz is used.

The threshold between the linear and logarithmic scale is defined by the number of linear bands. The frequency scale used in MFCC calculation depends on four variables: M_l (number of linear bands), M_{nl} (number of nonlinear bands), f_h (highest frequency of the filter bank), and l_b (length of the linear band). As a result of the above, we propose the following definition of the filter bank with M filters ($M = M_l + M_{nl}$):

$$H_m(k) = \begin{cases} 0, & k < f[m-1], \\ \frac{k - f[m-1]}{f[m] - f[m-1]}, & f[m-1] \le k \le f[m], \\ \frac{f[m+1] - k}{f[m+1] - f[m]}, & f[m] \le k \le f[m+1], \\ 0, & k > f[m+1], \end{cases} \tag{5}$$

where k is the linear frequency and m is the filter number ($1 \le m \le M$). The boundary points $f[n]$ are

$$f[n] = D[n] \times \begin{cases} 1 & \text{if female}, \\ Ex, & \text{if male}, \end{cases} \tag{6}$$

where

$$D[n] = \begin{cases} 0, & n = 0, \\ f[n-1] + l_b, & 1 \le n \le M_l + 1, \\ \text{Mel}^{-1}\Big(\text{Mel}(f_l) & \\ +n\frac{\text{Mel}(f_h) - \text{Mel}(f_l)}{M+1}\Big), & M_l{+}1 < n \le M{+}1. \end{cases} \tag{7}$$

$$\text{Mel}(f) = 2595 \log\left(1 + \frac{f}{700}\right), \tag{8}$$

$$f_l = M_l \times (l_b + 1). \tag{9}$$

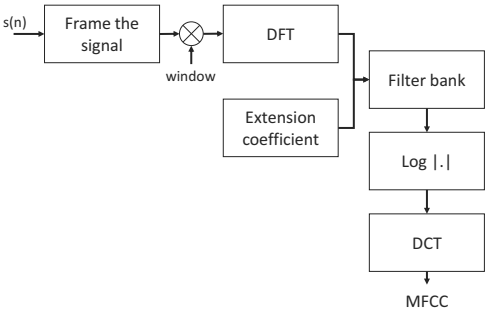

Fig. 1. Modified MFCC extraction process.

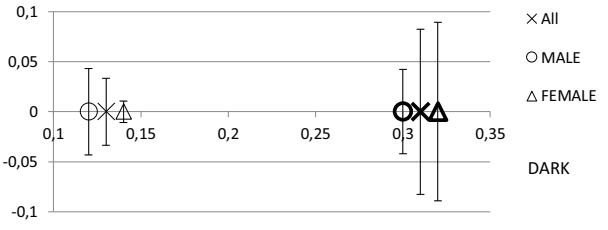

Fig. 2. Values of parameter 7 (number of samples exceeding the r_3 threshold) for dark and clear lateral allophones averaged for all speakers, and males and females separately. Clear allophones are marked with the bold line.

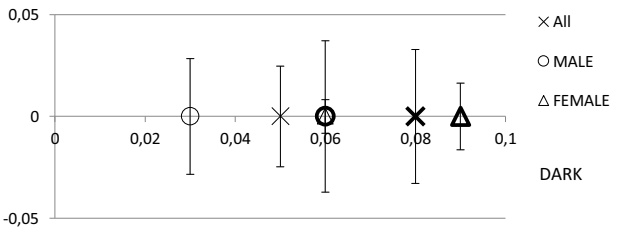

Fig. 3. Values of parameter 166 (audio spectrum kurtosis variance) for dark and clear lateral allophones averaged for all speakers, and males and females separately. Clear allophones are marked with the bold line.

The log-magnitude is calculated in order to obtain the real cepstrum. The energies from filters to cepstral coefficients are converted by the discrete cosine transform (DCT). To calculate modified MFCC features, the speech signal is divided into short-time segments.

3. Analysis

The starting point of the analyses was the parameterization of all audio samples and subsequent optimization of features in the context of their separability and usability in the allophone recognition process. In order to perform the analysis of features, they should be normalized first. It was decided to normalize the values to the range $[0, 1]$. The maximum and minimum values were found for each feature. The normalized values were calculated using the following formula:

$$z_i = \frac{x_i - \min(\boldsymbol{x})}{\max(\boldsymbol{x}) - \min(\boldsymbol{x})}, \tag{10}$$

where $\boldsymbol{x} = (x_1, \ldots, x_i)$ is the vector of non-normalized values for the selected feature, z_i is the normalized value, i is the number of samples.

An analysis of average values for dark and clear allophones was performed for each feature. Some of parameters clearly differentiate between these two groups (the number of samples exceeding threshold r_3 in Fig. 2), while for others (i.e., the audio spectrum kurtosis variance presented in Fig. 3) these trends are not observed.

Also, the difference of averaged values between clear and dark /l/ was calculated for every parameter as follows:

$$\text{difference} = \text{dark}/l/ - \text{clear}/l/. \tag{11}$$

Selected results are included in Figs. 4–6. A positive value of the difference indicates that values of a parameter are higher for dark /l/, while negative that values are higher for clear allophones. For some features, the difference is positive (Fig. 4) or negative (Fig. 5) for all speakers while for others (i.e., Fig. 6) it varies between speakers.

Principal component analysis was performed to achieve possibly most orthogonal dimensions (Smith, 2017). The PCA was applied to the set consisting of 232 parameters (Table A1) calculated for all samples. As a result, we arrived at 120 components that are sufficient to contain 99% of the information. All of the PCA calculations were performed in the MATLAB environment. In addition, a correlation analysis was performed. Based on the correlation coefficient (>0.5), sets of features correlated with all speakers, male and female groups, were created. The various sets of features presented in Table 4 were used in the allophone recognition stage employing the ANN, kNN and SOM algorithms. Some of the examined vectors are based on our previous research (Piotrowska *et al.*, 2018), while others were introduced specifically for this study.

4. Automatic classification of allophones

Extracting an allophone from the audio signal containing a whole word or a sentence is an arduous process, resulting in uncertain data. In addition, the basic difficulties lie in identifying a set of features that are unique for a selected allophone. Parameterization of allophones is still not a well-researched area, as an allophone is a very short utterance; thus typical descriptors used in the automatic speech recognition domain may not be applicable. Such conditioning leads to two conclusions: first of all, the machine learning approach is a good choice to deal with uncertain data, and secondly, for finding an appropriate feature vector, parameters from the speech and the music domain should be applied, modified according to the needs and tested along with several machine learning algorithms.

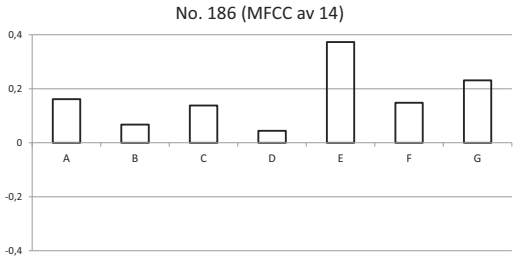

Fig. 4. Difference between dark and clear /l/ calculated for individual speakers for parameter 186 (MFCC 14 average).

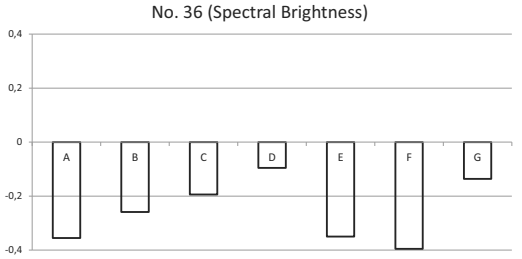

Fig. 5. Difference between dark and clear /l/ calculated for individual speakers for parameter 36 (spectral brightness).

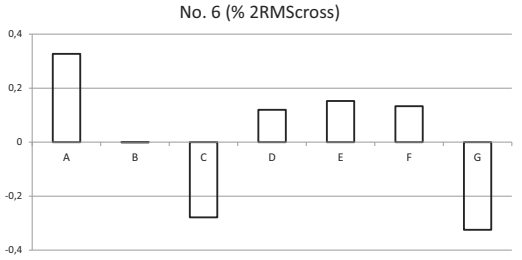

Fig. 6. Difference between dark and clear /l/ calculated for individual speakers for parameter 6 (number of samples exceeding the level r_2).

For the purpose of the present study three algorithms were chosen: artificial neural networks, the k-nearest neighbor and self-organizing maps. The ANN algorithm has an excellent generalization capability to learn from the set of data and has been used for decades in automatic speech recognition. Moreover, recently ANNs have been employed in automatic phoneme recognition, hence the research very closely related to allophone extraction and classification (Mitterer *et al.*, 2018; Kozierski *et al.*, 2016).

The kNN algorithm is very simple and versatile. Moreover, its main feature is that it is a non-parametric learning algorithm. Thus it does not make any assumptions on the underlying data distribution. Moreover, it is often utilized as a kind of benchmark for more complex classifiers such as ANNs. Thus this was the motivation behind employing the kNN in the experiment. The last mentioned algorithm was chosen, since the editing process involved both a sound engineer and a phonology expert.

Contrarily to what was said before, the editing process was highly controlled in the present study. All extracted allophone samples after the editing process were evaluated, corrected if necessary and then approved by the expert as correct. Thus, even though the editing process remains to be to some extent subjective and uncertain, still it enables us to employ SOMs for unsupervised learning to organize objects in the form of a low-dimensional map. As this method is data driven and finds clusters without any external guidance, such properties have led to the implementation of SOMs in related areas, among others, to speech (Venkateswarlu and Kumari, 2011; Wang and Van Hamme, 2011) and music information retrieval (Pampalk *et al.*, 2002). Moreover, SOMs are strongly related to human perception; thus they are especially adequate for discerning speech elements, automatically. Lastly, selected algorithms enabled performing a comparison between learning and the self-driven approach.

Table 4. Sets of features used in the analyses.

Set label	Features
232	All
PCA	120 components achieved from PCA analysis
time	2–28
time+SPEC	2–36,159–172
time+ SPEC +MFCC	2–36,159–212
time+ SPEC +ASE	2–36,159–172, 57–116
time+ SPEC +SFM	2–36, 117–172
time+ SPEC +modMFCC	2–36, 159–172, 213–232
ASE	57–116
SFM	117–158
SPEC	29–36, 159–172
C_All	Set based on correlation for all speakers
C_Male	Set based on correlation for all male speakers
C_Female	Set based on correlation for all female speakers

4.1. Feedforward neural network-based classification. The ANN-based classification was performed using the *nntool* within the Matlab environment. The created set was divided randomly into three subsets: training (70%), validation (15%) and testing (15%). A feed-forward ANN with one hidden layer was trained to classify clear/dark /l/ allophones. Various feature vectors listed in Table 4 were fed to the input of the ANN. Various configurations of ANNs were tested, but the best results were obtained for a network with 10 neurons in the hidden layer. The PCA did not improve the results for the analyzed set and was therefore removed from further analysis. The analyses were performed for the whole group of speakers, for male and female separately, as well as for individual ones (see Tables 5 and 6).

4.2. k-Nearest neighbor results. In the presented study, the value k was set to 7. The optimum value for k was established by performing a series of preliminary tests. An experiment was repeated 50 times for each case, and the arithmetic mean was calculated. The results for various sets of features are presented in Tables 7 and 8.

The accuracy tests for particular speakers were also performed. The highest result of 97% was achieved for

Table 5. ANN accuracy results for all speakers for various sets of features (the feature vectors are labeled according to Table 4).

Set of parameters	All [%]	Male [%]	Female [%]
232 features	97	95	98
PCA (120)	96	95	97
C_All/ C_Male/ C_Female	98	99	99
Time-related	89	86	92
ASE	82	88	78
SFM	61	60	67
Spec	75	80	82
time+SPEC	92	94	94
time+SPEC+MFCC	95	96	94
time+SPEC+ASE	94	97	95
time+SPEC+SFM	91	93	94
time+SPEC+ modMFCC	98	97	98

Table 6. ANN-based classification accuracy for single-speaker dark/clear /l/ distinction for different sets of features (the feature vectors are labeled according to Table 4).

Speaker	232 [%]	C_All [%]	C_Male/ C_Female [%]
A	98	98	98
B	94	94	93
C	94	94	94
D	97	96	96
E	93	92	91
F	89	89	90
G	90	90	90

Speaker D, while Speaker F scored only 86% (Table 7).

4.3. Self-organized map clusterization. The SOM-based clusterization was performed using the Matlab Neural Networks Tool. A network with a rectangular topology and two output classes were defined. A training including 500 iterations was performed. The results of SOM classification are included in Tables 9 and 10. Different SOM input feature vectors were tested. In addition to the feature vectors of 232 parameters and components obtained from the PCA, some specially created sets were also used. The features were selected on the basis of large and coherent differences for all the speakers. The feature sets are described in detail in Table 4. The PCA did improve the results by only 2–3% for the set analyzed and was therefore excluded from further analysis. The analyses were performed for the whole group of speakers, as well as for male and female ones separately.

Table 7. kNN accuracy results for all speakers for various sets of features (the feature vectors are labeled according to Table 4).

Set of parameters	All [%]	Male [%]	Female [%]
232 features	93	94	92
C_All/ C_Male/ C_Female	93	95	91
Time-related	94	94	91
ASE	74	79	90
SFM	82	87	77
SPEC	82	80	86
time+SPEC	94	91	94
time+SPEC+MFCC	93	92	93
time+SPEC+ASE	94	91	94
time+SPEC+SFM	93	92	93
time+SPEC+ modMFCC	95	94	95

Table 8. kNN classification accuracy for single-speaker dark-clear /l/ distinction for different sets of features (the feature vectors are labeled according to Table 4).

Speaker	232 [%]	C_All [%]	C_Male/ C_Female [%]
A	91	92	94
B	94	94	95
C	93	93	95
D	95	95	97
E	93	91	91
F	84	86	86
G	93	93	93

4.4. Comparison of results. Various sets of features were tested during the course of the present study. The best results for all classifiers were obtained for vector time+SPEC+modMFCC (Figs. 7–9), which includes time- and spectral-based features along with the modified MFCCs proposed by the authors. This comparison is presented in Table 11. The highest accuracy results were obtained using the ANN, although for the Time+SPEC feature vectors the kNN and ANN returned very similar scores. Still, for the whole group of features, kNN accuracy was the highest (94% compared with 92% for the ANN and 72% for the SOM). For all feature sets, the lowest accuracy was achieved with SOMs (between 70% and 80%), which can be explained by its characteristic feature, namely, the lack of supervised training.

Table 9. SOM accuracy results for all speakers for various sets of features (the feature vectors are labeled according to Table 4).

Set of parameters	All [%]	Male [%]	Female [%]
232 features	75	75	87
PCA (120)	77	73	80
C_All/ C_Male/ C_Female	61	81	79
Time-related	61	80	70
ASE	63	72	58
SFM	54	54	58
Spec	65	74	81
time+SPEC	72	76	85
time+SPEC+MFCC	80	77	86
time+SPEC+ASE	66	74	87
time+SPEC+SFM	71	74	85
time+SPEC+ modMFCC	81	77	87

Table 10. SOM classification results for different sets of features (the feature vectors are labeled according to Table 4).

Speaker	232 [%]	C_All [%]	C_Male/ C_Female [%]
A	87	87	87
B	88	91	92
C	83	90	92
D	90	90	93
E	90	91	92
F	62	64	65
G	91	91	91

Table 11. SOM, ANN and kNN highest accuracy results for set of features with the highest accuracy (time+SPEC+modMFCC)).

Classifier	All [%]	Male [%]	Female [%]
ANN	98	97	98
kNN	95	94	95
SOM	81	77	87

Table 12. List of words containing the dark /l/ allophone used in the final stage of the experiment.

Dark /l/	also, hell, shell, sick girl, kill, feel, steal, jail, shield, field

Table 13. Results of automatic evaluation of dark lateral allophones.

Classifier	ANN [%]	kNN [%]	SOM [%]
Accuracy	83	52	62
Precision	85	78	82
Recall	95	73	65
F1 score	89	75	73

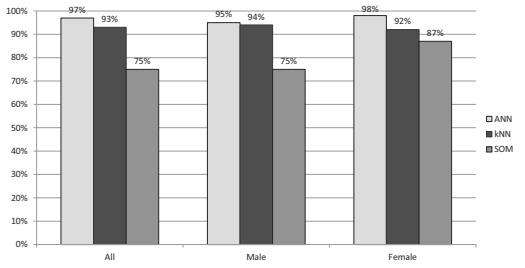

Fig. 7. Comparison of accuracy results for various automatic classification methods using the 232 vector of features.

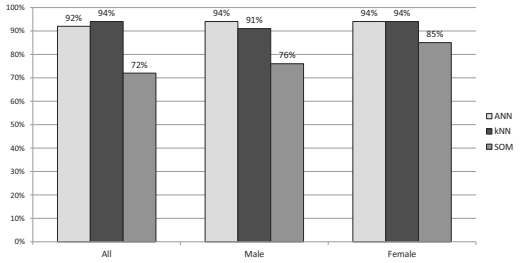

Fig. 8. Comparison of accuracy results for various automatic classification methods using the Time+SPEC vector of features.

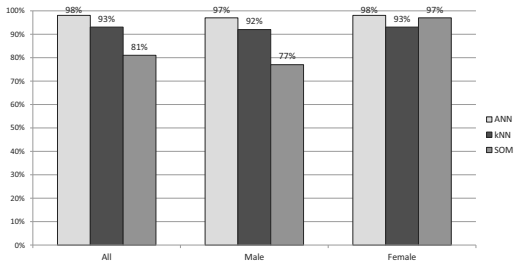

Fig. 9. Comparison of accuracy results for various automatic classification methods using the Time+SPEC+modMFCC vector of features.

5. Automatic evaluation of dark /l/ pronunciation

At this stage of the experiment, dark /l/ performances of non-native English speakers were analyzed using Time+SPEC+modMFCC vector of features. Recordings for nine non-native speakers of 11 words containing dark /l/ listed in Table 12 were executed.

The performance of speakers was evaluated by a phonology expert and treated as a ground truth for automatic classification. The data set consisted of 190 dark /l/ performances, including 44 incorrect trials. The conducted study allowed determining the set of features describing correct dark and clear lateral allophones; the ANN, kNN and SOM methods were implemented to analyze whether non-native performances fit into dark /l/ criteria. The accuracy of these analyses is presented in Table 13. As can be seen, apart from accuracy, three additional performance measures were calculated. Precision and recall give us exactness and completeness of the classifiers, respectively, while the F1 score shows us the balance between precision and recall.

In order to determine whether the apparent differences in performance of the algorithms are statistically significant, McNemar's test is used (McNemar, 1947). Gillick and Cox (1989) recommend to use it in comparison of algorithms that classify isolated words. This was the main reason for choosing this test. The results of McNemar's test are given in Table 14.

The obtained p-values of Table 14 are compared with the test significance level α. In the experiment, α equals 0.05. According to the test results, the differences

Table 14. Results of McNemar's test for the ANN, kNN and SOM methods.

Classifiers	ANN/kNN	ANN/SOM	SOM/kNN
p-value	<0.0001	<0.0001	0.0525

between the ANN and SOM as well as between the ANN and kNN are considered to be extreme cases of statistical significance. Meanwhile, the differences between the SOM and kNN are considered to be statistically insignificant. This means that the SOM and kNN algorithms have more errors in common compared with ANNs.

6. Conclusions

The results obtained in this study show that utilizing parameterization as a pre-processing stage in the classification process enables an automatic assignment of selected allophones of lateral /l/ to appropriate phonological group with high accuracy, i.e., over 98% for the ANN, 95% for the kNN and over 80% for the SOM.

Also, the results demonstrate that the proposed set of parameters represents differences between dark and clear /l/ allophones. The ANN, kNN and SOM accuracy for all speakers, male and female, was the highest for the selected time and frequency domain-based parameters supplemented with the modified MFCCs constructed by the authors. Phonetic experiments showed that dark /l/ is characterized by observable convergence of formants F_1 and F_2 compared with clear /l/. This is in line with the current automatic classification, which uses energy distribution in consecutive signal bands. The study shows that the PCA did not improve the SOM clustering accuracy for the performed study. However, this should be further checked with other group of allophones, as the PCA typically helps to constrain data redundancy and at the same time improves accuracy.

The results obtained in this study lead to the conclusion that a separate analysis for male and female speakers is possible. Noteworthy is the fact that better accuracy was achieved for female speakers, although due to the limited number of speakers these observations cannot be considered a general rule. The analysis conducted separately for individual speakers shows remarkable differences among them. It was also observed that some speakers' utterances were more difficult for all classifiers (e.g., Speaker F). Although the lowest accuracy was achieved with SOMs, they still returned a high accuracy between 70% and 80%. Compared with the ANN and kNN methods, which require supervised training (92–98% accuracy for best performing feature vectors), the SOM-based results are promising and will be included in further research.

An application of deep learning is also planned, although it would need a recording of a much larger spoken material volume, and this, in turn, entails big effort required for its annotation. Finally, making the annotating process along with feature space extraction automatic would also need approaches used in big data analytics. As described by Stefanowski *et al*. (2017), deep neural networks may help in automating tasks such as feature space construction, as this becomes an inherent part of the training process. However, this is true from the big data perspective: in the case of a classical approach to automatic speech recognition (ASR), instead of performing feature extraction, it is possible to use 2D feature spaces derived from signal spectrograms. In this way the data representation grows considerably and it makes it possible to use a 2D feature representation along with convolutional neural networks (CNNs) in the speech/allophone automatic recognition process (Korvel *et al*., 2019). We will work on these issues in the future.

Acknowledgment

This research was sponsored by the National Science Center in Poland (2015/17/B/ST6/01874).

References

Ali, A.A., Van der Spiegel, J., Mueller, P., Haentjens, G. and Berman, J. (1999). An acoustic-phonetic feature-based system for automatic phoneme recognition in continuous speech, *Proceedings of the 1999 IEEE International Symposium on Circuits and Systems, ISCAS'99, Orlando, FL, USA*, Vol. 3, pp. 118–121.

Almajai, I., Cox, S., Harvey, R. and Lan, Y. (2016). Improved speaker independent lip reading using speaker adaptive training and deep neural networks, *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China*, pp. 2722–2726.

Aubanel, V. and Nguyen, N. (2010). Automatic recognition of regional phonological variation in conversational interaction, *Speech Communication* **52**(6): 577–586.

Baghdasaryan, A.G. and Beex, A. (2011). Automatic phoneme recognition with segmental hidden Markov models, *2011 Conference Record of the 45th Asilomar Conference on Signals, Systems and Computers (ASILOMAR), Pacific Grove, CA, USA,* pp. 569–574.

Baken, R.J. and Orlikoff, R.F. (2000). *Clinical Measurement of Speech and Voice*, 2nd Edn., Singular Thomson Learning, San Diego, CA.

Benezeth, Y., Bachman, G., Le-Jan, G., Souviraà-Labastie, N. and Bimbot, F. (2011). *BL-Database: A French Audiovisual Database for Speech Driven Lip Animation Systems*, PhD thesis, INRIA, Rennes.

Biswas, A., Sahu, P.K. and Chandra, M. (2015). Multiple camera in car audio-visual speech recognition using phonetic and

visemic information, *Computers & Electrical Engineering* **47**(2015): 35–50.

Brocki, Ł. and Marasek, K. (2015). Deep belief neural networks and bidirectional long-short term memory hybrid for speech recognition, *Archives of Acoustics* **40**(2): 191–195.

Cooke, M., Barker, J., Cunningham, S. and Shao, X. (2006). An audio-visual corpus for speech perception and automatic speech recognition, *The Journal of the Acoustical Society of America* **120**(5): 2421–2424.

Czyzewski, A., Bratoszewski, P., Hoffmann, P., Lech, M. and Szczodrak, M. (2017a). The project IDENT: Multimodal biometric system for bank client identity verification, *International Conference on Multimedia Communications, Services and Security, Poznań, Poland*, pp. 16–32.

Czyzewski, A., Kostek, B., Bratoszewski, P., Kotus, J. and Szykulski, M. (2017b). An audio-visual corpus for multimodal automatic speech recognition, *Journal of Intelligent Information Systems* **49**(2): 167–192.

Czyzewski, A., Kostek, B., Ciszewski, T. and Majewicz, D. (2013). Language material for English audiovisual speech recognition system development, *The Journal of the Acoustical Society of America* **134/5**: 4069.

Dalka, P., Bratoszewski, P. and Czyzewski, A. (2014). Visual lip contour detection for the purpose of speech recognition, *2014 International Conference on Signals and Electronic Systems (ICSES), Poznań, Poland*, pp. 1–4.

Fox, N.A., O'Mullane, B.A. and Reilly, R.B. (2005). Valid: A new practical audio-visual database, and comparative results, *International Conference on Audio and Video-Based Biometric Person Authentication, Rye Brook, NY, USA*, pp. 777–786.

Giegerich, H.J. (1992). *English Phonology: An Introduction*, Cambridge University Press, Cambridge.

Giles, S.B. and Moll, K.L. (1975). Cinefluorographic study of selected allophones of English /i/, *Phonetica* **31**(3–4): 206–227.

Gillick, L. and Cox, S.J. (1989). Some statistical issues in the comparison of speech recognition algorithms, *1989 International Conference on Acoustics, Speech, and Signal Processing, ICASSP-89, Glasgow, UK*, pp. 532–535.

Jadczyk, T. and Ziółko, M. (2015). Audio-visual speech processing system for polish with dynamic Bayesian network models, *Proceedings of the World Congress on Electrical Engineering and Computer Systems and Science (EECSS 2015), Barcelona, Spain*, pp. 13–14.

Kim, H.-G., Moreau, N. and Sikora, T. (2006). *MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval*, John Wiley & Sons, Chichester.

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**(1): 5.

Korvel, G., Kurowski, A., Kostek, B. and Czyzewski, A. (2019). Speech analytics based on machine learning, *in* G. Tsihrintzis *et al.* (Eds.), *Machine Learning Paradigms*, Springer, Cham, pp. 129–157.

Kostek, B., Kupryjanow, A., Zwan, P., Jiang, W., Raś, Z.W., Wojnarski, M. and Swietlicka, J. (2011). Report of the ISMIS 2011 contest: Music information retrieval, *International Symposium on Methodologies for Intelligent Systems, Warsaw, Poland*, pp. 715–724.

Kozierski, P., Sadalla, T., Drgas, S. and Dąbrowski, A. (2016). Allophones in automatic whispery speech recognition, *21st International Conference on Methods and Models in Automation and Robotics (MMAR), Międzyzdroje, Poland*, pp. 811–815.

Kunka, B., Kupryjanow, A., Dalka, P., Bratoszewski, P., Szczodrak, M., Spaleniak, P., Szykulski, M. and Czyzewski, A. (2013). Multimodal English corpus for automatic speech recognition, *Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA), Poznań, Poland*, pp. 106–111.

Kupryjanow, A. and Czyzewski, A. (2013). Real-time speech signal segmentation methods, *Journal of the Audio Engineering Society* **61**(7/8): 521–534.

Makowski, R. and Hossa, R. (2014). Automatic speech signal segmentation based on the innovation adaptive filter, *International Journal of Applied Mathematics and Computer Science* **24**(2): 259–270, DOI: 10.2478/amcs-2014-0019.

Marasek, K. and Gubrynowicz, R. (2005). Multi-level annotation in SpeeCon Polish speech database, *in* L. Bolc *et al.* (Eds.), *Intelligent Media Technology for Communicative Intelligence*, Springer, Berlin/Heidelberg, pp. 58–67.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages, *Psychometrika* **12**(2): 153–157.

Mermelstein, P. (1976). Distance measures for speech recognition, psychological and instrumental, *in* C.H. Chen (Ed.), *Pattern Recognition and Artificial Intelligence*, Vol. 116, Academic Press, New York, NY, pp. 374–388.

Misra, H., Ikbal, S., Bourlard, H. and Hermansky, H. (2004). Spectral entropy based feature for robust ASR, *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Montreal, Canada*, EPFL-CONF-83132.

Mitterer, H., Reinisch, E. and McQueen, J.M. (2018). Allophones, not phonemes in spoken-word recognition, *Journal of Memory and Language* **98**(2018): 77–92.

Mroueh, Y., Marcheret, E. and Goel, V. (2015). Deep multimodal learning for audio-visual speech recognition, *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, QLD, Australia*, pp. 2130–2134.

Noda, K., Yamaguchi, Y., Nakadai, K., Okuno, H.G. and Ogata, T. (2015). Audio-visual speech recognition using deep learning, *Applied Intelligence* **42**(4): 722–737.

Pampalk, E., Rauber, A. and Merkl, D. (2002). Using smoothed data histograms for cluster visualization in self-organizing maps, *International Conference on Artificial Neural Networks, Madrid, Spain*, pp. 871–876.

Panek, D., Skalski, A., Gajda, J. and Tadeusiewicz, R. (2015). Acoustic analysis assessment in speech pathology detection, *International Journal of Applied Mathematics and Computer Science* **25**(3): 631–643, DOI: 10.1515/amcs-2015-0046.

Piotrowska, M., Korvel, G., Kostek, B., Rojczyk, A. and Czyzewski, A. (2018). Objectivization of phonological evaluation of speech elements by means of audio parametrization, *2018 11th International Conference on Human System Interaction (HSI), Gdańsk, Poland*, pp. 325–331.

Plewa, M. and Kostek, B. (2015). Music mood visualization using self-organizing maps, *Archives of Acoustics* **40**(4): 513–525.

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

Song, Y., Wang, W.-H. and Guo, F.-J. (2009). Feature extraction and classification for audio information in news video, *International Conference on Wavelet Analysis and Pattern Recognition, ICWAPR 2009, Baoding, China*, pp. 43–46.

Stefanowski, J., Krawiec, K. and Wrembel, R. (2017). Exploring complex and big data, *International Journal of Applied Mathematics and Computer Science* **27**(4): 669–679, DOI: 10.1515/amcs-2017-0046.

Trojanová, J., Hrúz, M., Campr, P. and Železnỳ, M. (2008). Design and recording of Czech audio-visual database with impaired conditions for continuous speech recognition, *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco*, pp. 1–5.

Venkateswarlu, R. and Kumari, R.V. (2011). Novel approach for speech recognition by using self-organized maps, *2011 International Conference on Emerging Trends in Networks and Computer Communications (ETNCC), Udaipur, India*, pp. 215–222.

Wang, Y. and Van Hamme, H. (2011). Gaussian selection using self-organizing map for automatic speech recognition, *International Workshop on Self-Organizing Maps, Espoo, Finland*, pp. 218–227.

Żelasko, P., Ziółko, B., Jadczyk, T. and Skurzok, D. (2016). AGH corpus of Polish speech, *Language Resources and Evaluation* **50**(3): 585–601.

Ziółko, B. and Ziółko, M. (2009). Time durations of phonemes in Polish language for speech and speaker recognition, *Language and Technology Conference, Poznań, Poland*, pp. 105–114.

Magdalena Piotrowska received the doctoral degree with distinction at the Faculty of Electronics, Telecommunications and Informatics, Gdansk University of Technology, Poland. Her thesis, entitled *Automatic Mood Indexing of Music Excerpts Based on Correlation Between Subjective Evaluation and Feature Vector*, was dedicated to music information retrieval. She is a member of the Audio Engineering Society, where she had served as the governor and recently as the chair of the Education Committee. Her main scientific interests are music information retrieval, psychoacoustics, signal processing, and connections between science and practical applications in the field of audio technology.

Gražina Korvel received her BSc degree in mathematics and her MSc degree in informatics (with honors) from the Lithuanian University of Educational Sciences in 2007 and 2009, respectively. She received the doctoral degree from the Vilnius University Institute of Data Science and Digital Technologies (former Institute of Mathematics and Informatics) in 2013. Currently she works in that institution. Her research interests include speech signal processing, developing mathematical models, applications of soft computing and computational intelligence.

Bożena Kostek holds a professorship at the Faculty of Electronics, Telecommunications and Informatics, Gdansk University of Technology, Poland. She is a corresponding member of the Polish Academy of Sciences and a fellow of the Audio Engineering Society. Her main scientific interests are signal processing, music information retrieval, psychoacoustics, multimedia, as well as applications of soft computing to the domains mentioned. She is a recipient of many prestigious awards for research, including those of the Prime Minister of Poland, the Ministry of Science and the Polish Academy of Sciences. She is the editor-in-chief of the *Journal of the Audio Engineering Society*.

Tomasz Ciszewski (PhD, associate professor) works at the Gdansk University of Technology, Faculty of Electronics, Telecommunication and Informatics, and at the University of Gdansk, Faculty of Languages, Institute of English and American Studies. He is a University of Lodz graduate (1995), and his PhD thesis (2000) was devoted to phonological analysis of the English stress system in a non-linear conditions-and-parameters approach. He is an author of several papers published in domestic and international journals and conference proceedings on English phonetics and theoretical phonology. He has also published two books: *The English Stress System: Conditions and Parameters* and *The Anatomy of the English Metrical Foot: Acoustics, Perception and Structure* (PeterLang Verlag).

Andrzej Czyżewski (PhD, DSc, Eng) is a full professor at the Faculty of Electronics, Telecommunication and Informatics of the Gdansk University of Technology. He is an author or a co-author of more than 600 scientific papers in international journals and conference proceedings. He has supervised more than 30 R&D projects funded by the Polish government and has participated in 7 European projects. He is also an author of 15 Polish and 7 international patents. He has extensive experience in soft computing algorithms and their applications in sound and image processing. He is a recipient of many prestigious awards, including the first prize of the Prime Minister of Poland for research achievements (in 2000 and 2015). Andrzej Czyzewski chairs the Multimedia Systems Department at the Gdansk University of Technology.

Appendix

Audio features

A full list of 232 features employed in the presented research is included in Table A1.

Table A1. List of calculated features.

No.	Feature
1	Number of samples
2	Temporal centroid
3	Zero crossing rate
4	Root mean square energy
5–7	Normalized number of samples exceeding $r_1/r_2/r_3$ threshold
8–13	Normalized mean values and variations of samples exceeding $r_1/r_2/r_3$, averaged for 10 frames
14	Peak to the RMS ratio
15–16	Normalized mean values and variations of the peak to the RMS ratio averaged for 10 frames
17–20	Normalized number of signal crossings in relation to $0/r_1/r_2/r_3$
21–28	Normalized mean values and variations of $0/r_1/r_2/r_3$ crossing averaged for 10 frames
29	Centroid (first moment of the spectrum)
30	Spread (standard deviation of the data)
31	Skewness (third central moment of the spectrum)
32	Kurtosis (fourth standardized moment of the spectrum)
33	Flatness (ratio between the geometric mean and the arithmetic mean)
34	Entropy (the relative Shannon entropy of the input)
35	Rolloff (the frequency such that a certain fraction (here 0.85) of the total energy is contained below that frequency)
36	Brightness (fixing the cut-off frequency, and measuring the amount of energy above that frequency)
37–56	20 Mel-frequency cepstral coefficients
57–85	Audio spectrum envelope (ASE) mean values in 29 frequency bands)
86	ASE mean value (averaged for all frequency bands)
87–115	ASE variance values in 29 frequency bands
116	Mean ASE variance parameters
117–136	Spectral flatness measure (SFM) mean values for 20 frequency bands
137	SFM mean value (averaged for all frequency bands)
138–157	Spectral flatness measure (SFM) variance values for 20 frequency bands
158	Mean SFM variance parameters
159–160	Audio spectrum centroid (mean value and variance)
161–162	Audio spectrum spread (mean value and variance)
163–164	163-164 Audio Spectrum Skewness (mean value and variance)
165–166	Audio spectrum kurtosis (mean value and variance)
167–168	Spectral entropy (mean value and variance)
169–170	Spectrum rolloff (mean value and variance)
171–172	Spectral brightness (mean value and variance)
173–192	20 Mel-frequency cepstral coefficients (mean values)
193–212	20 Mel-frequency cepstral coefficients (variance)
213–232	20 Modified mel-frequency cepstral coefficients (mean values)

Received: 16 July 2018
Revised: 17 October 2018
Re-revised: 9 November 2018
Accepted: 29 November 2018