Highlighting interlanguage phoneme differences based on similarity matrices and convolutional neural network

Gražina Korvel, Povilas Treigys, and Bożena Kostek

Citation: The Journal of the Acoustical Society of America **149**, 508 (2021); doi: 10.1121/10.0003339
View online: https://doi.org/10.1121/10.0003339
View Table of Contents: https://asa.scitation.org/toc/jas/149/1
Published by the Acoustical Society of America

JASA ARTICLE

Highlighting interlanguage phoneme differences based on similarity matrices and convolutional neural network[a)]

Gražina Korvel,[1,b)] Povilas Treigys,[1] and Bożena Kostek[2]
[1]*Institute of Data Science and Digital Technologies, Vilnius University, Vilnius, Lithuania*
[2]*Audio Acoustics Laboratory, Faculty of Electronics, Telecommunications and Informatics, Gdańsk University of Technology, Gdańsk, Poland*

ABSTRACT:
The goal of this research is to find a way of highlighting the acoustic differences between consonant phonemes of the Polish and Lithuanian languages. For this purpose, similarity matrices are employed based on speech acoustic parameters combined with a convolutional neural network (CNN). In the first experiment, we compare the effectiveness of the similarity matrices applied to discerning acoustic differences between consonant phonemes of the Polish and Lithuanian languages. The similarity matrices built on both an extensive set of parameters and a reduced set after removing high-correlated parameters are used. The results show that higher accuracy is obtained by the similarity matrices without discarding high-correlated parameters. In the second experiment, the averaged accuracies of the similarity matrices obtained are compared with the results provided by spectrograms combined with CNN, as well as the results of the vectors containing acoustic parameters and two baseline classifiers, namely *k*-nearest neighbors and support vector machine. The performance of the similarity matrix approach demonstrates its superiority over the methods used for comparison.
 https://doi.org/10.1121/10.0003339
(Received 28 May 2020; revised 19 December 2020; accepted 22 December 2020; published online 22 January 2021)
[Editor: James F. Lynch] Pages: 508–523

I. INTRODUCTION

The use of modern communication technologies brings with it new challenges. In particular, the need for interactive voice systems with an inter-language environment is growing. Differences within languages, as well as the lack of multilingual datasets, make the progress of this day below expectations. Researchers are focusing much effort on overcoming the language barrier, such as the development of methods to share acoustic models between languages (Byrne *et al.*, 2000), automatic language identification (Pellegrino and André-Obrecht, 2000; Gelly and Gauvain, 2017), recognition of the similarities/differences between languages, e.g., the variability of foreign-accented speech of second language learners (Xie and Jaeger, 2020) or in expressing basic human emotions (Fu *et al.*, 2020; Vryzas *et al.*, 2020), conducting a cross-dialectal acoustic study (Schoormann *et al.*, 2017). However, little attention (if any) has so far been paid to differences between Polish and Lithuanian speech acoustical properties even though there are bilingual Lithuanian and Polish speakers learning both languages in early childhood. In this research, we will focus on acoustic analysis of Lithuanian and Polish languages originating from the same family of Indo-European languages but belonging to two groups of languages. This will apply to Polish from the Lechic group, the Slavic subfamily of Slavic languages, and Lithuanian belonging to the Baltic (or more precisely, Eastern Baltic) languages. However, due to the similarities in terms of grammar, the Baltic languages are included with the Slavic languages in one Balto-Slavic subfamily, so that is why the models will also be tested in the context of yet another language, i.e., English.

The paradigm of intelligent speech processing consisting in pre-processing, speech parameterization, and applying a conventional machine-learning algorithm starts to become obsolete in the deep learning era (Lauriola *et al.*, 2020; Strisciuglio *et al.*, 2019; Grozdić *et al.*, 2017). It should be noted that the aim of the signal parameterization is twofold, i.e., to create a feature vector, a compressed signal representation, and to find acoustic parameters that will describe minute nuances of speech signal structure in a meaningful way. In contrast, deep learning methods require extensive data collection, thus there is no longer a need to compress the data. Contrarily, two-dimensional (2D) maps are often presented at the algorithm input to augment data. Moreover, neural networks (NN) endeavor to retrieve features from the raw signal or relatively unprocessed data (Bhatt *et al.*, 2019; Bianco *et al.*, 2019) and may effectively use them in speech classification, recognition, synthesis, and other speech-processing-related tasks.

However, the acoustic parameters of the speech signal are still widely explored in the area of speech and audio

[a)]This paper should be part of the special issue on Machine Learning in Acoustics.
[b)]Electronic mail: grazina.korvel@mif.vu.lt, ORCID: 0000-0002-1931-6852.

https://doi.org/10.1121/10.0003339

processing. These parameters are time- and frequency-domain descriptors. Time-related parameters are pitch, formants, loudness, and energy. In the frequency domain, parameters are most often derived from the Fourier transform and describe the shape of the spectrum. Most often, the acoustic parameters are employed for comparison in speech recognition (Maučec and Žgank, 2011; Cho and Park, 2016; Menne *et al.*, 2018), speech emotion classification (Ntalampiras, 2020; Vrysis *et al.*, 2020), but also in tasks such as speaker diarization (Zewoudie *et al.*, 2018; Jothilakshmi *et al.*, 2009), automatic speaker verification (Sarkar *et al.*, 2014), multi-modal speech recognition (Czyzewski *et al.*, 2017; Noulas *et al.*, 2012), and the detection of Parkinson's disease (Braga *et al.*, 2019; Yaman *et al.*, 2020). To evaluate to what extent visual data can enhance recognition accuracy in the multi-modal approach, Cygert and his collaborators (Cygert *et al.*, 2018) used the standard Mel-Frequency Cepstrum Coefficients (MFCCs) for sound parameterization. MFCC is the most dominant method in the speech processing area (Rejaibi *et al.*, 2019; Tüske *et al.*, 2014) and is also used in this research (see Table II). In the analysis of English intonation, Hirst explored a set of acoustic parameters that vary as a function of time to be the output of the phonetic model (Hirst, 2018). The quality of allophone pronunciation for non-native English speakers was evaluated by utilizing a set of acoustic descriptors commonly employed in music information retrieval (Kostek *et al.*, 2017).

In our previous research (Korvel *et al.*, 2019a), we created and tested a set of acoustic parameters related to the differences between Polish and Lithuanian consonants. Two baseline classifiers, namely *k*-nearest neighbors (*k*-NN) and support vector machine (SVM), were used to test the effectiveness of the extracted parameters in the classification process. The obtained high accuracy of the classification indicates that the proposed parameters are useful in determining the interlanguage differences. In this article, based on our previous results and designed experiments employing deep learning algorithms, we propose a new way to explore the differences between languages based on extracted consonants. This approach combines a convolutional neural network (CNN) and acoustic parameters converted into images representing a self-similarity matrix.

CNN is a class of deep, feed-forward artificial neural networks. Convolution operations are employed on the input data to produce higher-level representations. The choice of this type of network was made upon the fact that CNN performs well in an image processing context (Koller *et al.*, 2016; Korvel *et al.*, 2018). This was already proved when utilizing 2D speech representations such as spectrograms, cepstrograms, mel-cespstrograms, chromagrams, etc., as input to CNN (Korvel *et al.*, 2018; Vryzas *et al.*, 2020).

Foote (1999) and later Foote and Cooper (2001) defined acoustic similarity as repeating or similar elements that may be discerned in an audio recording or across recordings and visualized as a 2D representation. Their proposal was to identify music structure characteristic for a given music excerpt. It should be pointed out that in the literature, both "self-similarity" and "similarity" matrix terms are used interchangeably, thus both terms are to be found throughout the text.

Thus, the idea of a self-similarity matrix was borrowed by us from the music information retrieval (MIR) domain (Foote *et al.*, 2002). Foote *et al.* (2002) used the self-similarity method to characterize the rhythm and tempo of music. In the approach of Foote *et al.* (2002), a matrix is constructed based on dividing the sound signal into windows and calculating the similarity between each of them. The resulting image is a visualization of the sound structure and is useful when comparing music pieces according to their musical information. In the case of the analysis of short speech signals, we are more focused on examining the acoustic differences, so our resulting images are derived from the acoustic characteristics of the speech signals (for example, the location of the formant frequencies, voicing, articulation).

To our knowledge, self-similarity matrices created upon acoustic features were not used for speech signal classification before. However, it should be recalled that such an approach was recently utilized in the speech area, for example, in the context of visualization of pseudonymization performance for speech signal (Noé *et al.*, 2020) or exploitation of self-similarity matrix matching technique to estimate the speech intelligibility of cleft lip and palate (Kalita *et al.*, 2018).

The objective of this research is the exploration of consonant phonemes in the context of inter-language differences. It is well known that vowels carry on an essential part of the speech and language characteristics (Pellegrino and André-Obrecht, 2000). In contrast, consonants contain language-related information and are necessary to identify lexical meaning (Nespor *et al.*, 2003). That is why it is important to discern and highlight differences resulting from language specifics. For comparison of results, we created similarity matrices as well as spectrograms derived from the same dataset. That is because spectrograms are extensively investigated in tasks related to speech processing (Satt *et al.*, 2017; Rafaely and Alhaiany, 2018). Moreover, in the authors' previous research concerning the use of deep learning networks applied to speech recognition, spectrograms brought the highest accuracies (Korvel *et al.*, 2018).

The structure of the paper is as follows. In Sec. II, we describe the datasets utilized in experiments. Section III shows the parameters that were utilized for the evaluation of the interlanguage differences used in our previous study; the construction of the self-similarity matrices based on acoustic parameters is outlined in the third section as well. Section IV depicts building issues of a feature space based on spectrograms. A convolutional neural network with multiple layers, employed as a machine learning algorithm, along with its architecture, is described in Sec. V. The experimental results are reported in Sec. VI. Finally, concluding remarks are contained in Sec. VII, outlining further research plans.

II. DATA AND THEIR CHARACTERISTICS

The basic unit of text is grapheme. The phoneme, construing the basic unit of phonology, represents the smallest unit of speech sound that may cause a change of meaning within a language, but that does not have the meaning of its own. Lithuanian language consists of 20 consonant graphemes that are the written equivalent of phonemes. The Polish language includes 23 ones. Because the uttered signal is represented by phonemes, in most studies, grapheme to phoneme conversion is performed. The Lithuanian and Polish languages share many of the same consonant phonemes. Despite this, these shared phonemes may have different articulation. The aim of this research study is to detect acoustic differences between the consonant phonemes of the Polish and Lithuanian languages.

A. Participant methods

The recordings encompass speech of eight native speakers (four females and four males) with the same dialects within a language. Most of the speakers were monolingual, who learned foreign languages at school. It should be noted that no speakers were using both Polish and Lithuanian. The subjects were aged from 21 to 45 years. All the recordings were acquired in a controlled acoustic environment (a professional recording studio). The sentences were not the same across languages, but when selecting the sentences, intonation was taken into account. Alongside word stress and rhythm, intonation is a crucial element of linguistic prosody. That is why the recording scenario included read sentences with different prosody (indicative, imperative, and questioning utterances). The recording scenario included read sentences with different prosody (indicative, imperative, and questioning utterances). The consonants reside in various positions: at the beginning, in the middle, or at the end of a word. Efforts were made to select such sentences that the consonant phonemes would be found in different positions of words approximately in the same proportion, disregarding the contextual variant of a particular phoneme. The phonemes were extracted manually, thus this tedious process limited creating a larger set for the analysis. In the case of the Polish language, the speakers were asked to read 25 sentences. For Lithuanian, we consider 32 sentences to ensure the maximum possible number of samples for lesser-occurring consonants and to maintain the same proportion of Polish and Lithuanian consonants. The audio files were recorded with a 22 kHz/16-bit resolution.

The English language was tested across Polish and Lithuanian in the experiments performed. For this purpose, the TIMIT Acoustic-Phonetic Continuous Speech Corpus was used (Garofolo *et al.*, 1993). This corpus contains read speech recordings of 630 speakers of eight major dialects of American English, each reading ten sentences. In our research study, recordings of a dialect named New York City was used. We chose recordings with only one dialect in order to avoid differences between dialects in English recordings. This dialect region was selected randomly. Concerning English, we took the same ratio of females and males recordings, as in the case of Polish and Lithuanian.

B. Experimental setup

The experiments were based on consonant phonemes extracted from the recordings. The annotation was conducted manually. In the experiments, only phonemes that appear in Polish and Lithuanian languages were considered. The list of phonemes used in the experiments is given in Table I.

The number of samples for each phoneme was different, while for each phoneme, the same number of samples was taken from each language (see Table I). As a result, 1341 phoneme transcribed were extracted for each language separately. The block diagram of the experimental setup is presented in Fig. 1.

As we see from Fig. 1, the accuracy of the CNN method is evaluated by performing statistical analysis of results and comparing results with the two baseline methods, i.e., k-NN and SVM. These methods employes the same extracted speech parameters as in the case of CNN.

Deep learning methods require a sufficiently large amount of training data. The standard image augmentation techniques (e.g., rotation, translation, scaling, or horizontal shearing) cannot be applied to speech recordings because geometric linear transformations cause loss of time orientation.

Even though other techniques may be employed in audio signal augmentation for deep neural networks, such as time-shifting, pitch shifting, dynamic range compression, or background noise (Koszewski and Kostek, 2020; Pereira *et al.*, 2020; Salamon and Bello, 2017; Vryzas *et al.*, 2020), they also may influence the acoustic parameters. In our future experiments, we would like to expand data (in the context of scalability of the method proposed), focusing specifically on generating synthetic parameter sequence and

TABLE I. Consonant phoneme used in the experiments.

Type of consonant	SAMPA symbol	IPA symbol	Number of samples for each language
Plosives	/p/	/p/	54
	/t/	/t/	92
	/d/	/d/	21
	/k/	/k/	122
	/g/	/g/	28
Affricates	/tS/	/tʂ/	25
Fricatives	/f/	/f/	52
	/v/	/v/	49
	/s/	/s/	86
	/z/	/z/	17
	/S/	/ʂ/	83
	/Z/	/ʐ/	17
Nasal	/m/	/m/	78
	/n/	/n/	223
Liquids	/r/	/r/	203
	/l/	/l/	143
Glides	/j/	/j/	48

https://doi.org/10.1121/10.0003339

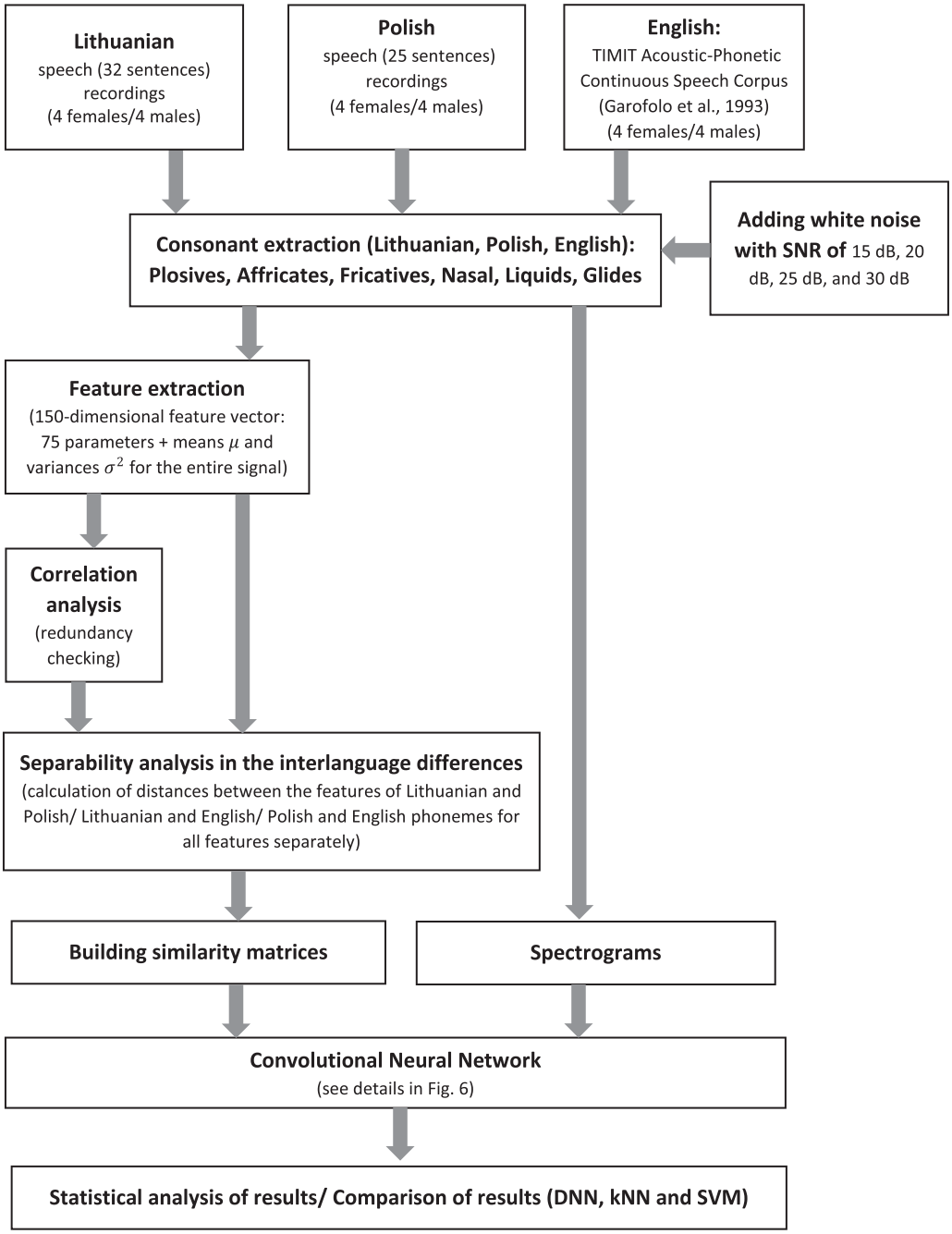

FIG. 1. Block diagram of the experimental setup.

creating upon them self-similarity matrices to augment the amount of data.

Therefore, in our research, the dataset was extended by adding a noise signal. To augment the datasets, we added white noise to the speech signal with the following levels: 15, 20, 25, and 30 dB. As a result, the initial dataset was extended five times.

III. INPUT FEATURES BASED ON ACOUSTIC PARAMETERS

In the proposed speech signal representation, phonemes are first described using acoustic parameters. Based on these parameters, the self-similarity matrices are created and introduced as 2D space features at the CNN input.

A. Speech parameterization

Due to the different nature of the consonant signal, a large set of parameters should be considered in order to extract the differences between languages. In this study, the set of parameters employed is the same as in our previous study in which the comparison of Lithuanian and Polish consonants was performed (Korvel *et al.*, 2019a). These parameters are standard MPEG 7 speech and audio signal descriptors (Kim *et al.*, 2006) along with those from the

TABLE II. The acoustic parameters for evaluation of interlanguage differences.

No.	Abbreviation	Description
1	rms	Root-mean-square (rms) energy
2	TC	Temporal Centroid
3	ZCR	Zero-Crossing Rate
4-6	k_1, k_2, k_3	The number of samples exceeding levels rms, 2× rms, 3×rms
7	Peak to rms	Peak to rms calculated as the mean value of the ratio calculated in 10 sub-frames
8–11	p_1, p_2, p_3, p_4	The mean values of signal crossings in relation to zero, rms, 2×rms, 3×rms averaged for 10 sub-segments
12–15	q_1, q_2, q_3, q_4	The variance values of signal crossings in relation to zero, rms, 2×rms, 3×rms averaged for 10 sub-segments
16	ASC	Audio spectral centroid
17	ASSp	Audio spectral spread
18	ASSk	Audio spectral skewness
19	ASK	Audio spectral kurtosis
20	Entropy	Spectral entropy
21	RollOff	Spectral roll-off
22	Brightness	Spectral brightness
23–51	ASE1–ASE29	Audio spectrum envelope calculated on 29 sub-bands
52–55	F1–F4	The first four formants
56–75	MFCC1-MFCC20	Mel-Frequency Cepstral Coefficients

music area (Kostek *et al.*, 2011). The set of the analyzed parameters is given in Table II.

The parameters are calculated on short-time segments of the speech signal. For this purpose, the speech signal is divided into segments with a length of 512 samples (i.e., 23 ms). For each frame, the Hamming window was applied. An overlap between contiguous segments was 256 samples (50%). In total, there are 75 separate parameters (see Table II) for each short-time segments of the speech signal. Moreover, statistical characteristics, i.e., means (μ) and variances (σ^2), are calculated for all short-term segments for the entire speech signal, resulting in a 150-dimensional vector of parameters is obtained. All parameters contained in the vectors get normalized to the interval [0 1].

The experiments were performed on the whole set (see Table II), as well as on the optimized parameter set. The optimized parameter set in the context of a particular consonant (see the the Appendix), aimed at distinguishing interlanguage differences, was prepared in our previous research (Korvel *et al.*, 2019a). The parameter reduction consisted of rejecting high-correlated parameters. For this purpose, the matrix of correlation coefficients was calculated. The parameters for which correlation coefficients are larger than 0.75 are discarded (the correlation coefficient was set by carrying out initial experiments). Then, the remaining parameters are used for the separability analysis in the interlanguage differences recognition process. For this purpose, the distances between the parameters of Lithuanian and Polish phonemes are calculated. Therefore, the last step in vector optimization rejects parameters that have the smallest differences of the averaged values between the parameters of the different languages (Korvel *et al.*, 2019a).

A detailed description of the selected parameters, as well as the formulas for their calculation, can be found in the authors' publications (Korvel *et al.*, 2019a; Korvel *et al.*, 2019b; Kostek *et al.*, 2011).

B. Similarity matrix construction

After the consonant phonemes were parameterized, they were then represented in a 2D form, called a similarity matrix. The matrix was constructed from the pairwise distances between parameters. The Euclidean distance method is used for this purpose. Let $\boldsymbol{p}_i = (p_{i1}, p_{i2}, \ldots, p_{iN})$ and $\boldsymbol{p}_j = (p_{j1}, p_{j2}, \ldots, p_{jN})$ be two parameters calculated on N short-time intervals, $i, j \in [1, M]$, M denotes the number of parameters (in this study $M = 150$, i.e., means and variances calculated for 75 separate parameters given in Table II). The Euclidean distance between these parameters is calculated by the following formula:

$$d(\boldsymbol{p}_i, \boldsymbol{p}_j) = \sqrt{\sum_{n=1}^{N} (p_{in} - p_{jn})^2}. \quad (1)$$

The similarity matrix represents the Euclidean distances [see Eq. (1)] between all possible acoustic parameters. A graphical representation of the matrix structure, which results from the location of the distance measures in a 2D representation, is shown in Fig. 2.

Each pixel in the matrix obtains a greyscale value corresponding to the distance between parameter pairs. The

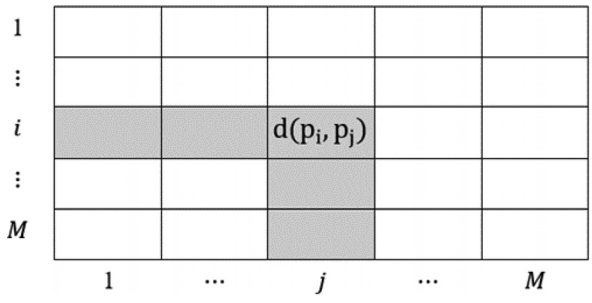

FIG. 2. Visualization of the similarity matrix construction.

https://doi.org/10.1121/10.0003339

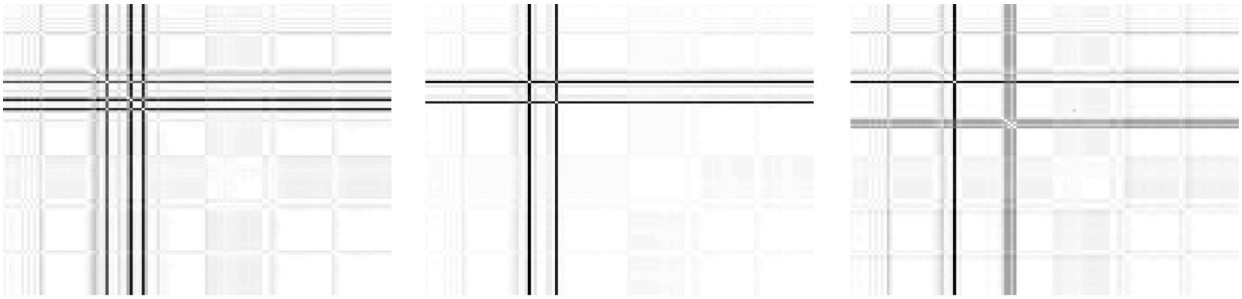

FIG. 3. The similarity matrices of the phoneme /l/ uttered by (a) Lithuanian female, (b) Polish female, and (c) English female calculated based on parameters given in Table II.

darkest color refers to the smallest distance. Basically, the diagonal shows the darkest shade because each parameter is as similar as possible conceptually.

Figures 3 and 4 show examples of similarity matrices of the phoneme /l/ calculated based on the parameters given in Table II and the Appendix. To enhance the visibility of the printed images, we inverted the colors in the examples shown in Figs. 3 and 4. However, the original images with unreversed colors were used as input data for the network. The darker the point is in the inverted color visualization, the more similar are instances i and j. Similar regions are dark, while dissimilar ones are light. Repetitive similarities are seen as checkboard patterns (Foote, 1999).

By visually exploring the differences, one may say that the phoneme /l/ uttered by Lithuanian, Polish, and English females show interlanguage differences.

Figure 3 reveals the largest similarity values related to Audio Spectrum Envelope parameters, namely ASE13 and ASE21 for both Lithuanian and Polish phonemes /l/ and ASE13 for English phoneme /l/. In the case of the English language, high similarity values are found in relation to $F2$ and $F4$, i.e., second and fourth formants.

In the case of the example given in Fig. 4, parameters such as μ(ASE7), μ(Peak to rms), σ^2(MFCC6), and σ^2(MFCC13) are strongly related to Polish and Lithuanian phonemes /l/, while for an English phoneme /l/, large similarity values are obtained for μ(F4) and μ(MFCC2) descriptors.

IV. FEATURE SPACE BASED ON SPECTROGRAMS

As an alternative method of consonant phoneme representation in the 2D space, we chose a spectrogram. The reason why we decided to use this type of representation is the effectiveness of spectrograms in solving signal processing tasks (Li *et al.*, 2018; Korvel *et al.*, 2018; Yenigalla *et al.*, 2018). As in the case of the similarity matrix, a spectrogram is also constructed from a series of short-time segments. In order to obtain a spectrogram, the speech segment is converted to the frequency domain, and the log-spectra are calculated according to the formula

$$Sk(n) = \log(Xk(n)), \tag{2}$$

where $Xk(n)$ $(n = 1, \ldots, N)$ are Fourier transform coefficients of frame k, and N is the number of these coefficients.

An example of the log-spectra spectrogram [see Eq. (2)] of the phoneme /l/ is given in Fig. 5.

V. CONVOLUTIONAL NEURAL NETWORK

In the previous two sections, we described speech signal converting techniques into images (similarity matrices and spectrograms), i.e., obtained a new representation of audio data. Such data conversion enables the use of algorithms whose direct application area is image analysis and processing. Over the past few years, machine learning algorithms—especially artificial neural networks—outperform the classical,

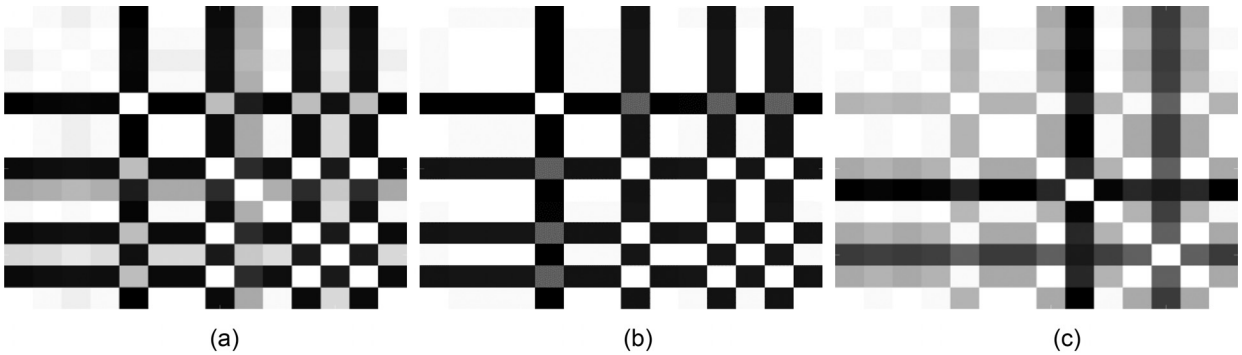

FIG. 4. The similarity matrices of the phoneme /l/ uttered by (a) Lithuanian female, (b) Polish female, and (c) English female calculated based on parameters contained in the Appendix.

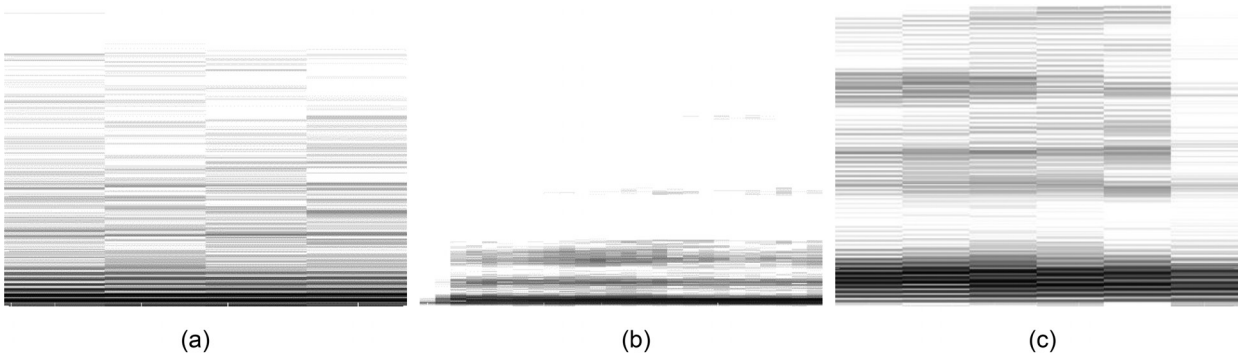

FIG. 5. The spectrogram of the phoneme /l/ uttered by (a) Lithuanian female, (b) Polish female, and (c) English female (Hamming window, fast Fourier transform (FFT) size 512, overlap 256, sampling rate 22 kHz).

baseline algorithms in a variety of areas. By obtaining the phoneme similarity matrix, the interlanguage differences can be investigated in the spatial data domain while solving classification tasks by application of a convolutional neural network. A convolutional neural network is a state-of-the-art algorithm and a dominant approach (Gu *et al.*, 2018), especially in image processing tasks. To classify the 2D representations prepared, we applied a simple CNN made up of multiple layers. Each layer is convolved with the randomly initialized convolution kernel, followed by the max-pooling operation. Max-pooling decreases the number of features passed to the next layer, retaining only the most robust and important ones. Then the receiving layer operation is the same. The scheme can be repeated arbitrary times until the flattening takes place. Flattening is the reorganization of 2D representation into a vector. Then the last output layer typically acts as a classifier learning of a possibly non-linear classification function. A CNN architecture comprised three convolutional layers and two dense layers. The first convolutional layer aggregated 32 filters, the subsequent two, 64. The first dense layer was composed of 64 units, and the last one was a compound of two units acting as a classier. The pool kernel size was the same as the stride kernel size, i.e., 3×3. A graphical representation of the CNN architecture is given in Fig. 6, where the numbers in brackets indicate the number of used filters and stride size, respectively.

The CNN architecture, as well as other network parameters, were selected by the careful calibration procedure. The rectified linear unit (ReLu) layer, which acts as an activation function, is used in the convolutional layers and the first dense layer. Meanwhile, the softmax function was implemented through the second dense layer to normalize the output to a probability distribution over predicted output classes. To prevent network overfitting, batch normalization has been employed after every layer change.

In our work, similarity matrix images are scaled to the size 512×256 and used as network input. All images were resized using the nearest neighbor method. The training process of the network is controlled by assessing the classification rate on the validation set. Training is stopped when an error in the validation set has not been improved for ten epochs. Binary cross-entropy is used as a loss function. For cross-entropy minimization, the Adam method, introduced by Kingma and Ba (2014), is applied. The performance of neural network classifiers is determined by a number of hyperparameters (settings) such as weight initialization, momentum, batch size, and learning rate. They were set by careful parameter calibration procedure to achieve the best learning rate accuracy. We initialized the network settings from a baseline network, using, e.g., the same learning rate as in our earlier research. The process of learning the appropriate settings for the experiment designed continued up to the moment when such hyperparameters were found that allow achieving the best learning rate accuracy. The learning rate was set to 0.0001, and the $\beta 1$ and $\beta 2$ parameters were set to 0.95 and 0.999, respectively. Calculations were performed using the Python programming language and the

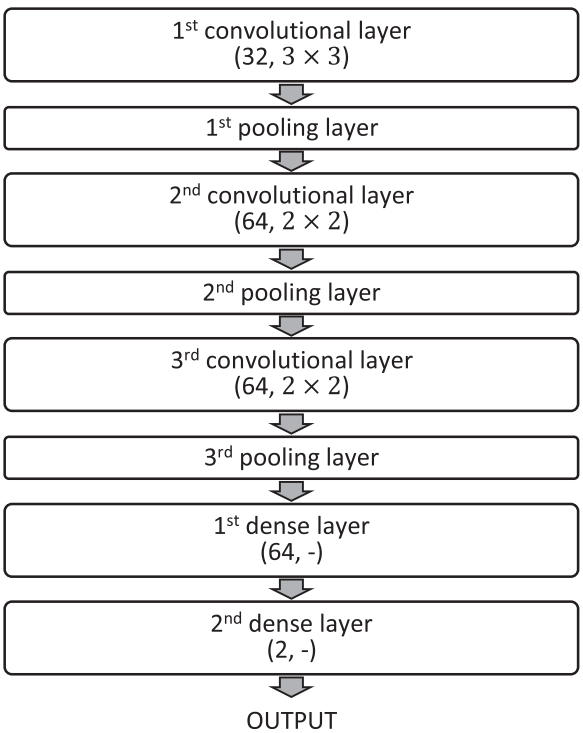

FIG. 6. Graphical representation of the CNN architecture (numbers in brackets indicate the number of used filters and stride sizes, respectively).

https://doi.org/10.1121/10.0003339

Keras Python deep learning library with the TensorFlow library (TensorFlow library, Keras library) on a GeForce RTX 2080 GPU card.

VI. EXPERIMENT RESULTS

This article reports the results of two experiments that compared the effectiveness of similarity matrices in the context of automatic phoneme classification.

A. Experiment 1: The comparison of classification results obtained by matrices build on an extensive set of parameters and those with discarding high-correlated parameters

In the first experiment, similarity matrices calculated based on the parameters given in Table II as well as those calculated based on the parameters contained in the Appendix, were constructed, and CNN was applied. The effectiveness of similarity matrices was tested for each phoneme class separately, with the following scenarios being considered here:

(A) Samples of the Lithuanian and Polish languages
(B) Samples of the Lithuanian and English languages
(C) Samples of the Polish and English languages

The obtained results averaged for all speakers are presented in Table III, where the overall accuracy (Acc.), validation accuracy (Val. acc.), and three class-specific measures, namely, class precision (Precision), class recall (Recall), and $F1$-measure, are given.

The highest accuracies presented in Table III are highlighted in bold. This experiment does not present a comparison of the results of the classification by gender due to space-saving in the paper. The obtained classification accuracy for men and women separately will be shown in Experiment 2, where the comparison of classification results obtained by different parameters and classifiers is carried out.

To evaluate the classification performance, it is necessary to analyze the obtained accuracies of Polish and Lithuanian phonemes because the set of parameters was constructed based on the acoustic characteristics of these two languages. As we can see from the results given in

TABLE III. Results of CNN classification based on the acoustic parameters.

	Similarity matrix based on parameters given in Table II					Similarity matrix based on parameters contained in the Appendix				
	Acc.	Val. acc	Precision	Recall	$F1$	Acc.	Val. acc	Precision	Recall	$F1$
Phoneme /p/										
A	**0.9971**	1	1	1	1	**0.9971**	1	1	1	1
B	**0.9932**	1	1	1	1	0.9829	1	1	1	1
C	**1**	1	1	1	1	**1**	1	1	1	1
Phoneme /t/										
A	**1**	1	1	1	1	0.9983	1	1	1	1
B	**0.9854**	1	1	1	1	**0.9854**	1	1	1	1
C	0.9895	1	1	1	1	**0.9937**	1	1	1	1
Phoneme /d/										
A	**1**	1	1	1	1	0.9931	1	1	1	1
B	**1**	1	1	1	1	0.9574	1	1	1	1
C	0.9149	1	1	1	1	**1**	1	1	1	1
Phoneme /k/										
A	**0.9910**	1	1	1	1	0.9770	0.9948	0.9800	0.9800	0.9800
B	**0.9929**	1	1	1	1	0.9887	0.9714	0.9700	0.9600	0.9600
C	**0.9986**	1	1	1	1	0.9929	1	1	1	1
Phoneme /g/										
A	**1**	1	1	1	1	0.9500	0.8636	0.9400	0.9300	0.9300
B	**0.9340**	1	1	1	1	0.9151	1	1	1	1
C	**0.8774**	1	1	1	1	0.7925	0.9615	0.9700	0.9700	0.9700
Phoneme /tS/										
A	**1**	1	1	1	1	0.9881	1	1	1	1
B	**1**	1	1	1	1	0.9935	0.9474	0.9600	0.9600	0.9600
C	**0.9935**	1	1	1	1	0.9870	1	1	1	1
Phoneme /f/ (52)										
A	**0.9970**	1	1	1	1	0.9850	1	1	1	1
B	**1**	1	1	1	1	**1**	0.9886	0.9900	0.9900	0.9900
C	**1**	1	1	1	1	0.9907	1	1	1	1
Phoneme /v/										
A	**0.9841**	1	1	1	1	0.9236	0.8974	0.9400	0.9300	0.9300
B	**0.9677**	1	1	1	1	0.9283	0.9565	0.9500	0.9500	0.9500
C	**0.9892**	1	1	1	1	0.9785	0.9420	0.9200	0.9100	0.9100

TABLE III. (*Continued*)

	Similarity matrix based on parameters given in Table II					Similarity matrix based on parameters contained in the Appendix				
	Acc.	Val. acc	Precision	Recall	*F*1	Acc.	Val. acc	Precision	Recall	*F*1
Phoneme /s/										
A	**0.9964**	1	1	1	1	**0.9964**	1	1	1	1
B	**0.9781**	0.8542	0.8900	0.8600	0.8600	**0.9781**	0.9931	0.9900	0.9900	0.9900
C	**0.9964**	1	1	1	1	0.9927	0.9861	0.9800	0.9800	0.9800
Phoneme /z/										
A	**1**	1	1	1	1	**1**	1	1	1	1
B	**1**	1	1	1	1	**1**	1	1	1	1
C	**1**	1	1	1	1	0.9130	1	1	1	1
Phoneme /S/										
A	**0.9944**	1	1	1	1	0.9925	1	1	1	1
B	0.9699	0.9848	1	1	1	**0.9831**	1	1	1	1
C	**1**	1	1	1	1	0.9737	0.9924	0.9900	0.9900	0.9900
Phoneme /Z/										
A	**1**	1	1	1	1	**1**	1	1	1	1
B	**1**	1	1	1	1	**1**	1	1	1	1
C	**1**	1	1	1	1	**1**	1	1	1	1
Phoneme /m/										
A	**0.9980**	1	1	1	1	0.9340	0.6371	0.8000	0.6700	0.6300
B	**0.9880**	1	1	1	1	0.9440	0.9194	0.9500	0.9500	0.9500
C	**0.9940**	1	1	1	1	0.9700	0.8387	0.8800	0.8400	0.8400
Phoneme /n/										
A	**0.9986**	1	1	1	1	0.9034	0.7697	0.8100	0.7100	0.6800
B	**0.9901**	1	1	1	1	0.9022	0.9208	0.9400	0.9300	0.9300
C	**0.9975**	1	1	1	1	0.9466	0.9736	0.9800	0.9800	0.9800
Phoneme /r/										
A	**0.9934**	1	1	1	1	0.9635	0.9867	0.9900	0.9900	0.9900
B	**0.9921**	1	1	1	1	0.9693	0.9920	0.9800	0.9800	0.9800
C	**0.9982**	1	1	1	1	0.9746	1	0.9900	0.9900	0.9900
Phoneme /l/										
A	**0.9956**	1	1	1	1	0.8479	0.8622	0.9100	0.9000	0.9000
B	**0.9920**	1	1	1	1	0.9109	0.9078	0.9100	0.9000	0.9000
C	**0.9966**	1	1	1	1	0.9704	0.9495	0.9600	0.9600	0.9600
Phoneme /j/										
A	**0.9870**	1	1	1	1	0.9416	0.9211	0.9600	0.9600	0.9600
B	**1**	1	1	1	1	0.9772	0.9692	0.9800	0.9800	0.9800
C	**0.9848**	1	1	1	1	0.9810	0.9692	0.9800	0.9800	0.9800

Table III, in all comparisons, higher or equal accuracies are obtained based on the self-similarity method without parameter reduction (i.e., for the similarity matrix based on parameters given in Table II). From this, we can conclude that the deep learning network picks the most useful parameters itself. The reduction of the input dimension is not needed beforehand, as opposed to supervised machine learning methods, where input parameters are chosen manually.

The same trend is observed when comparing the classification results of Polish and English phonemes and Lithuanian and Polish ones (see Table III). But here are some exceptions, i.e., phoneme /s/ in the case of the Lithuanian and English languages and phonemes /t/ and /d/ in the case of the Polish and English languages. These exceptions allow us to conclude that the acoustic parameters established for Polish and Lithuanian are not entirely useful for assessing the differences between these languages and the language, which was not used to develop a set of acoustic parameters. In other words, the results showed that the established acoustic parameters are language-dependent.

B. Experiment 2: The comparison of classification results obtained by different parameters and classifiers

In this experiment, we compared the obtained averaged accuracies of similarity matrices with results provided by spectrograms and CNN, as well as the findings from our previous experiments (Korvel *et al.*, 2019a). These latter ones were obtained using the parameters contained in the Appendix and employing two classifiers: *k*-NN and SVM. The classification accuracies are given in Table IV.

The highest accuracies (see Table IV) are highlighted in bold. One of the results obtained, which should be discussed, is the phoneme /n/. Apparently, looking at the results presented in Table IV, it may seem that the self-similarity method

https://doi.org/10.1121/10.0003339

TABLE. IV. Classification performance of *k*-NN, SVM, and CNN methods in Polish, Lithuanian, and English phoneme classification using acoustic parameters, similarity matrices based on acoustic parameters, and spectrograms.

		k-NN	SVM	CNN		
		Parameters contained in the Appendix	Parameters contained in the Appendix	Similarity matrix based on parameters given in Table II	Similarity matrix based on parameters contained in the Appendix	Spectrograms
Phoneme /p/						
A	*All*	0.9780	0.9480	**0.9971**	**0.9971**	0.9706
	Female	0.9750	0.9750	**1**	0.9854	0.9806
	Male	0.9460	**1**	**1**	0.9861	**1**
B	*All*	0.9250	0.9930	**0.9932**	0.9829	0.9829
	Female	0.8630	0.9880	**1**	0.9667	0.9833
	Male	0.8930	0.6610	**1**	0.9912	0.9912
C	*All*	0.9850	**1**	**1**	**1**	0.9933
	Female	0.9630	0.9750	**0.9946**	0.9785	0.9839
	Male	0.9460	0.8930	0.9912	**1**	0.9912
Phoneme /t/						
A	*All*	0.9630	0.8810	**1**	0.9983	0.9949
	Female	0.9380	0.9130	**1**	0.9801	**1**
	Male	0.9640	0.9640	**0.9966**	0.9865	**0.9966**
B	*All*	0.8960	0.8660	0.9854	0.9854	**1**
	Female	0.8380	0.8880	**1**	0.9793	0.9959
	Male	0.8750	0.9460	0.9831	0.9747	**1**
C	*All*	0.9550	0.9700	0.9895	0.9937	**0.9937**
	Female	0.9500	0.9630	**1**	**1**	0.9959
	Male	0.9460	0.9640	**1**	0.9958	**1**
Phoneme /d/						
A	*All*	0.9550	0.9100	**1**	0.9931	**1**
	Female	0.9750	0.9380	0.9808	**1**	0.9615
	Male	**1**	0.9290	**1**	**1**	0.9762
B	*All*	0.8130	0.8580	**1**	0.9574	0.9681
	Female	0.6750	0.8750	0.9167	0.9167	**0.9444**
	Male	0.7680	0.7860	0.9828	0.8793	**1**
C	*All*	0.9700	0.9850	0.9149	**1**	**1**
	Female	0.9500	**0.9630**	0.9167	0.9167	0.9444
	Male	0.9460	0.9640	**1**	0.9483	0.9310
Phoneme /k/						
A	*All*	0.9550	0.9250	**0.9910**	0.9770	0.9898
	Female	0.7500	0.7380	**1**	0.9674	0.9930
	Male	0.9290	0.9460	**0.9889**	0.9694	**0.9889**
B	*All*	0.7610	0.8060	0.9929	0.9887	**0.9972**
	Female	0.6750	0.7130	**0.9974**	0.9710	0.9921
	Male	0.7680	0.8210	**0.9969**	0.9847	0.9939
C	*All*	0.9630	0.9930	**0.9986**	0.9929	0.9858
	Female	0.8250	0.8380	0.9868	**1**	0.9921
	Male	0.9110	0.9290	0.9725	**0.9969**	0.9847
Phoneme /g/						
A	*All*	0.9550	0.8880	**1**	0.9500	0.9833
	Female	0.9500	0.8500	**0.9889**	0.9222	0.9778
	Male	0.9460	0.8930	**0.9889**	0.9556	0.9778
B	*All*	0.7690	0.8810	**0.9340**	0.9151	0.8962
	Female	0.7380	0.9000	0.9636	**1**	0.9091
	Male	0.7680	**0.9820**	0.9231	0.8654	0.8654
C	*All*	0.9550	**0.9700**	0.8774	0.7925	0.9151
	Female	**0.9750**	0.9630	0.9273	0.9455	0.9273
	Male	0.8930	**0.9460**	0.7885	0.8846	0.8077
Phoneme /tS/						
A	*All*	0.9400	0.8660	**1**	0.9881	**1**
	Female	0.8380	0.7250	0.9231	**1**	**1**
	Male	0.8930	0.9110	**1**	0.9762	0.9524

TABLE. IV. (*Continued*)

		k-NN	SVM	CNN		
		Parameters contained in the Appendix	Parameters contained in the Appendix	Similarity matrix based on parameters given in Table II	Similarity matrix based on parameters contained in the Appendix	Spectrograms
B	*All*	0.7610	0.7910	**1**	0.9935	**1**
	Female	0.6130	0.7380	0.9718	0.9859	**1**
	Male	0.6790	0.7500	0.9643		0.9881
C	*All*	0.9480	0.9850	**0.9935**	0.9870	**0.9935**
	Female	0.8880	0.8750	**1**	0.9437	**1**
	Male	0.7680	0.8930	**1**	0.9643	**1**
Phoneme /f/						
A	*All*	0.8960	0.9180	0.9970	0.9850	**1**
	Female	0.8630	0.8880	**1**	0.9957	0.9871
	Male	0.8930	0.7860	**1**	**1**	**1**
B	*All*	0.7840	0.8510	**1**	**1**	0.9969
	Female	0.8250	0.7500	**0.9953**	0.9767	0.9907
	Male	0.8210	0.6790	0.9636	**1**	**1**
C	*All*	0.9400	0.9480	**1**	0.9907	0.9969
	Female	0.8250	0.8750	**1**	0.9767	**1**
	Male	0.7860	0.6250	**1**	**1**	**1**
Phoneme /v/						
A	*All*	0.9400	0.9180	0.9841	0.9236	**0.9968**
	Female	0.9880	0.9500	**1**	0.9489	**1**
	Male	0.9110	0.9460	**1**	0.9730	**1**
B	*All*	0.7760	0.8130	0.9677	0.9283	**0.9857**
	Female	0.7500	0.8000	**0.9865**	0.9662	0.9595
	Male	0.8750	0.8750	**1**	0.9318	0.9773
C	*All*	0.9400	0.9630	**0.9892**	0.9785	0.9749
	Female	0.9250	0.9500	**1**	0.9662	0.9730
	Male	0.8750	0.9110	**1**	0.9848	**1**
Phoneme /s/						
A	*All*	0.9700	0.9250	**0.9964**	**0.9964**	0.9875
	Female	0.7500	0.7000	0.9795	**1**	0.9836
	Male	0.8750	0.7500	**1**	0.9838	0.9935
B	*All*	0.9100	0.8730	0.9781	0.9781	**0.9945**
	Female	0.7380	0.8380	**0.9959**	0.9918	0.9713
	Male	0.6960	0.8930	0.9869	0.9902	0.9934
C	*All*	0.9850	0.9630	**0.9964**	0.9927	0.9945
	Female	0.8500	0.8880	0.9795	**1**	0.9836
	Male	0.8750	0.8750	0.9967	**1**	0.9967
Phoneme /z/						
A	*All*	0.9700	0.9180	**1**	**1**	**1**
	Female	0.7750	0.7630	**1**	**1**	**1**
	Male	0.9110	0.6960	**1**	**1**	**1**
B	*All*	0.8960	0.8510	**1**	**1**	**1**
	Female	0.6880	0.7130	**1**	**1**	**1**
	Male	0.6790	0.6790	**1**	**1**	**1**
C	*All*	0.9630	0.9550	**1**	0.9130	**1**
	Female	0.8500	0.9000	**1**	**1**	**1**
	Male	0.8390	0.7500	**1**	**1**	**1**
Phoneme /S/						
A	*All*	0.9780	0.9850	**0.9944**	0.9925	0.9925
	Female	0.7500	0.9000	0.9836	0.9959	**1**
	Male	0.8930	0.8040	0.9595	0.9797	**0.9899**
B	*All*	0.8960	0.8660	0.9699	0.9831	**1**
	Female	0.7130	0.7000	**0.9918**	0.9836	0.9877
	Male	0.8040	0.6070	0.9932	0.9730	**0.9966**

https://doi.org/10.1121/10.0003339

TABLE. IV. (*Continued*)

		k-NN	SVM	CNN		
		Parameters contained in the Appendix	Parameters contained in the Appendix	Similarity matrix based on parameters given in Table II	Similarity matrix based on parameters contained in the Appendix	Spectrograms
C	*All*	0.9550	0.9700	**1**	0.9737	0.9925
	Female	0.9130	0.9630	0.9959	0.9836	**1**
	Male	0.9640	0.8570	**0.9966**	0.9831	**0.9966**
Phoneme /Z/						
A	*All*	0.9780	0.9700	**1**	**1**	0.9130
	Female	0.7630	0.9130	**1**	**1**	**1**
	Male	0.8930	0.7860	**1**	**1**	**1**
B	*All*	0.9030	0.8660	**1**	**1**	**1**
	Female	0.7130	0.7630	**1**	**1**	**1**
	Male	0.8040	0.5890	**1**	**1**	0.9500
C	*All*	0.9780	0.9780	**1**	**1**	**1**
	Female	0.9000	0.9750	**1**	**1**	**1**
	Male	0.9640	0.8930	**1**	**1**	**1**
Phoneme /m/						
A	*All*	0.9780	0.9030	**0.9980**	0.9340	**0.9880**
	Female	0.9750	0.9880	**0.9963**	0.9222	0.9852
	Male	0.9460	0.9640	**1**	0.9708	0.9875
B	*All*	0.8660	0.8510	**0.9880**	0.9440	**0.9880**
	Female	0.8380	0.8630	**0.9815**	0.9444	**0.9815**
	Male	0.8390	0.9110	0.9958	0.9292	**1**
C	*All*	0.9850	0.9550	0.9940	0.9700	**0.9980**
	Female	0.9630	0.9250	**1**	0.9852	0.9926
	Male	0.9290	0.9640	**1**	0.9958	0.9917
Phoneme /n/						
A	*All*	0.9780	**1**	0.9986	0.9034	0.9881
	Female	0.9500	0.9750	0.9847	0.8903	**0.9894**
	Male	0.9820	**1**	**1**	0.9525	0.9881
B	*All*	**1**	**1**	0.9901	0.9022	0.9910
	Female	**1**	**1**	0.9848	0.9351	0.9876
	Male	0.9820	**1**	0.9980	0.9372	0.9980
C	*All*	**1**	**1**	0.9975	0.9466	0.9836
	Female	0.9880	**1**	0.9959	0.9461	0.9972
	Male	**1**	**1**	0.9939	0.9858	0.9960
Phoneme /r/						
A	*All*	0.9700	0.9780	**0.9934**	0.9635	0.9867
	Female	0.9380	0.9750	0.9904	0.9807	**0.9952**
	Male	0.9820	**1**	0.9983	0.9727	0.9796
B	*All*	0.9100	0.8810	**0.9921**	0.9693	0.9861
	Female	0.8250	0.7880	**0.9981**	0.9962	0.9962
	Male	**1**	0.9820	0.9835	0.9773	0.9690
C	*All*	0.9700	0.9850	**0.9982**	0.9746	0.9882
	Female	0.9500	0.9500	**1.0000**	0.9920	0.9904
	Male	0.9820	**1**	0.9897	0.9692	0.9733
Phoneme /l/						
A	*All*	0.9700	0.9480	0.9956	0.8479	**0.9989**
	Female	0.9290	0.9820	**0.9938**	0.8409	0.9814
	Male	0.8810	**0.9930**	0.9815	0.9699	0.9769
B	*All*	0.8500	0.9500	**0.9920**	0.9109	0.9897
	Female	0.6610	0.5360	**0.9916**	0.9241	0.9873
	Male	0.9700	**1**	0.9950	0.8953	0.9925
C	*All*	0.9500	0.9880	**0.9966**	0.9704	0.9829
	Female	0.9110	0.9460	**0.9958**	0.9707	0.9958
	Male	0.9820	**1**	**1**	0.9875	0.9975
Phoneme /j/						
A	*All*	0.9630	0.9700	0.9870	0.9416	**0.9968**
	Female	0.9880	0.9750	**0.9932**	0.9324	0.9595

TABLE. IV. (*Continued*)

		k-NN	SVM	CNN		
		Parameters contained in the Appendix	Parameters contained in the Appendix	Similarity matrix based on parameters given in Table II	Similarity matrix based on parameters contained in the Appendix	Spectrograms
	Male	0.9460	0.9820	**1**	**1**	**1**
B	*All*	0.8810	0.9630	**1**	0.9772	0.9886
	Female	0.8750	0.9500	0.9767	0.9380	**0.9922**
	Male	0.8210	0.6610	0.9926	0.9481	**1**
C	*All*	0.9780	**1**	0.9848	0.9810	**0.9696**
	Female	0.9500	0.9750	**1**	0.9612	0.9845
	Male	0.9640	0.9290	**1**	0.9778	**1**

does not work well with this phoneme. In fact, the classification performance of the phoneme /n/ is similar to the classification efficacy of other phonemes. The SVM classifier achieved the highest accuracy in almost all cases for the phoneme /n/, and it was clearly better than the remaining methods.

To provide a clearer picture of the given results, we calculated the average accuracies. Since the performance of the phonemes cannot be compared with each other due to the unequal number of samples used, we calculated the overall accuracies for all phonemes compared to gender and language. The results are given in Table V.

The accuracies contained in Tables IV and V show that in all comparisons of Polish and Lithuanian phonemes, higher classification performance is obtained on a similarity matrix built on the parameters given in Table II, i.e., the so-called "extensive" set. The same tendencies are seen in the other two scenarios (samples of the Lithuanian and English, and those of Polish and English). An interesting fact worth mentioning is that in the case of the Polish and English languages, the efficiency of classification for consonant phonemes based on spectrograms exceeds the efficiency obtained by the similarity matrices. This was brought for all utterances regardless of gender. In contrast, in the case of classification, taking into account speech samples of the females and males separately, the highest accuracies occurred when employing the similarity matrices (see Table V).

VII. CONCLUSIONS

In this research, we proposed a new method based on similarity matrices for highlighting the acoustic differences between the Lithuanian and Polish languages. In the experiments, we computed the performance scores of these methods for Lithuanian, Polish, and English consonant classification.

In the first experiment, we investigated the effectiveness of similarity matrices applied to discerning acoustic differences between consonant phonemes of the Polish and Lithuanian languages. They were built on both an extensive set of parameters and a reduced set after removing high-correlated parameters. It was found that in all Lithuanian and Polish consonant phoneme comparisons, higher accuracy is obtained by the similarity matrices without reducing the parameters. On this basis, we can conclude that the deep learning network needs more extensive data from which it automatically extracts useful information. This also indicates that the similarity matrices cover the broadest possible range of parameters.

The classification results of Polish and English phonemes and Lithuanian and Polish ones showed that the established acoustic parameters are language-dependent.

In the second experiment, the averaged accuracies of the similarity matrices obtained were compared with the results provided by spectrograms combined with a CNN, as well as with the outcomes of the vectors containing acoustic parameters and two baseline classifiers, namely *k*-NN and SVM. In the case of Polish and Lithuanian, as well as in the

TABLE. V. The overall classification performance of *k*-NN, SVM, and CNN methods in Polish, Lithuanian, and English phoneme classification using acoustic parameters, similarity matrices based on acoustic parameters, and spectrograms.

		k-NN	SVM	CNN		
		Parameters contained in the Appendix	Parameters contained in the Appendix	Similarity matrix based on parameters given in Table II	Similarity matrix based on parameters contained in the Appendix	Spectrograms
A	*All*	0.9076	0.8806	**0.9407**	0.9106	0.9326
	Female	0.8386	0.8416	**0.9341**	0.9090	0.9330
	Male	0.8773	0.8528	**0.9397**	0.9265	0.9337
B	*All*	0.8109	0.8311	**0.9324**	0.9120	0.9314
	Female	0.7238	0.7668	**0.9307**	0.9153	0.9266
	Male	0.7801	0.7679	**0.9310**	0.8987	0.9286
C	*All*	0.9122	0.9227	0.9295	0.9143	**0.9313**
	Female	0.8653	0.8847	**0.9329**	0.9203	0.9312
	Male	0.8650	0.8552	**0.9294**	0.9263	0.9259

https://doi.org/10.1121/10.0003339

case of Lithuanian and English, the highest classification accuracy was achieved for similarity matrices. Comparing Polish and English, the classification performance of males and females together based on spectrograms exceeded the performance obtained by the similarity matrices, while in the case of classification, taking into account female and male samples separately, the highest accuracy was obtained when applying the similarity matrices.

A further testing option could include tests using utterances by people for whom the given language is not their mother tongue. This may show to what extent an incorrect accent, including changes in intonation and rhythm of speech, would demonstrate the impact on the similarity method. Moreover, the same experiment may be performed for the case when the position of a consonant in a word is taken into account, i.e., especially looking at whether it is the intimal or final consonant.

Last, we would like to extend the approach based on similarity matrices and machine learning by generating synthetic feature sequences and creating upon them self-similarity matrices to augment the amount of data and check the scalability of the method.

ACKNOWLEDGMENTS

The authors are thankful for the high-performance computing resources provided by the Information Technology Open Access Center at the Faculty of Mathematics and Informatics of Vilnius University Information Technology Research Center.

APPENDIX: PARAMETERS ALLOWING FOR INTERLANGUAGE DIFFERENCE ANALYSIS (KORVEL *ET AL.* 2019A)

/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)
/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)

Bhatt, G., Jha, P., and Raman, B. (**2019**). "Representation learning using step-based deep multi-modal autoencoders," Pattern Recogn. **95**, 12–23.

Bianco, M. J., Gerstoft, P., Traer, J., Ozanich, E., Roch, M. A., Gannot, S., and Deledalle, C. A. (**2019**). "Machine learning in acoustics: Theory and applications," J. Acoust. Soc. Am. **146**(5), 3590–3628.

Braga, D., Madureira, A. M., Coelho, L., and Ajith, R. (**2019**). "Automatic detection of Parkinson's disease based on acoustic analysis of speech," Eng. Appl. Artif. Intell. **77**, 148–158.

Byrne, W., Beyerlein, P., Huerta, J. M., Khudanpur, S., Marthi, B., Morgan, J., Peterek, N., Picone, J., Vergyri, D., and Wang, T. (**2000**). "Towards language independent acoustic modeling," in *Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing*, June 5–9, Istanbul, Turkey, pp. I1029–II1032.

Cho, J. W., and Park, H.-M. (**2016**). "Independent vector analysis followed by HMM-based feature enhancement for robust speech recognition," Signal Process. **120**, 200–208.

Cygert, S., Szwoch, G., Zaporowski, S., and Czyzewski, A. (**2018**). "Vocalic segments classification assisted by mouth motion capture," in *Proceedings of the 2018 11th International Conference on Human System Interaction (HSI)*, July 4–6, Gdansk, Poland, pp. 318–324.

Czyzewski, A., Kostek, B., Bratoszewski, P., Kotus, J., and Szykulski, M. (**2017**). "An audio-visual corpus for multi-modal automatic speech recognition," J. Intell. Inf. Syst. **49**(2), 167–192.

Foote, J., Cooper, M. L., and Nam, U. (**2002**). *Audio retrieval by rhythmic similarity* (ISMIR, Canada).

Foote, J. (**1999**). "Visualizing music and audio using self-similarity," in *Proceedings of ACM Multimedia '99*, October 30–November 5, Orlando, FL, pp. 77–80.

Foote, J., and Cooper, M. L. (**2001**). "Visualizing musical structure and rhythm via self-similarity," in *Proceedings of the International Conference on Computer Music*, September 17–22, Habana, Cuba.

Fu, C., Dissanayake, T., Hosoda, K., Maekawa, T., and Ishiguro, H. (**2020**). "Similarity of speech emotion in different languages revealed by a neural network with attention," in *Proceedings of the 2020 IEEE 14th International Conference on Semantic Computing (ICSC)*, February 3–5, San Diego, CA, pp. 381–386.

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

Gelly, G., and Gauvain, J. L. (**2017**). "Spoken language identification using LSTM-based angular proximity," in *Proceedings of INTERSPEECH*, August 20–24, Stockholm, Sweden, pp. 2566–2570.

Grozdić, ĐT., Jovičić, S. T., and Subotić, M. (**2017**). "Whispered speech recognition using deep denoising autoencoder," Eng. Appl. Artif. Intell. **59**, 15–22.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., and Chen, T. (**2018**). "Recent advances in convolutional neural networks," Pattern Recogn. **77**, 354–377.

Hirst, D. (**2018**). "Phonological and acoustic parameters of English intonation," in *Intonation in Discourse* (Routledge, London), pp. 19–34.

Jothilakshmi, S., Ramalingam, V., and Palanivel, S. (**2009**). "Speaker diarization using autoassociative neural networks," Eng. Appl. Artif. Intell. **22**(4-5), 667–675.

Kalita, S., Prasanna, S. R. M., and Dandapat, S. (**2018**). "Self-similarity matrix based intelligibility assessment of cleft lip and palate speech," in *Proceedings of Interspeech 2018*, September 2–6, Hyderabad, India.

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

Kingma, D. P., and Ba, J. (**2014**). "Adam: A method for stochastic optimization," arXiv:1412.6980.

Koller, O., Ney, H., and Bowden, R. (**2016**). "Deep hand: How to train a CNN on 1 million hand images when your data is continuous and weakly labelled," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 27–30, Las Vegas, NV, pp. 3793–3802.

Korvel, G., Kurowski, A., Kostek, B., and Czyzewski, A. (**2019a**). "Speech analytics based on machine learning," in *Machine Learning Paradigms. Intelligent Systems Reference Library*, edited by G. Tsihrintzis, D. Sotiropoulos, and L. Jain (Springer International Publishing, Cham).

Korvel, G., Kurasova, O., and Kostek, B. (**2019b**). "Comparison of Lithuanian and Polish consonant phonemes based on acoustic analysis–preliminary results," Arch. Acoust. **44**(4), 693–707.

Korvel, G., Treigys, P., Tamulevicus, G., Bernataviciene, J., and Kostek, B. (**2018**). "Analysis of 2D feature spaces for deep learning-based speech recognition," J. Audio Eng. Soc. **66**(12), 1072–1081.

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

Kostek, B., Piotrowska, M., Ciszewski, T., and Czyzewski, A. (**2017**). "Comparative study of self-organizing maps vs subjective evaluation of quality of allophone pronunciation for non-native English speakers," in *Proceedings of the Audio Engineering Society Convention 143*, October 18–21, New York.

Koszewski, D., and Kostek, B. (**2020**). "Musical instrument tagging using data augmentation and effective noisy data processing," J. Audio Eng. Soc. **68**(1/2), 57–65.

Lauriola, I., Gallicchio, C., and Aiolli, F. (**2020**). "Enhancing deep neural networks via multiple kernel learning," Pattern Recogn. **101**, 107194.

Li, Y., Pi, S., and Xiao, N. (**2018**). "Speech recognition method based on spectrogram," in *Proceedings of the International Conference on Mechatronics and Intelligent Robotics* (Springer, Cham), pp. 889–897.

Maučec, M. S., and Žgank, A. (**2011**). "Speech recognition system of Slovenian broadcast news," *Speech Technologies* (InTech, Rijeka, Croatia), pp. 221–236.

Menne, T., Tüske, Z., Schlüter, R., and Ney, H. (**2018**). "Learning acoustic features from the raw waveform for automatic speech recognition," in *DEGA*, March 19–28, München, Germany, pp. 1533–1536.

Nespor, M., Peña, M., and Mehler, J. (**2003**). "On the different roles of vowels and consonants in speech processing and language acquisition," Lingue Linguaggio **2**(2), 203–230.

Noé, P. G., Bonastre, J. F., Matrouf, D., Tomashenko, N., Nautsch, A., and Evans, N. (**2020**). "Speech pseudonymisation assessment using voice similarity matrices," https://www.researchgate.net/publication/344015535_Speech_Pseudonymisation_Assessment_Using_Voice_Similarity_Matrices (Last viewed September 19, 2020).

Noulas, A., Englebienne, G., and Krose, B. J. (**2012**). "Multi-modal speaker diarization," IEEE Trans. Pattern Anal. Mach. Intell. **34**(1), 79–93.

Ntalampiras, S. (**2020**). "Toward language-agnostic speech emotion recognition," J. Audio Eng. Soc. **68**(1/2), 7–13.

Pellegrino, F., and André-Obrecht, R. (**2000**). "Automatic language identification: An alternative approach to phonetic modelling," Signal Process. **80**(7), 1231–1244.

Pereira, I., Distante, C., Silveira, L. F., and Gonçalves, L. (**2020**). "Using neural networks to compute time offsets from musical instruments," J. Audio Eng. Soc. **68**(3), 157–167.

Rafaely, B., and Alhaiany, K. (**2018**). "Speaker localization using direct path dominance test based on sound field directivity," Signal Process. **143**, 42–47.

Rejaibi, E., Komaty, A., Meriaudeau, F., Agrebi, S., and Othmani, A. (**2019**). "MFCC-based recurrent neural network for automatic clinical depression recognition and assessment from speech," arXiv:1909.07208.

Salamon, J., and Bello, J. P. (**2017**). "Deep convolutional neural networks and data augmentation for environmental sound classification," IEEE Signal Process. Lett. **24**(3), 279–283.

Sarkar, A. K., Do, C. T., Le, V. B., and Barras, C. (**2014**). "Combination of cepstral and phonetically discriminative features for speaker verification," IEEE Signal Process. Lett. **21**(9), 1040–1044.

Satt, A., Rozenberg, S., and Hoory, R. (**2017**). "Efficient emotion recognition from speech using deep learning on spectrograms," in *Proceedings of INTERSPEECH 2017*, August 20–24, Stockholm, Sweden, pp. 1089–1093.

Schoormann, H. E., Heeringa, W. J., and Peters, J. (**2017**). "A cross-dialectal acoustic study of Saterland Frisian vowels," J. Acoust. Soc. Am. **141**(4), 2893–2908.

Strisciuglio, N., Vento, M., and Petkov, N. (**2019**). "Learning representations of sound using trainable COPE feature extractors," Pattern Recogn. **92**, 25–36.

Tüske, Z., Golik, P., Schlüter, R., and Ney, H. (**2014**). "Acoustic modeling with deep neural networks using raw time signal for LVCSR," in *Proceedings of the Fifteenth Annual Conference of the International Speech Communication Association*, September 14–18, Singapore.

Vrysis, L., Tsipas, N., Thoidis, I., and Dimoulas, C. (**2020**). "1D/2D Deep CNNs vs. temporal feature integration for general audio classification," J. Audio Eng. Soc. **68**(1/2), 66–77.

Vryzas, N., Vrysis, L., Matsiola, M., Kotsakis, R., Dimoulas, C., and Kalliris, G. (**2020**). "Continuous speech emotion recognition with convolutional neural networks," J. Audio Eng. Soc. **68**(1/2), 14–24.

Xie, X., and Jaeger, T. F. (**2020**). "Comparing non-native and native speech: Are L2 productions more variable?," J. Acoust. Soc. Am. **147**, 3322.

Yaman, O., Ertam, F., and Tuncer, T. (**2020**). "Automated Parkinson's disease recognition based on statistical pooling method using acoustic features," Medical Hypotheses **135**, 109483.

Yenigalla, P., Kumar, A., Tripathi, S., Singh, C., Kar, S., and Vepa, J. (**2018**). "Speech emotion recognition using spectrogram & phoneme embedding," in *Proceedings of INTERSPEECH 2018*, September 2–6, Hyderabad, India, pp. 3688–3692.

Zewoudie, A. W., Luque, J., and Hernando, J. (**2018**). "The use of long-term features for GMM-and i-vector-based speaker diarization systems," EURASIP J. Audio Speech Music Process. **2018**(1), 14.