Int. J. Appl. Math. Comput. Sci., 2023, Vol. 33, No. 3, 479–492
DOI: 10.34768/amcs-2023-0035

INVESTIGATION OF THE LOMBARD EFFECT BASED ON A MACHINE LEARNING APPROACH

GRAŽINA KORVEL [a], POVILAS TREIGYS [a], KRZYSZTOF KĄKOL [b], BOŻENA KOSTEK [c,*]

[a] Institute of Data Science and Digital Technologies
Vilnius University
Akademijos str. 4, LT-08412 Vilnius, Lithuania
e-mail: {grazina.korvel,povilas.treigys}@mif.vu.lt

[b] PGS Software
ul. Sucha 3, 50-086 Wrocław, Poland
e-mail: krzysztofkakol@gmail.com

[c] Audio Acoustics Laboratory, Faculty of Electronics, Telecommunications and Informatics
Gdańsk University of Technology
ul. G. Narutowicza 11/12, 80-233 Gdańsk, Poland
e-mail: bozena.kostek@pg.edu.pl

The Lombard effect is an involuntary increase in the speaker's pitch, intensity, and duration in the presence of noise. It makes it possible to communicate in noisy environments more effectively. This study aims to investigate an efficient method for detecting the Lombard effect in uttered speech. The influence of interfering noise, room type, and the gender of the person on the detection process is examined. First, acoustic parameters related to speech changes produced by the Lombard effect are extracted. Mid-term statistics are built upon the parameters and used for the self-similarity matrix construction. They constitute input data for a convolutional neural network (CNN). The self-similarity-based approach is then compared with two other methods, i.e., spectrograms used as input to the CNN and speech acoustic parameters combined with the k-nearest neighbors algorithm. The experimental investigations show the superiority of the self-similarity approach applied to Lombard effect detection over the other two methods utilized. Moreover, small standard deviation values for the self-similarity approach prove the resulting high accuracies.

Keywords: Lombard effect, speech detection, noise signal, self-similarity matrix, convolutional neural network.

1. Introduction

Nature has equipped humans with a mechanism to communicate more effectively in noise conditions. The main keywords related to the Lombard effect (LE) are Lombard speech, acoustics communication, noise effects, vocal modifications, or vocal plasticity (Hotchkin and Parks, 2013). This is because the Lombard effect is defined as the unintended tendency of an interlocutor to increase the level of an utterance in noise conditions to improve audibility and intelligibility. Although the definition of the LE is mainly focused on the increase in the utterance level, there are many additional indications in the acoustic analysis that LE increases. This concerns the average amplitude of the signal, modifications in the formant frequencies, changes in the length of utterance, and shifts in the energy from low frequency to high and medium frequency bands.

Careful analyses helped observe additional effects such as flattening the slope of the spectrum, increasing speech intelligibility by increasing the range of formant frequencies, fundamental frequency, or changing the duration of words and vowels, as well as the length of the entire utterance (Hansen, 1994; Kleczkowski *et al.*, 2017; Summers *et al.*, 1988). The LE is also associated with non-acoustic effects, such as more prominent facial muscle movements when speaking in noise (Stathopoulos

*Corresponding author

et al., 2014; Chiu *et al*., 2020). Furthermore, the LE involves cognitive functions of the human brain and, consequently, vocal-motor control over own voice (Kim and Davis, 2014; Luo *et al*., 2018). Our study follows the well-established fact in the scientific community that LE-affected speech becomes more intelligible to the listener (Boril and Hansen, 2009; Garnier and Henrich, 2014). It is also worth noting that recent research on Lombard intelligibility has shown that despite subtle native language influences on non-native Lombard speech, both native and non-native speech provide the benefit (Marcoux *et al*., 2022).

In contrast, the LE may create problems when detecting speech in noise automatically, but not trained on data related to the LE (Vlaj and Kacic, 2011; Marxer *et al*., 2018; Korvel *et al*., 2020; Maheswari *et al*., 2020). Such hyper-articulation impairs the performance of the speech recognition systems (Maheswari *et al*., 2020), so it is essential to train them on data, including Lombard-related features. However, a question remains about what the most representative Lombard-related characteristics are. This was explicitly articulated by Chiu *et al*. (2020). They said that if a speech recognition system (ASR) is trained on data containing Lombard speech, then statistical models built upon them may improve recognition results (Chiu *et al*., 2020).

The motivation for undertaking this study is to find an efficient method for detecting the LE in uttered speech under noise and interference conditions. This is based on searching for the best combination of speech representation and classification algorithm. To that end, several neural network architectures optimized for the type of interfering noise, type of room, and gender of the person being recorded, along with signal representations, are exploited and compared with a baseline algorithm combined with a feature vector. For the purpose of this study, self-similarity matrices are built upon acoustic parameters derived from speech signal processing (Rybka and Janicki, 2013; Panek *et al*., 2015; Gama *et al*., 2021).

The idea of a self-similarity matrix is borrowed from the music information retrieval (MIR) domain, where it is used to characterize the rhythm and tempo of the music (Foote, 1999; Wei *et al*., 2019). Recently, such an approach was employed in the speech area in the context of visualization of speech disfluencies (Esmaili *et al*., 2016) and interlanguage phoneme differences (Korvel *et al*., 2021), as well as pseudonymization performance assessment applied to the speaker's privacy preservation (Noé *et al*., 2022). In addition, it has been shown that taking into account the structural similarity of the image itself, it can be applied in evaluating results based on machine learning in areas not related to speech processing (Dong *et al*., 2012; Wang *et al*., 2018).

In our research, self-similarity matrices are constructed based on acoustic parameters differentiating between neutral speech and produced by the Lombard effect. The resulting graphical representations are derived from the differences between acoustic characteristics of the non-Lombard and Lombard speech signals and are used as an input to a convolutional neural network (CNN), widely used in image classification (Bernardo *et al*., 2021; Kowal and Korbicz, 2019). It should be pointed out that using self-similarity matrices is a novel approach to investigating the Lombard effect phenomenon in uttered speech. In the literature, commonly the short-time Fourier transform (STFT) with both linear and Mel scales, the constant-Q transform (CQT) and the continuous Wavelet transform (CWT), are used in conjunction with the CNN method to represent the signal in the time and frequency domain (Huzaifah, 2017; Choi *et al*., 2018). Also a raw waveform-based approach has been explored to learn hierarchical characteristics of audio directly (Lee *et al*., 2018).

Another motivation behind the experiments is to efficiently detect the LE, because this phenomenon can be used for creating synthesized Lombard speech in the case of noisy environments when better intelligibility is needed. Therefore, it should be determined whether it is present in the input signal to avoid unnecessary speech modifications when the speech is naturally Lombard in its character. This may then be applied to communication systems, public address systems, or hearing aids (Saba and Hansen, 2022).

The paper starts with a brief description of extracting parameters that show changes in time, frequency, and intensity level analysis when the LE occurs. Then, mid-term statistics are built upon these parameters and used for the self-similarity matrix construction, constituting a CNN's input. The main assumptions of the detection-based experiment are shown. The following section deals with the analysis and interpretation of the results obtained in the LE classification. Training accuracies are calculated across the ten splits of the data set of the recorded utterances. The self-similarity-based approach is compared with two other methods, i.e., spectrograms used as input to the CNN and speech acoustic parameters combined with the k-nearest neighbor (kNN) algorithm. Detailed tables containing the classification metrics for all tested methods are presented. On this basis, conclusions are derived, showing the superiority of the self-similarity approach applied to Lombard effect detection over two other methods utilized. Finally, future directions of the experiment development are provided.

2. Measured data

To gather well-controlled data, speech recordings were made in two rooms with different acoustic characteristics. One of the rooms was an acoustically treated studio with

suppressed reverberation (we call this room 'Treat'). In contrast, the second room was an interior with slight acoustic treatment (we call it 'UnTreat'). A microphone was fixed approximately 1 meter away and pointed towards the mouth of the person being recorded. Eight speakers (four males and four females) were asked to read 10 separate words. The format of the speech recordings was the .wav file with the following parameters: 48 kHz, 32 bit, mono.

To obtain utterances of normal speech and with the Lombard effect, the assumed recording scenario was repeated with varying conditions. The normal speech utterances were recorded without additional noise played back. The utterances with the Lombard effect were recorded by playing interference noise via the headphones during the recording process. Before each recording session, measurement of the level was performed using headphones placed on the B&K (head and torso) manikin, type 4128C-002. Then, during the recording, the Brüel & Kjaer 2260 Investigator was employed to monitor the noise level continuously. For listening to noises, we used Philips Stereo Headphones SBC HP195. Speech recordings were made by the Panasonic AG–MC200 microphone and the ZOOM H6 recorder. Two types of noise, i.e., pink noise generated using the noise generator and the natural language samples of babble speech (also known as cocktail-party-effect), were used. These conditions resulted in three types of recordings with the Lombard effect:

(i) playing noise with a cocktail-party effect of an approximately 80 dBA signal level,

(ii) playing broadband pink noise of an approximately 73 dBA signal level,

(iii) playing broadband pink noise of an approximately 84 dBA signal level.

The recording session duration did not exceed 15 minutes; therefore, there was no problem with prolonged exposure to noise. As a result, the recordings consist of audio samples divided according to several categories, namely: the type of interfering noise, the type of room, and the gender of the person being recorded.

As noise levels are given for each of the tested noise contamination scenarios, it is helpful to know the associated signal levels to assume the expected SNRs and foresee the problem difficulty (Tsardoulias *et al.*, 2016; Dimoulas *et al.*, 2006). The combined noise and signal waveform plot is shown in Fig. 1, where the standardized recording is mixed with pink noises at SNR $= 5$ dB and with the energy level maintained.

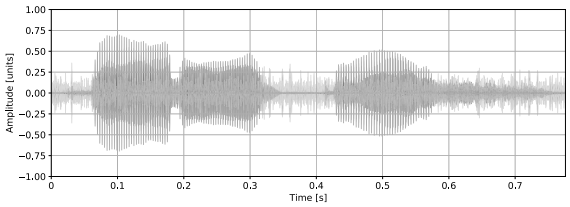

Fig. 1. Noise-free signals and pink noises at SNR $= 5$ dB.

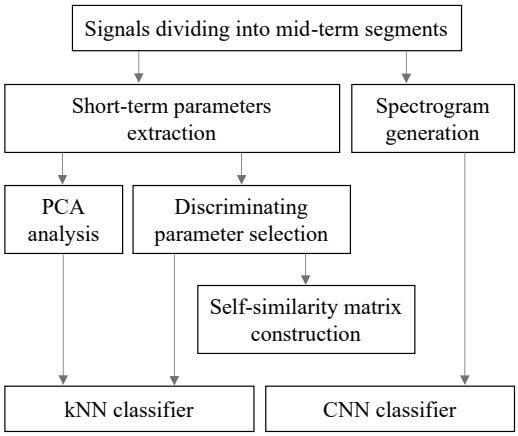

Fig. 2. Block diagram of the combinations of the signal representation and the classification method.

3. Lombard effect detection

This section refers to the primary goal of this study, which concerns detecting the Lombard effect in uttered speech. The method proposed is based on searching for a best combination of speech representation and classification methods. We advocate self-similarity construction based on short- and mid-term statistical properties of speech parameters. The signal representation exploited, along with the algorithm combined, are given in Fig. 2.

3.1. Self-similarity matrix construction. The concept behind the self-similarity method employed in this research is to transform the speech signal into a vector of parameters that are capable of accounting for changes caused by the Lombard effect, and then compare each parameter of the vector with all other parameters of the same vector. The similarity matrix is constructed based on similarity scores between parameters and constitutes the input for a CNN.

Generally, a similarity matrix represents the distance between two vectors of parameters. The distance is determined by the Euclidean distance formula, i.e.,

$$d(\mathbf{p},\mathbf{q}) = \sqrt{\sum_{i=1}^{L}(\mathbf{p}_i - \mathbf{q}_i)^2}, \tag{1}$$

Algorithm 1. Preparation procedure for self-similarity matrix construction.

Step 1. Signal dividing into short-term segments.

Step 2. Signal dividing into mid-term segments.

Step 3. Extraction of short-term parameters related to the Lombard effect.

Step 4. Building mid-term statistics upon short-term parameters.

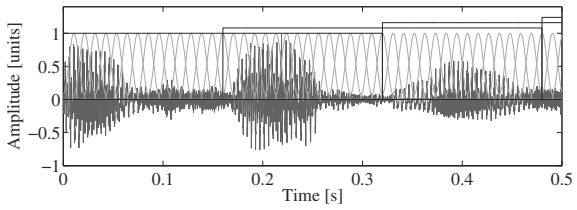

Fig. 3. Speech signal decomposition into short-term and mid-term segments (the grey line denotes short-term segments, the black line refers to mid-term segments).

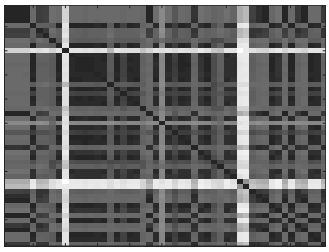

Fig. 4. Example of the self-similarity matrix.

where $\mathbf{p}$ and $\mathbf{q}$ are two vectors of parameters and L denotes the number of parameters.

In this research, the self-similarity matrix is used. This means that the distance computation is performed for the same set of parameters, i.e., $\mathbf{p} = \mathbf{q}$.

Before the matrices are built, the 4-step procedure represented as Algorithm 1 is performed.

The self-similarity matrix construction process starts with dividing the speech signal into short-term and mid-term segments (see Steps 1 and 2 of Algorithm 1). An example of the signal division is given in Fig. 3. In this stage the following settings are used: the length of the short-term segment is 1024 samples, the overlap between segments is 50%, and the number of the short-term segments in a mid-term segment is equal to 30. The short-term segment length and overlap were chosen based on our previous studies related to signal parameterization (Korvel *et al.*, 2019), while the number of intervals was derived from several initial tests.

In Step 3, the acoustic parameters related to the Lombard effect are calculated for each short-term segment. The procedure of parameter calculation is given in Section 3.1. In the last step, for each mid-term segment, the short-term parameter statistics are calculated. The mean value as a mid-term statistic is employed. Mid-term statistics cover general changes in parameters over time and are commonly used in speech analysis (Smailis *et al.*, 2016; Piotrowska *et al.*, 2021).

The self-similarity matrix is constructed for each mid-term segment based on its mid-term statistics. An example of the self-similarity matrix for mid-term segment of speech with Lombard effect is given in Fig. 4, where the x and y-axes represent the parameter numbers. Each pixel is given a grey-scale value proportional to the distance value. Darker colors mean a higher similarity between parameters.

3.2. Short-term parameters. As already mentioned, the Lombard effect modifies the volume of the uttered speech, fundamental frequency, formant frequency, spectral tilt, and duration. Therefore, the acoustic characteristics, which reflect this phenomenon both in the time- and the frequency domains, should be examined. The time-domain parameters are extracted directly from the samples of the speech signal. The frequency characteristics are calculated from the Fourier spectrum:

$$X\left(k\right) = \sum_{n=0}^{N-1} x(n) w\left(n\right) e^{\frac{-2\pi \mathrm{j} k n}{N}}, \tag{2}$$

where $X(k)$ are the Fourier transform coefficients, $k = 0, \ldots, K-1$ (K is the number of the Fourier transform coefficients), $x(n)$ means the samples of a short-time segment of the speech signal, N stands for the length of the short-time segment, $w\left(n\right)$ is the Hamming window function, and j is the imaginary unit.

We investigate an extensive set of parameters to evaluate the Lombard effect appropriately. The investigated parameters are given in Table 1. The set of parameters was constructed based on our previous experience with acoustic analysis of speech signals. The parameters utilized are typical speech parameters as well as features borrowed from the music information retrieval (MIR) area and the MPEG-7 standard (Downie, 2003; Schedl *et al.*, 2014; Kim *et al.*, 2005). Our previous experiments showed that music domain-derived features benefit speech signal processing (Korvel *et al.*, 2019; Piotrowska *et al.*, 2019).

Overall, we have 106 extracted parameters for each short-term segment (see Table 1). The time-domain representation shows the time-varying behavior of the signal. The temporal centroid (TC), i.e., the first parameter, represents the time point where half of the signal energy of the short-time speech segment $x(n)$ occurs,

$$\mathrm{TC} = \frac{\sum_{n=1}^{N} n x^2(n)}{\sum_{n=1}^{N} x^2(n)}, \tag{3}$$

Table 1. Parameters extracted from the speech signal.

No.	Parameter
Time-domain parameters	
1	Temporal centroid (TC)
2	Zero crossing rate (ZCR)
3	Root mean square (RMS) energy
4-6	The number of samples exceeding levels RMS, 2× RMS, 3×RMS $(p_1,\ p_2,\ p_3)$
7-12	The mean and variance of samples exceeding levels RMS, 2×RMS, 3×RMS averaged for 10 sub-segments $(\mu(p_1),\ \mu(p_2),\ \mu\,(p),$ $\sigma^2(p_1), \sigma^2(p_2), \sigma^2\,(p_3))$
13	Peak to RMS
14-17	The number of the signal crossings in relation to zero, RMS, 2×RMS, 3×RMS $(q_1,\ q_2,\ q_3,\ q_4)$
18-25	The mean and variance of signal crossings in relation to zero, RMS, 2×RMS, 3×RMS averaged for 10 sub-segments $(\mu(q_1), \mu(q_2),$ $\mu\,(q_3),\ \ \mu(q_4),\ \ \sigma^2(q_1),\ \ \sigma^2(q_2),\ \ \sigma^2\,(q_3),$ $\sigma^2(q_4))$
Frequency-domain parameters	
31	Audio spectral centroid (ASC)
32	Audio spectral spread (ASSp)
33	Audio spectral skewness (ASSk)
34	Audio spectral kurtosis (ASK)
35	Spectral entropy
36	Spectral roll-off
37	Spectral brightness
38-66	Audio spectrum envelope calculated on 29 sub-bands (ASE1-ASE29)
67	Mean audio spectrum envelope (MASE)
68-85	Spectral flatness measure calculated on 18 sub-bands (SFM1-SFM18)
86	Mean spectral flatness measure (MSFM)
87-106	Mel-frequency cepstral coefficients (MFCC1- MFCC20)

where N is the length of the short-time segment.

The second parameter (the zero crossing rate ZCR) is the number of the time axis crossings of the signal,

$$\mathrm{ZC} = \frac{1}{N-1}\sum_{n=2}^{N}|s_n - s_{n-1}|, \tag{4}$$

where

$$s_n = \begin{cases} 1 & \text{if } x\,(n) > 0, \\ 0 & \text{if } x\,(n) \le 0. \end{cases} \tag{5}$$

Root mean square (RMS) energy represents the average power of the analyzed short-time speech segment $x\,(n)$

$$\mathrm{RMS} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}x^2(n)}. \tag{6}$$

The RMS parameter is also used to extract information on the speech temporal behavior (see parameters no. 4–25). These parameters were proposed by Kostek *et al*. (2011) and are based on the analysis of the distribution of the sound sample values in relation to the RMS levels.

The frequency domain parameters show how the signal energy is distributed within the frequency range. The parameters are derived from the power spectrum

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

where $X(k)$ are Fourier transform coefficients calculated by Eqn. (2), $k = 0,\ \ldots, K-1$ (K is the number of Fourier transform coefficients), 're' and 'im' mean real and imaginary parts, respectively.

Parameters 31–37 are the so-called spectral shape parameters. These measures are based on an octave frequency scale centered at 1 kHz (Kim *et al*., 2005; Korvel *et al*., 2019).

Parameters related to the audio spectrum envelope (nos. 38–67) give a compact representation of the power spectrum of the speech signal, while the spectral flatness measure parameters (nos. 68–86) let separate voiced and unvoiced speech. These parameters are considered on a sub-band level. The audio spectrum envelope (ASE) in a single band l $(l = 1,\ \ldots,\ 29)$ is calculated as follows:

$$\mathrm{ASE}(l) = \begin{cases} \sum\limits_{k=0}^{P_1}\mathrm{PS}(k), & l = 1, \\ \sum\limits_{k=P_{l-1}}^{P_l}\mathrm{PS}(k), & 2 \le l \le L+1, \\ \sum\limits_{k=P_{l+1}}^{f_s/2}\mathrm{PS}(k), & l = L+2, \end{cases} \tag{8}$$

where $L = 29$, $\mathrm{PS}(k)$ is the power spectrum of the short-time segment calculated by Eqn. (7), f_s is the sampling frequency, P_l means band frequency range values.

The spectral flatness measure (SFM) in a single band l $(l = 1,\ \ldots,\ 18)$ is calculated by

$$\mathrm{SFM}(l) = \frac{\left[\prod_{k=P_l}^{P_{l+1}}\mathrm{PS}(k)\right]^{\frac{1}{N}}}{\frac{1}{P_{l+1}-P_l+1}\sum_{k=P_l}^{P_{l+1}}\mathrm{PS}(k)}, \tag{9}$$

where $\mathrm{PS}(k)$ is the short-time power spectrum, P_l are the edges of the frequency bands.

The sub-band edges P_l (see Eqns. (8) and (9)) are logarithmically distributed corresponding to a specific

octave frequency. The calculation procedure of the sub-bands is given by Korvel *et al*. (2019).

The extraction process of mel-frequency cepstral coefficients (MFCC1–MFCC20), parameters no. 87–106, begins with filtering the short-time power spectrum (see Eqn. (7)) by triangle bandpass filters. The scale of filters is linear up to 1 kHz and logarithmic above this frequency. Then, a log magnitude is calculated to obtain the real cepstrum. The cepstral coefficients are obtained by applying the discrete cosine transform (DCT):

$$c_j = \sum_{i=0}^{K-1} m_i \cos\left(\frac{\pi j(i-\frac{1}{2})}{M}\right), \tag{10}$$

where m_i are the log filterbank amplitudes, M is the number of triangle bandpass filters, j is the index of the cepstral coefficient ($j = 1, \ldots, 20$).

A detailed description of the parameters used is given by Kostek *et al*. (2011) and Korvel *et al*. (2019). Before performing an analysis of these parameters, they should be normalized. The normalization to the range [0, 1] is used. This process can be described by the following formula:

$$\widetilde{v}_l = \frac{v_l - \min(\mathbf{V})}{\max(\mathbf{V}) - \min(\mathbf{V})}, \tag{11}$$

where $\mathbf{V} = (v_1, \ldots, v_L)$ is a vector of non-normalized parameters, the values v_l and $\widetilde{v}_l$ refer to the normalized and non-normalized l-th parameter, respectively, $l \in [1, L]$, L signifies the number of parameters.

3.2.1. Parameter vector dimensionality reduction. We aim to disclose which of the extracted acoustic parameters concern the Lombard effect the most. To check which parameters of the 106 extracted ones are important, we implement a dimensionality reduction technique by employing a discriminating parameter selection procedure. According to this procedure, correlation coefficients are calculated between the parameters extracted from the speech signal with and without the Lombard effect. The parameters for which correlation coefficients are greater than a threshold prescribed are rejected. To find out which threshold yields the best results, systematic and controlled experiments were carried out. As a result, an optimized feature vector is determined, which shows differences between signals recorded with the Lombard effect and without it. The correlation coefficient was 0.5. The analysis considers the dependence of the type of interfering noise, the type of room in which the recordings took place, and the gender of the person exposed to noise while recorded. A set of the optimized acoustic parameters in the context of the Lombard effect is given in Table 2. The self-similarity matrix is constructed based on these parameters. An example of a graphical representation of separation based on these parameters is given in Fig. 6.

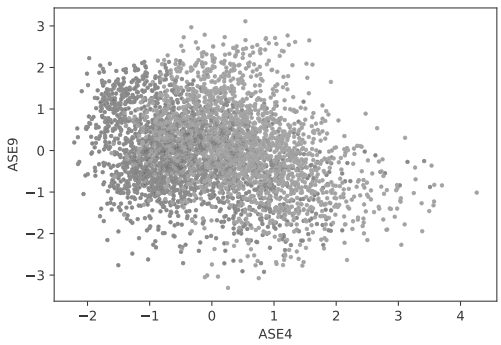

Fig. 5. Separation of the Lombard effect based on Audio Spectrum Envelope calculated on four and nine sub-bands (the brighter circles denote the recordings without the Lombard effect, the darker circles—recordings with the Lombard effect (cocktail-party-effect type noise; level of 80 dBA)).

Table 2. Set of the optimized acoustic parameters in terms of Lombard effect detection.

No.	Parameter
Time-domain parameters	
4-6	The number of samples exceeding levels RMS, 2× RMS, 3×RMS (p_1, p_2, p_3)
13	Peak to RMS
Frequency-domain parameters	
35	Spectral entropy
39-55, 62-64	Audio spectrum envelope calculated on 29 sub-bands (ASE2–ASE17, ASE25–ASE27)
69-76, 82-85	Spectral flatness measure calculated on 18 sub-bands (SFM2–SFM8, SFM15–SFM18)
86	Mean spectral flatness measure (SFM)
87-88, 97-106	Mel-frequency cepstral coefficients (MFCC1–MFCC2, MFCC11–MFCC20)

In addition, principal component analysis (PCA) was performed to achieve possibly the most orthogonal dimensions (Kherif and Latypova, 2020; Diamantaras, 2002). The PCA method was applied to the set consisting of 106 parameters (Table 1) calculated for all mid-term segments. As a result, we obtained 58 components sufficient to contain 99% of the information.

3.3. Spectrogram generation. The spectrogram is the most often used representation of a speech signal (Ouyang *et al*., 2019; Nugraha *et al*., 2020). A spectrogram is constructed from a series of short-time Fourier transforms (see Eqn. (7)), which are computed along the time domain waveform of the analyzed speech signal. The obtained values are collected together, and a spectrogram image is built up. A graphical representation of the spectrogram obtained is given in Fig. 6.

3.4. Convolutional neural network selection. In our research, a convolutional neural network (CNN) is investigated and trained on self-similarity matrices to obtain a model that can most precisely detect the Lombard effect. The model architectures were investigated by applying a hyperparameter optimization framework (O'Malley *et al*., 2019) and the hyperband optimization technique (Li *et al*., 2017). The structure of the search space was selected as follows: every convolutional layer was followed by MaxPooling and batch normalization. The maximum number of convolutional layers was limited to five. Then up to five more dense layers can be added before flattening and the classifying dense layer, which is placed at the end of the architecture.

In each convolutional layer the number of filters ranged from to 8 to 128, the kernel size was in the range from 1 to 7, and the activations explored were ELU, ReLU, TanH and Sigmoid. Pool sizes in MaxPooling varied in the range from 2 to 4. The number of dense layer units selected by the optimization algorithm varied from to 8 to 128 and the investigated set of possible activation functions was the same as those tested for every convolutional layer. The Glorot uniform algorithm (Glorot and Bengio, 2010) was used for kernel weight initialization (convolutional and dense layers). The model was compiled using a binary cross-entropy loss function and the Adam (adaptive moment) optimizer (Kingma and Ba, 2014) at varying learning rate steps: 0.1, 0.01, 0.001, and 0.0001. The adaptive moment estimation parameters β_1 and β_2 were set to 0.95 and 0.999, respectively. Input images are scaled to 256×256 and resized using the nearest neighbor method.

The summaries of the best-found architectures for different types of recordings with the Lombard effect are presented in Tables 6 to 8. After obtaining the best architecture from the hyperband optimization algorithm, each architecture's robustness was inspected further by training the model 100 times. To prevent the neural network from over training and to obtain the best results, early stopping, together with a reduced learning rate on plateau techniques, were used. The reduced learning rate on the plateau parameter factor was set to 0.2, and the minimum learning rate was set to 0.001. During each architecture training, validation, and test stage, data samples were selected randomly while keeping data split proportions: 64% of the data were used for model training, 16% for model validation, and 20% of total data for model testing. Finally, the best-performing model performance is visualized in Fig. 10. The model is constructed using the Python programming language and the Keras Python deep learning library with the TensorFlow library (TensorFlow library, Keras library) (Manaswi *et al*., 2018).

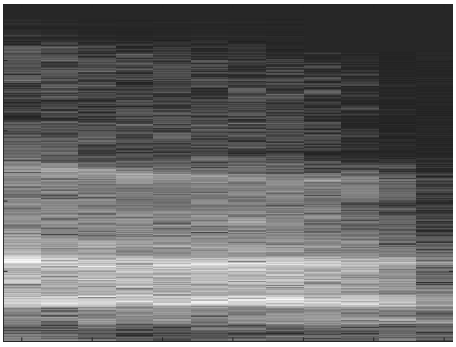

Fig. 6. Example of the spectrogram representation.

Table 3. Split of mid-term segments with regard to the different types of interfering noise.

Type of noise	No. of segms.
Wihout interfering noise	2233
Pink noise of 73 dBA signal level	2333
Pink noise of 84 dBA signal level	2233
The cocktail-party-effect of 80 dBA	1518

Table 4. Split of mid-term segments with regard to the room type.

Type of room	No. of segms.
An acoustically treated room	4196
A studio with light acoustic treatment	4121

Table 5. Split of mid-term segments with regard to the speaker the gender.

Type of room	No. of segms.
An acoustically treated room	4521
A studio with light acoustic treatment	3796

4. Experimental results

The experiment is performed on speech recordings divided into mid-term segments. In this way, 8317 mid-term segments were obtained. Information about the split of mid-term segments with regard to the different types of noise used during the recording process is given in Table 3. Accordingly, Tables 4 and 5 show the split of mid-term segments regarding the room type and the gender of the person being recorded, respectively.

In the first part of the experiment the effectiveness of the optimized acoustic parameter set (see Table 2) and PCA components obtained from all extracted parameters (Table 1) is evaluated. For this purpose, the kNN algorithm is employed (Zhang *et al*., 2017). The evaluation of the accuracy measure is performed separately for interference noise, room type, and gender. To obtain results robust to random sampling, a 100-fold random sub-sampling method is used (Berrar, 2019). Based on this method, 100 pairs of training and testing sets are generated. The learning function is applied to

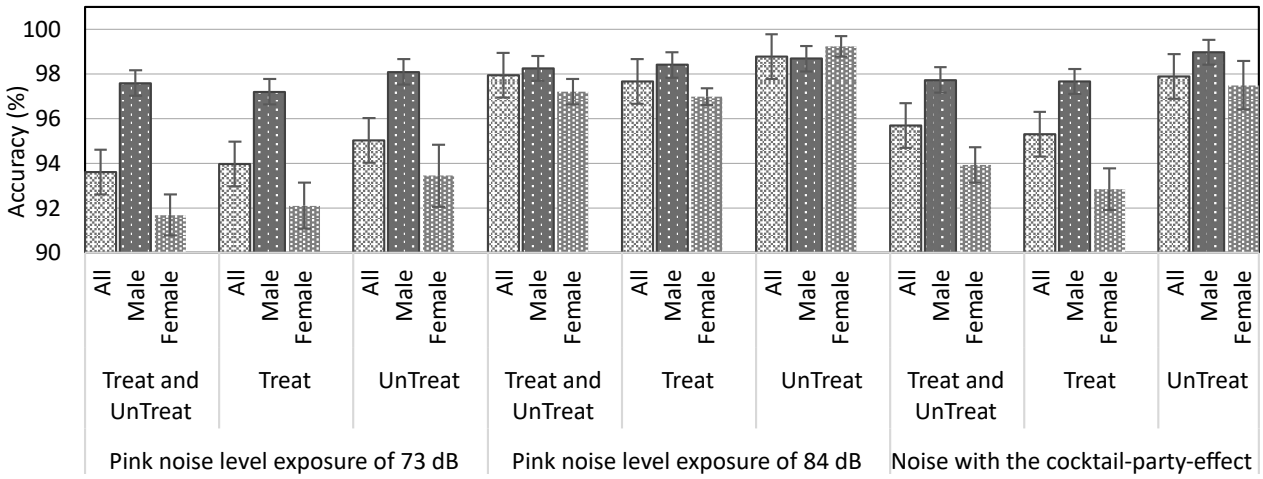

Fig. 7. Performance metrics [%] of feature vector containing acoustic parameters from Table 2 combined with the kNN across the ten splits for recordings with the Lombard effect ('Treat' denotes an acoustically treated room, 'UnTreat' is a studio with a light acoustic treatment).

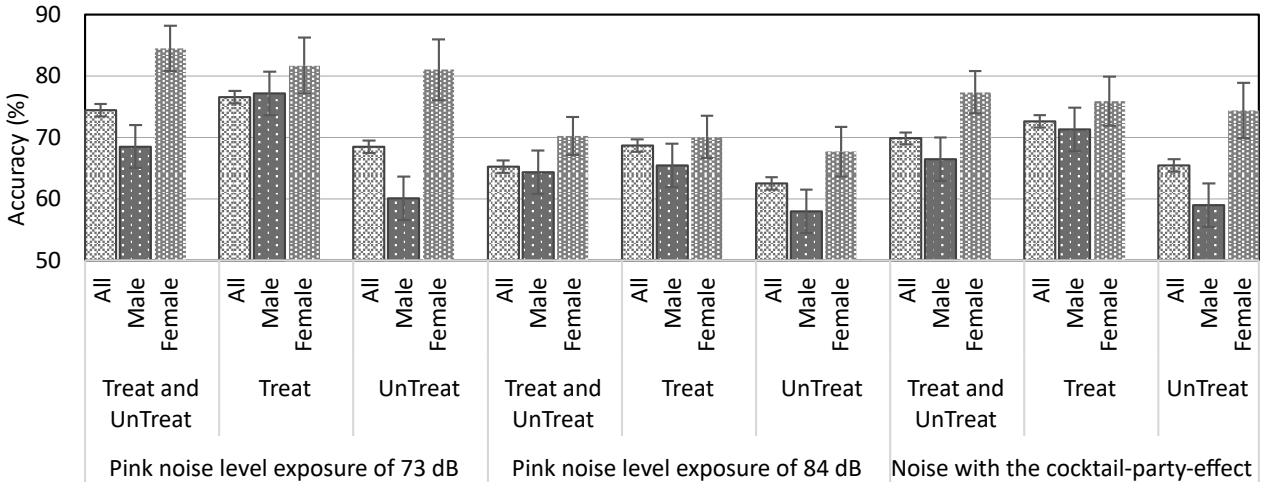

Fig. 8. Performance metrics [%] of 58 PCA components combined with the kNN across the ten splits for recordings with the Lombard effect.

each training set, and the resulting model is then applied to the corresponding test set. The performance is estimated as the average over 100 test sets. The classification results are given in Figs. 7 and 8.

As can be seen from the results, the classification accuracies obtained by using the optimized acoustic parameters set are higher than those using PCA components; therefore, the self-similarity matrices based on the optimized acoustic parameters set are created and introduced as 2D space features at the CNN input. For comparing the results obtained, spectrogram representations, an alternative method of speech signal representation in the 2D space, are used. The different CNN models were trained for each type of interfering noise, the type of room, and the gender of the person being recorded. The architectures of the models trained for each kind of noise without a split into the gender of the speaker and room type are given in Tables 6–8, where the names of all layers, along with hyperparameters, are presented. The obtained results are presented in Figs. 9 and 10.

As can be seen from the results, the highest classification accuracies were achieved for the self-similarity approach. The feature vector containing acoustic parameters related to the Lombard effect combined with the kNN also produces good results, with minor differences.

When comparing the obtained results in the context of interference noise, the following tendencies are seen: while playing broadband pink noise with a higher noise level (i.e., 84 dBA), the classification performance of

Table 6. CNN architecture for recordings with the Lombard effect recordings playing broadband pink noise at an approximately 73 dBA signal level.

Layer name	Filters	Units	Kernel size	Pool size	Padding	Stride	Activation
Conv2D	56		3,3		same	1,1	ReLU
MaxPooling2D				3,3		3.3	
BatchNormalization							
Conv2D	48		3,3		same	1,1	TanH
MaxPooling2D				2,2		2,2	
BatchNormalization							
Dense		96					ELU
BatchNormalization							
Dense		96					ELU
BatchNormalization							
Flatten							
Dense		2					Sigmoid

Table 7. CNN architecture for recordings with the Lombard effect recordings playing broadband pink noise at an approximately 84 dBA signal level.

Layer name	Filters	Units	Kernel size	Pool size	Padding	Stride	Activation
Conv2D	32		5,5		same	1,1	ReLU
MaxPooling2D				2,2		2,2	
BatchNormalization							
Conv2D	64		5,5		same	1,1	Sigmoid
MaxPooling2D				3,3		3,3	
BatchNormalization							
Dense		80					Sigmoid
BatchNormalization							
Flatten							
Dense		2					Sigmoid

all methods employed is better in comparison with the case of recordings with a lower pink noise level (i.e., 73 dBA). This led us the conclusion that the higher the background noise level, the stronger the Lombard effect (Bottalico *et al.*, 2017; 2022). In the case of the noise with the cocktail-party effect, the classification scores for methods based on acoustic parameters are more effective than recordings with pink noise distortion (regardless of the noise level). This is not confirmed for the joint spectrogram-CNN-based approach, but, notably, the scores of recordings with the cocktail-party-effect noise are significantly higher compared with the same level of pink noise.

The analysis of the results obtained regarding the room type revealed that for the method based on acoustic parameters, the classification accuracies of recordings obtained in a studio with a light acoustic treatment are better than those obtained in an acoustically treated room. For the spectrograms, this tendency related to room type is not observed.

Comparing the results obtained regarding the speaker gender, we can observe that the classification accuracy using spectrograms and the CNN and acoustic-based features together with the kNN is significantly different for male and female speakers, while the difference is not significant using the similarity approach. Moreover, it can be seen that the difference between male and female speakers, in the case of spectrograms, shows better accuracy when evaluating female and male recordings separately than when considering all recordings together.

Different CNN models were trained for each type of interfering noise, room type, and gender of the person being recorded to check whether the proposed feature space generalizes well. Following the analysis of the results obtained, we can see that the models generalize well to all data sets. As a result, we have inspected a single generalized model of merged data. According to Table 3 to represent the negative class (C1), we took all the data without interfering noise. To represent the positive class (C2), one third of the data were taken from the samples representing 73 dBA, another one third representing 84 dBA and the final one third representing the cocktail party effect of a 80 dBA signal level. Such a scheme enforced a positive-negative class data balance. The procedure for architecture selection and best-model evaluation was applied as described in Section 3.4. We

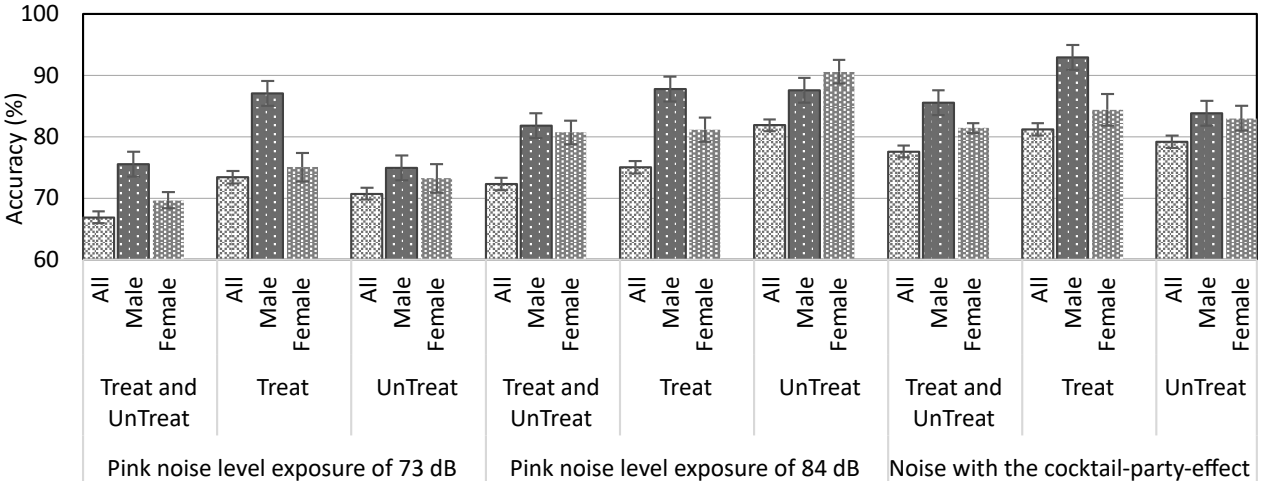

Fig. 9. Performance metrics [%] of spectrograms and the CNN across the ten splits for recordings with the Lombard effect.

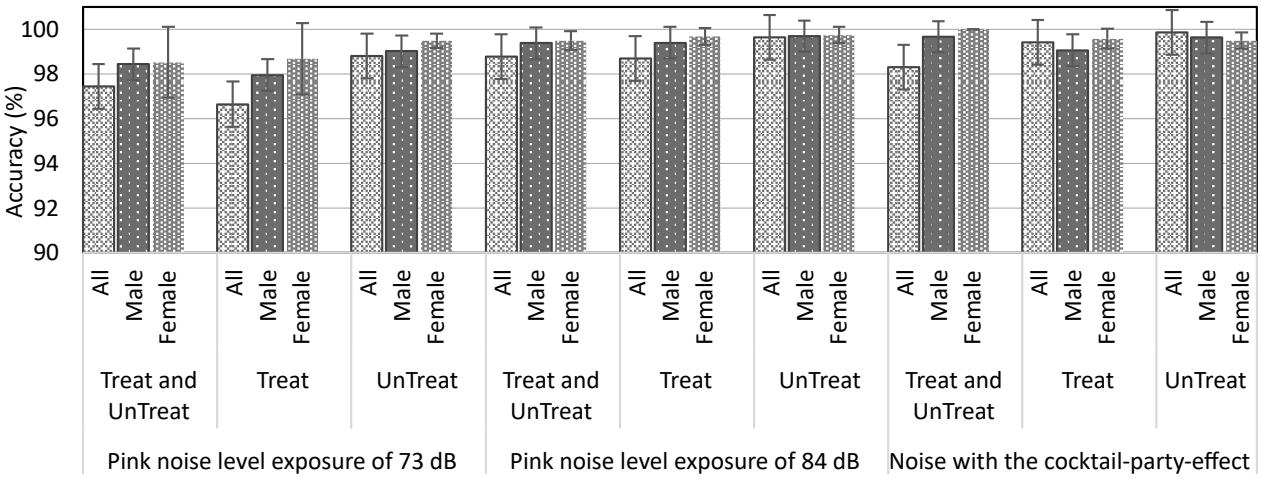

Fig. 10. Performance metrics [%] of self-similarity matrices and the CNN across the ten splits for recordings with the Lombard effect.

assume that the 100-times train-validated model resolves the room type and speaker gender data variability. The obtained architecture of the model is given in Table 9. The performance of the model is shown in Fig. 11, where the binary confusion matrix is visualized. The overall accuracy derived from the confusion matrix was 98.99% and statistics for 100 train/validation/test trials test set classification results converged to an accuracy mean of 97.2%. at the standard deviation of 0.7%.

5. Conclusions

In the paper, we have evaluated the performance of the proposed self-similarity approach for Lombard effect detection. For comparison of results, classification employing kNN and CNN methods using acoustic parameters and spectrograms, respectively, as inputs is performed. The experimental investigations showed that regardless of different factors involved (i.e., type of interfering noise, type of room, and gender of the person being recorded), the highest accuracies were achieved by the self-similarity approach. Moreover, the proposed self-similarity feature space shows good classifier generalization properties at different noise types and levels. Classification results converged to the mean accuracy of 97.2% at the standard deviation of 0.7% while classifying data of different factors involved. This leads us to conclude that, despite the simplicity of the selected network architecture, the self-similarity shows the superiority over the other two methods utilized in context of the Lombard effect detection. Moreover, the small standard deviation for the self-similarity approach indicates that the accuracies obtained are quite precise.

An additional conclusion related to the range of

Table 8. CNN architecture for recordings with the Lombard effect recordings playing the noise with the cocktail-party effect at an approximately 80 dBA signal level.

Layer name	Filters	Units	Kernel size	Pool size	Padding	Stride	Activation
Conv2D	40		1,1		same	1,1	ReLU
MaxPooling2D				2,2		2,2	
BatchNormalization							
Conv2D	64		5,5		same	1,1	ELU
MaxPooling2D				2,2		2,2	
BatchNormalization							
Dense		56					TanH
BatchNormalization							
Flatten							
Dense		2					Sigmoid

Table 9. General CNN architecture for all types of recordings with the Lombard effect.

Layer name	Filters	Units	Kernel size	Pool size	Padding	Stride	Activation
Conv2D	72		3,3		same	1,1	ELU
MaxPooling2D				2,2		2,2	
BatchNormalization							
Conv2D	128		1,1		same	1,1	ELU
MaxPooling2D				2,2		2,2	
BatchNormalization							
Conv2D	112		3,3		same	1,1	ReLU
MaxPooling2D				3,3		3,3	
BatchNormalization							
Dense		112					ELU
BatchNormalization							
Flatten							
Dense		2					Sigmoid

parameters can be formulated. While in many cases, extensive use of possible parameters proves helpful, especially in the case of machine learning, in terms of detection of the Lombard effect, it has not been. This is confirmed by the fact that the Lombard effect detection based on specific Lombard phenomenon-related parameters has shown very good results. In contrast, spectrograms covering a wide range of parameters showed lower accuracy. The same can be said about traditional dimensionality reduction techniques (i.e., PCA) used in data mining. The dimensionality reduction of a data set by transforming it into a new coordinate system has not worked. The reason for this may be that the Lombard speech characteristics vary over time, and they are difficult to detect as they depend on several factors. We also found that the network learns to detect other phenomena at the same time. This is supported by the fact that, in the case of spectrograms, the accuracy of evaluating female and male recordings separately is better than that of assessing all recordings together. We can see that the network learns features related not only to the Lombard effect but also, as in the referred case, to the gender of the speaker. However, this statement requires a thorough analysis, which we should follow in future research.

Lastly, the idea of a system mimicking the natural way of speaking in noisy conditions by speech synthesis is to be pursued as the area of hearing aid may benefit from such an approach.

Acknowledgment

This research is funded by the European Social Fund under the grant no. 09.3.3-LMT-K-712: *Development of Competences of Scientists, Other Researchers and Students Through Practical Research Activities*.

References

Bernardo, L.S., Damaševičius, R., de Albuquerque, V.H.C. and Maskeliūnas, R. (2021). A hybrid two-stage SqueezeNet and support vector machine system for Parkinson's disease detection based on handwritten spiral patterns, *International Journal of Applied Mathematics and Computer Science* **31**(4): 549–561, DOI: 10.34768/amcs-2021-0037.

Berrar, D. (2019). Cross-validation, *in* S. Ranganathan *et al.* (Eds), *Encyclopedia of Bioinformatics and Computational Biology*, Academic Press, Oxford, pp. 542–545.

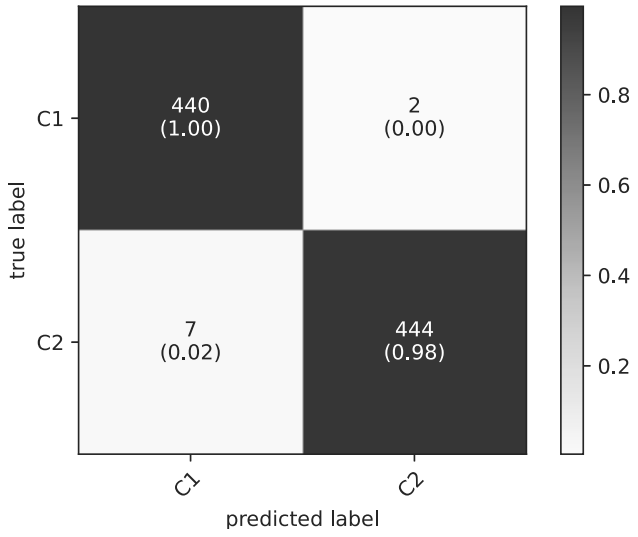

Fig. 11. Performance of each class (C1 denotes the recordings of normal speech, and C2 corresponds to the ones with the Lombard effect).

Boril, H. and Hansen, J.H. (2009). Unsupervised equalization of Lombard effect for speech recognition in noisy adverse environments, *IEEE Transactions on Audio, Speech, and Language Processing* **18**(6): 1379–1393.

Bottalico, P., Passione, I.I., Graetzer, S. and Hunter, E.J. (2017). Evaluation of the starting point of the Lombard effect, *Acta Acustica United With Acustica* **103**(1): 169–172.

Bottalico, P., Piper, R.N. and Legner, B. (2022). Lombard effect, intelligibility, ambient noise, and willingness to spend time and money in a restaurant amongst older adults, *Scientific Reports* **12**(1): 1–9.

Chiu, W., Xu, Y., Abel, A., Lin, C. and Tu, Z. (2020). Investigating the visual Lombard effect with Gabor based features, *Proceedings of INTERSPEECH*, pp. 4606–4610, (online).

Choi, K., Fazekas, G., Sandler, M. and Cho, K. (2018). A comparison of audio signal preprocessing methods for deep neural networks on music tagging, *26th European Signal Processing Conference (EUSIPCO), Rome, Italy*, pp. 1870–1874.

Diamantaras, K.I. (2002). Neural networks and principal component analysis, *in* Y.H. Hu and J.-N. Hwand (Eds), *Handbook of Neural Network Signal Processing*, CRC Press, Boca Raton, pp. 8.1–8.38, DOI: 10.1201/9781315220413.

Dimoulas, C., Kalliris, G., Papanikolaou, G. and Kalampakas, A. (2006). Novel wavelet domain wiener filtering de-noising techniques: Application to bowel sounds captured by means of abdominal surface vibrations, *Biomedical Signal Processing and Control* **1**(3): 177–218.

Dong, W., Zhang, L., Shi, G. and Li, X. (2012). Nonlocally centralized sparse representation for image restoration, *IEEE Transactions on Image Processing* **22**(4): 1620–1630.

Downie, J.S. (2003). Music information retrieval, *Annual Review of Information Science and Technology* **37**(1): 295–340.

Esmaili, I., Dabanloo, N.J. and Vali, M. (2016). Automatic classification of speech dysfluencies in continuous speech based on similarity measures and morphological image processing tools, *Biomedical Signal Processing and Control* **23**: 104–114.

Foote, J. (1999). Visualizing music and audio using self-similarity, *Proceedings of the 7th ACM International Conference on Multimedia (Part 1), Orlando, USA,* pp. 77–80.

Gama, R., Castro, M.E., van Lith-Bijl, J.T. and Desuter, G. (2021). Does the wearing of masks change voice and speech parameters?, *European Archives of Oto-Rhino-Laryngology* **2022**(279): 1701–1708, DOI: 10.1007/s00405-021-07086-9.

Garnier, M. and Henrich, N. (2014). Speaking in noise: How does the Lombard effect improve acoustic contrasts between speech and ambient noise?, *Computer Speech & Language* **28**(2): 580–597.

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks, *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics, Sardinia, Italy*, pp. 249–256.

Hansen, J.H. (1994). Morphological constrained feature enhancement with adaptive cepstral compensation (MCE-ACC) for speech recognition in noise and Lombard effect, *IEEE Transactions on Speech and Audio Processing* **2**(4): 598–614.

Hotchkin, C. and Parks, S. (2013). The Lombard effect and other noise-induced vocal modifications: Insight from mammalian communication systems, *Biological Reviews* **88**(4): 809–824.

Huzaifah, M. (2017). Comparison of time-frequency representations for environmental sound classification using convolutional neural networks, *arXiv:* 1706.07156.

Kherif, F. and Latypova, A. (2020). Principal component analysis, *in* A. Mechelli and S. Vieira (Eds), *Machine Learning*, Academic Press, Cambridge, pp. 209–225, DOI: 10.1016/B978-0-12-815739-8.00012-2.

Kim, H.-G., Moreau, N. and Sikora, T. (2005). *MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval*, Wiley, Chichester.

Kim, J. and Davis, C. (2014). Comparing the consistency and distinctiveness of speech produced in quiet and in noise, *Computer Speech & Language* **28**(2): 598–606.

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

Kleczkowski, P., Żak, A. and Król-Nowak, A. (2017). Lombard effect in Polish speech and its comparison in English speech, *Archives of Acoustics* **42**(4): 561–569.

Korvel, G., Kąkol, K., Kurasova, O. and Kostek, B. (2020). Evaluation of Lombard speech models in the context of speech in noise enhancement, *IEEE Access* **8**: 155156–155170, DOI: 10.1109/ACCESS.2020.3015421.

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

Korvel, G., Treigys, P. and Kostek, B. (2021). Highlighting interlanguage phoneme differences based on similarity matrices and convolutional neural network, *Journal of the Acoustical Society of America* **149**(1): 508–523.

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.

Kowal, M. and Korbicz, J. (2019). Refinement of convolutional neural network based cell nuclei detection using Bayesian inference, *41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany*, pp. 7216–7222.

Lee, J., Park, J., Kim, K.L. and Nam, J. (2018). SampleCNN: End-to-end deep convolutional neural networks using very small filters for music classification, *Applied Sciences* **8**(1): 1–14.

Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A. and Talwalkar, A. (2017). Hyperband: A novel bandit-based approach to hyperparameter optimization, *Journal of Machine Learning Research* **18**(1): 6765–6816.

Luo, J., Hage, S.R. and Moss, C.F. (2018). The Lombard effect: From acoustics to neural mechanisms, *Trends in Neurosciences* **41**(12): 938–949.

Maheswari, S.U., Shahina, A., Rishickesh, R. and Khan, A.N. (2020). A study on the impact of Lombard effect on recognition of hindi syllabic units using CNN based multimodal ASR systems, *Archives of Acoustics* **45**(3): 419–431.

Manaswi, N.K., Manaswi, N.K. and John, S. (2018). *Deep Learning with Applications Using Python*, Apress, Berkeley.

Marcoux, K., Cooke, M., Tucker, B.V. and Ernestus, M. (2022). The Lombard intelligibility benefit of native and non-native speech for native and non-native listeners, *Speech Communication* **136**: 53–62.

Marxer, R., Barker, J., Alghamdi, N. and Maddock, S. (2018). The impact of the Lombard effect on audio and visual speech recognition systems, *Speech Communication* **100**: 58–68.

Noé, P.-G., Nautsch, A., Evans, N., Patino, J., Bonastre, J.-F., Tomashenko, N. and Matrouf, D. (2022). Towards a unified assessment framework of speech pseudonymisation, *Computer Speech & Language* **72**: 101299.

Nugraha, A.A., Sekiguchi, K. and Yoshii, K. (2020). A flow-based deep latent variable model for speech spectrogram modeling and enhancement, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **28**: 1104–1117.

O'Malley, T., Bursztein, E., Long, J., Chollet, F., Jin, H. and Invernizzi, L. (2019). KerasTuner—A hyperparameter optimization framework, `https://github.com/keras-team/keras-tuner`.

Ouyang, Z., Yu, H., Zhu, W.-P. and Champagne, B. (2019). A fully convolutional neural network for complex spectrogram processing in speech enhancement, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK*, pp. 5756–5760.

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., Czyżewski, A., Ciszewski, T., Korvel, G., Kurowski, A. and Kostek, B. (2021). Evaluation of aspiration problems in L2 English pronunciation employing machine learning, *Journal of the Acoustical Society of America* **150**(1): 120–132.

Piotrowska, M., Korvel, G., Kostek, B., Ciszewski, T. and Czyżewski, A. (2019). Machine learning-based analysis of English lateral allophones, *International Journal of Applied Mathematics and Computer Science* **29**(2): 393–405, DOI: 10.2478/amcs-2019-0029.

Rybka, J. and Janicki, A. (2013). Comparison of speaker dependent and speaker independent emotion recognition, *International Journal of Applied Mathematics and Computer Science* **23**(4): 797–808, DOI: 10.2478/amcs-2013-0060.

Saba, J.N. and Hansen, J.H. (2022). The effects of Lombard perturbation on speech intelligibility in noise for normal hearing and cochlear implant listeners, *Journal of the Acoustical Society of America* **151**(2): 1007–1021.

Schedl, M., Gómez, E. and Urbano, J. (2014). Music information retrieval: Recent developments and applications, *Foundations and Trends® in Information Retrieval* **8**(2–3): 127–261.

Smailis, C., Sarafianos, N., Giannakopoulos, T. and Perantonis, S. (2016). Fusing active orientation models and mid-term audio features for automatic depression estimation, *Proceedings of the 9th ACM International Conference on Pervasive Technologies Related to Assistive Environments, Corfu, Greece*, pp. 1–4.

Stathopoulos, E.T., Huber, J.E., Richardson, K., Kamphaus, J., DeCicco, D., Darling, M., Fulcher, K. and Sussman, J.E. (2014). Increased vocal intensity due to the Lombard effect in speakers with Parkinson's disease: Simultaneous laryngeal and respiratory strategies, *Journal of Communication Disorders* **48**: 1–17.

Summers, W.V., Pisoni, D.B., Bernacki, R.H., Pedlow, R.I. and Stokes, M.A. (1988). Effects of noise on speech

production: Acoustic and perceptual analyses, *Journal of the Acoustical Society of America* **84**(3): 917–928.

Tsardoulias, E., Thallas, A.G., Symeonidis, A.L. and Mitkas, P.A. (2016). Improving multilingual interaction for consumer robots through signal enhancement in multichannel speech, *Journal of the Audio Engineering Society* **64**(7/8): 514–524.

Vlaj, D. and Kacic, Z. (2011). The influence of Lombard effect on speech recognition, *in* I. Ipšić (Ed), *Speech Technologies*, INTECH Open Access Publisher, London, pp. 151–168.

Wang, S., Wei, Y., Long, K., Zeng, X. and Zheng, M. (2018). Image super-resolution via self-similarity learning and conformal sparse representation, *IEEE Access* **6**: 68277–68287.

Wei, I.-C., Wu, C.-W. and Su, L. (2019). Generating structured drum pattern using variational autoencoder and self-similarity matrix, *20th International Society for Music Information Retrieval Conference, Delft, The Netherlands*, pp. 847–854.

Zhang, S., Li, X., Zong, M., Zhu, X. and Wang, R. (2017). Efficient kNN classification with different numbers of nearest neighbors, *IEEE Transactions on Neural Networks and Learning Systems* **29**(5): 1774–1785.

Gražina Korvel received her BS degree in mathematics and her MS degree in informatics from Vilnius Pedagogical University (currently Vytautas Magnus University Education Academy), Lithuania, in 2007 and 2009, respectively, and her PhD degree from the Institute of Data Science and Digital Technologies, Vilnius University, in 2013. She is currently a senior researcher with the Institute of Data Science and Digital Technologies. Her research interests include speech signal processing, natural language processing, development of mathematical models, applications of soft computing, and computational intelligence. She has published more than 30 papers. Since 2022 she has been a member of the Young Academy of the Lithuanian Academy of Sciences.

Povilas Treigys is a professor at the Faculty of Mathematics and Informatics of Vilnius University. He is a principal researcher and the head of the Signal and Image Analysis Group at the Institute of Data Science and Digital Technologies. His interests include image analysis, detection and object feature extraction in image processing, automated image objects segmentation, optimization methods, artificial neural networks, and software engineering. He is an author of more than 70 journal and conference articles.

Krzysztof Kąkol received his MS degree in sound engineering from the Gdańsk University of Technology in 2001, and his PhD in 2023. He has been working for many years as a software engineer, system analyst, developer and solution architect. His recent employer is PGS SOFTWARE—a Polish software house. He works there as a solutions architect and manager. His research and commercial interests include data pipelines, data analysis and processing, and data science, especially connected with neural networks.

Bożena Kostek is a professor in the Faculty of Electronics, Telecommunications and Informatics at the Gdańsk University of Technology (GUT), Poland. She is a corresponding member of the Polish Academy of Sciences and a fellow of the Audio Engineering Society (AES) and the Acoustical Society of America. Her main scientific interest include machine learning-based speech and music processing, music information retrieval, and cognitive processing. Professor Kostek is an author of more than 600 scientific papers. She has also published four books related to multimedia applications. She is the recipient of many prestigious awards for research, including those of the Prime Minister of Poland (twice), the Ministry of Science (twice) and the Polish Academy of Sciences. She has supervised more than 300 master theses and 20 doctoral theses. She has also led a number of research projects.

Received: 5 August 2022
Revised: 4 December 2022
Re-revised: 9 January 2023
Accepted: 3 March 2023