JAISCR, 2024, Vol. 14, No. 1, pp. 5 – 23
10.2478/jaiscr-2024-0001

INTERPRETING CONVOLUTIONAL LAYERS IN DNN MODEL BASED ON TIME–FREQUENCY REPRESENTATION OF EMOTIONAL SPEECH

Lukasz Smietanka* and Tomasz Maka

Faculty of Computer Science and Information Technology
West Pomeranian University of Technology, Szczecin
Zolnierska 52, 71-210, Szczecin, Poland

**E-mail: lsmietanka@zut.edu.pl*

Submitted: 14th June 2023; Accepted: 15th November 2023

Abstract

The paper describes the relations of speech signal representation in the layers of the convolutional neural network. Using activation maps determined by the Grad-CAM algorithm, energy distribution in the time–frequency space and their relationship with prosodic properties of the considered emotional utterances have been analysed. After preliminary experiments with the expressive speech classification task, we have selected the CQT-96 time–frequency representation. Also, we have used a custom CNN architecture with three convolutional layers in the main experimental phase of the study. Based on the performed analysis, we show the relationship between activation levels and changes in the voiced parts of the fundamental frequency trajectories. As a result, the relationships between the individual activation maps, energy distribution, and fundamental frequency trajectories for six emotional states were described. The results show that the convolutional neural network in the learning process uses similar fragments from time–frequency representation, which are also related to the prosodic properties of emotional speech utterances. We also analysed the relations of the obtained activation maps with time-domain envelopes. It allowed observing the importance of the speech signals energy in classifying individual emotional states. Finally, we compared the energy distribution of the CQT representation in relation to the regions' energy overlapping with masks of individual emotional states. In the result, we obtained information on the variability of energy distributions in the selected signal representation speech for particular emotions.

Keywords: convolutional neural networks, emotion recognition, audio features analysis, explainable machine learning

1 Introduction

Expressive speech, as the ability to express our feelings, is an essential component of effective interpersonal communication. The emotional state in the spoken statement can improve the interaction process in dialogue systems and may have a wide range of applications in voice-based human-computer interaction systems. Also, the prosodic content of the expressive speech strongly relates to the speaker's personality, and it can be used in multimodal biometric systems. In its basic form, the automatic classification of the expressive speech system is built based on a typical machine learning approach. In such a case, from a speech signal, a feature extraction phase is performed to create feature space, and then classic machine learning methods are used to generate a model and perform the classification. However, nowadays, such a process is replaced by the deep learning paradigm. This approach leads to higher classification efficiency compared to traditional machine learning techniques. The main factor causing this effect is unsupervised learning, where data attributes are determined automatically in the learning process.

The problem with this approach is that various neural network architectures lead to different representations in feature space and, hence to other discriminant properties for the same data source. A change in network architecture is needed to increase the classification accuracy and lead to a different signal representation. Another aspect related to this approach is the difficulty of defining dependencies of the internal representations of individual layers of the neural network and their relationship with the physical properties of the source signals. In such a case, analysing such systems is challenging in terms of optimisation, new input data, and robustness to acquisition conditions to keep or increase the classification effectiveness. On the other side, there is a lack of confidence in the results obtained with the use of neural networks because there are difficulties in finding relationships with the physical properties of the analysed phenomena and the possibility of explaining interactions between real objects.

Additionally, it is difficult to determine the system's effectiveness in the delivery of incomplete or distorted data input. There are also difficulties in determining the situation based on whose system makes decisions based on the impact of each attribute on the classification score. In actual conditions, interpreting the machine learning system is often necessary to increase confidence in his behaviour. The degree of interpretability of the model and the resulting classification accuracy compromise in a situation where the constraints resulting from functionalities are introduced into the machine learning model. Additionally, the functionality of deep neural networks is dependent on many nonlinear relationships between attributes in the feature space, which makes the explanation process of their mechanisms is difficult.

The main properties of speech signals include prosody and the speaker's anatomical properties based on fundamental frequency and resonances of the vocal tract. The temporal dynamics of the fundamental frequency constitute the basic information about the emotional load in spoken utterances. Therefore, this study presents the relationships between the layers of the convolutional networks and prosodic changes in emotional statements based on the dynamics of fundamental frequency trajectory.

Recently, a problem of interpretability [1] of the mechanisms leading to high classification accuracy obtained by the deep neural networks (DNN) appears more and more often in literature. The most intuitive approach to understanding how the neural network works is the visualization of feature layers in the network [2]. Using the proposed visualization technique, there is a possibility to tune the network architecture and diagnose possible performance issues. An analysis of the latent space to decrease computation cost and to determine its influence on the model is presented in [3]. The authors used the clustering of emotions in the latent space to analyse model behaviour. In [4], authors proposed a method to modify the convolutional neural network to an interpretable version, where each filter in layers represents a defined object part. In that way, the network automatically assigns each filter to an object part of the learning process and results in a better understanding of the results obtained in the learning process. Authors in [5] proposed interpretability measures for speech sentences and showed their qualitative and quantitative effectiveness. The analysis showed the connections between hidden unit vectors and prosody features by grouping them into different classes. Another interpretation technique applied to the urban sound classification task is presented in [6]. The authors used two audio representations to analyse DNN with layer-wise relevance propagation. The result determines the frequency content assigned with high relevance in feature sets and characterizes the high discriminative information. An interpretable group convolutional neural network (IG-CNN) was proposed in [7]. The presented mechanism is based on the separation of the learning processes of interpretable representation and autonomous representations. The proposed model outperforms the baseline of several popular datasets with emotional speech. A method to visualize and interpret intermediate layers in convolutional neural networks trained on raw speech signals is presented in [8]. The analysis of internal representations was performed using two architectures: WaveGan and ciwGAN. The interpretation concerned three acoustic properties of speech: periodic vibration, aperiodic vibration, and silence. As a result, the proposed visualization and interpretation approach for layers in the neural network make it possible to determine speech signals' prosodic properties. Also, using GAN-based architecture, the same authors in [9] show how to use the proposed method to perform unsupervised acoustic word classification. Because the information contained in individual layers of the convolutional neural network and architecture itself plays an essential role in data classification performance, an effect of depth and width on the analysis of the learned representations are presented in work [10]. In the other work [11], the authors present a multimodal architecture for emotion recognition, which contains two inputs for visual modality and speech as a raw waveform. As a part of the research, the authors compare cell activations from the proposed model and prosodic features. The results show that the network in the training process picks the information about energy, loudness, and fundamental frequency. Another work [12] proposes a set of techniques to interpret and visualize intermediate layers in generative CNNs trained on raw speech data. As a result, the authors show what properties of speech, like intensity or fundamental frequency, are encoded in the individual layers.

In this work, we have compared the time–frequency space selected by activation maps with the fundamental frequency trajectory dynamics and time-domain envelopes of speech signals. The key contributions of this paper include:

(I) A new technique for interpreting and visualizing knowledge of convolutional neural network;

(II) Use of the proposed technique to find out what information the convolutional neural network takes into account while classifying emotions;

(III) An illustration of the relationship between the fundamental frequency of speech signal and the decision process of convolutional neural network.

The rest of the paper is organized as follows. The methods of the proposed work, architecture of the neural network, and audio data used in experiments are explained in Section 2. Section 3 and its subsections contained descriptions of the performed experiments and an analysis of the results. Finally, the results are concluded in section 4.

2 Methodology

In the analysis of the structure of the convolutional layers, the number of parameter variables has been deliberately limited to reduce variability in the generated activation maps for time–frequency structures. These limitations are influenced by the choice of a simple network architecture and the method of dividing the source data. An essential element of this study is generating activation maps, which can be divided into several stages presented in Figure 1. In the first step, the selected time–frequency representation is generated for all recordings from the emotional speech database. Then, the audio data set is divided into two groups, which are processed independently in further stages. The first group includes representations of recordings from men, while the second group is women. Then, these groups are divided into three subsets: training, validation, and testing. This division is performed taking into account emotional states and speakers. Data from each speaker are grouped according to emotional states. Then, each resulting subgroup is divided into proportions of 50%, 25%, and 25% for the training, validation, and test sets, respectively. As a result, each of the three subsets contains data in equal proportions representing particular emotions and speakers. In the next stage, the proposed convolutional network architecture models are prepared and trained based on train and validation subsets. At this stage, due to the high randomness of operations in the training process, resulting, among other things, from the *Dropout* operation and use of GPU, ten different models are generated $(K_1, K_2, \ldots, K_{10})$ each separately for data from women and men. The scope of the obtained classification efficiency of the generated models for test subsets is shown in Figure 2. Based on the generated models and the selected method, activation maps (Grad-CAM), activation maps G are generated for individual data in the test subsets. Then, two groups of emotional

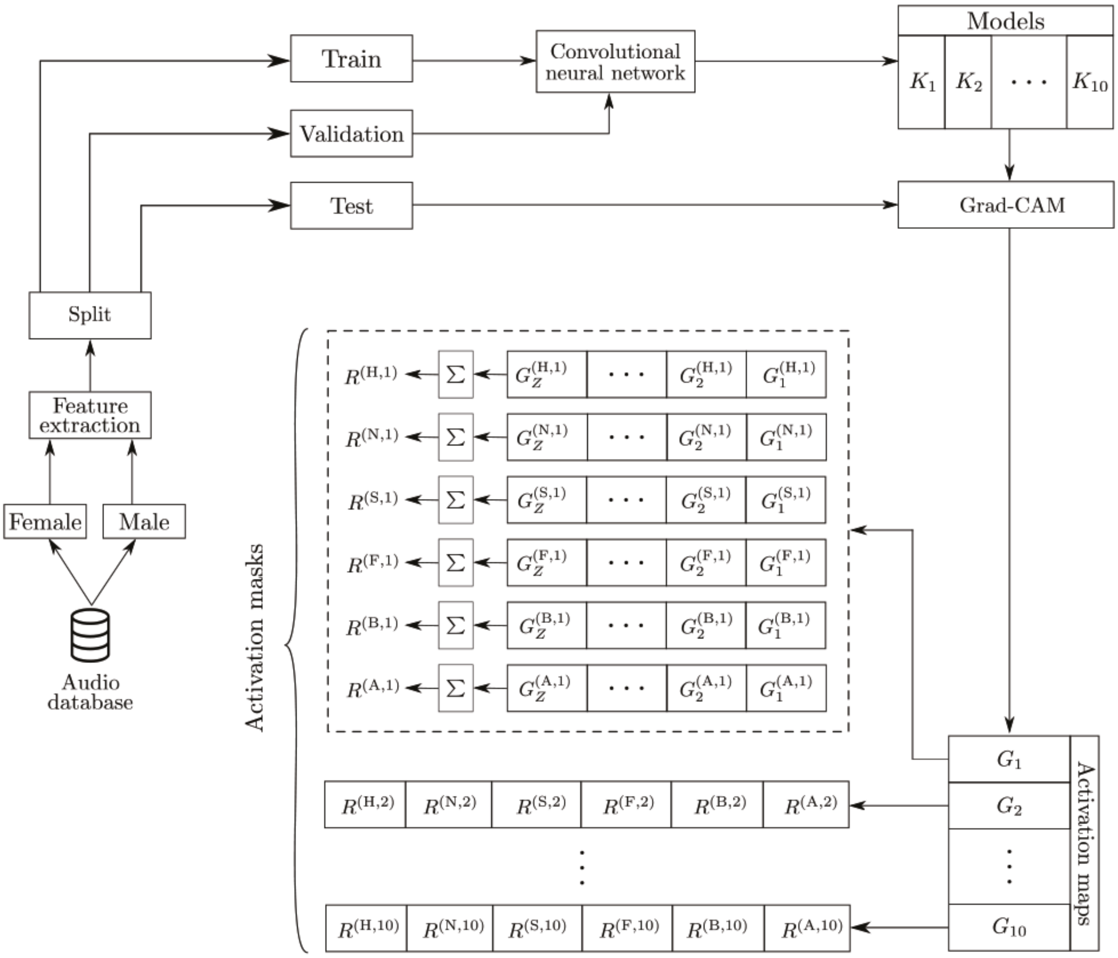

Train
Validation
Test
Convolutional neural network
Models
K_1 K_2 … K_{10}
Grad-CAM
Split
Feature extraction
Female
Male
Audio database
Activation masks
$R^{(H,1)}$ Σ $G_Z^{(H,1)}$ … $G_2^{(H,1)}$ $G_1^{(H,1)}$
$R^{(N,1)}$ Σ $G_Z^{(N,1)}$ … $G_2^{(N,1)}$ $G_1^{(N,1)}$
$R^{(S,1)}$ Σ $G_Z^{(S,1)}$ … $G_2^{(S,1)}$ $G_1^{(S,1)}$
$R^{(F,1)}$ Σ $G_Z^{(F,1)}$ … $G_2^{(F,1)}$ $G_1^{(F,1)}$
$R^{(B,1)}$ Σ $G_Z^{(B,1)}$ … $G_2^{(B,1)}$ $G_1^{(B,1)}$
$R^{(A,1)}$ Σ $G_Z^{(A,1)}$ … $G_2^{(A,1)}$ $G_1^{(A,1)}$
$R^{(H,2)}$ $R^{(N,2)}$ $R^{(S,2)}$ $R^{(F,2)}$ $R^{(B,2)}$ $R^{(A,2)}$
$R^{(H,10)}$ $R^{(N,10)}$ $R^{(S,10)}$ $R^{(F,10)}$ $R^{(B,10)}$ $R^{(A,10)}$
G_1
G_2
G_{10}
Activation maps

Figure 1. Mechanism for determining activation maps and generalized activation masks.

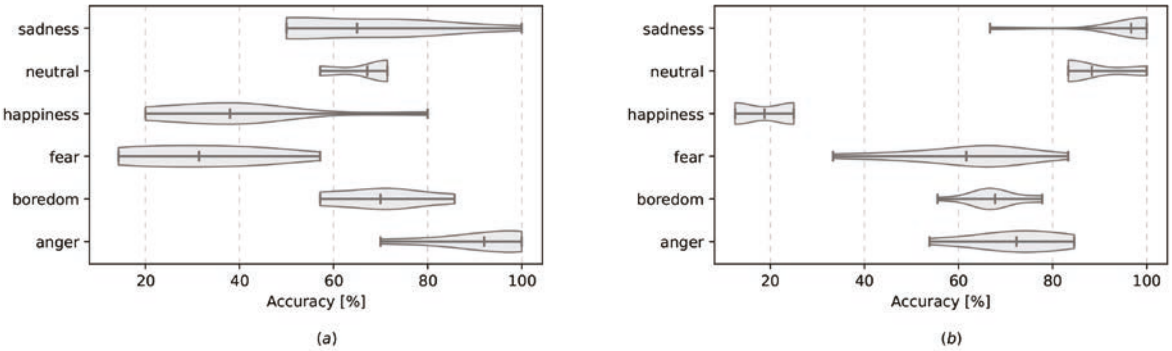


Figure 2. Classification accuracy for all models: females (a), males (b).

masks are formed based on the obtained activation maps. In the first group, masks R are determined as a sum of individual map activations from the model with the highest classification level. This sum is realized according to the following equation Eq. 1:

$$R^{(L,k)} = \sum_{n=1}^{Z} G_n^{(L,k)}, \tag{1}$$

where R and G are mask and activation maps generated based on the model k, L is an emotional state (A - *anger*, B - *boredom*, F - *fear*, N - *neutral*, H - *happiness* and S - *sadness*), Z is the number of samples in a test subset of a given emotion that are correctly classified. The second group of masks $\hat{R}$ is generated based on the sum of masks of individual emotion from all models according to Eq. 2:

$$\hat{R}^{(L)} = \sum_{k=1}^{10} R^{(L,k)}. \tag{2}$$

In Figure 3, example masks for selected emotional states are depicted, wherein the top row contains masks with the first group R, and in the bottom, masks from the second group $\hat{R}$.

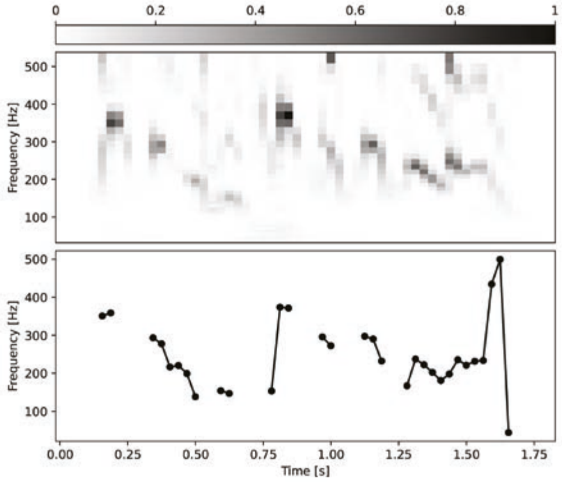


Figure 5. Examples of audio features: CQT-96 (top panel), fundamental frequency F_0 contour (bottom panel).

As can be seen, the masks from both groups are very similar. Therefore, for further research, we decided to use only activation maps and masks from models with the highest levels of classification for women and men, their confusion matrices are shown in Figure 4.

2.1 Audio features

According to our previous work [13], where we tested the popular representations of audio signals with a group of several neural networks, we have selected a constant-Q spectrogram with 96 bins (CQT-96). We have chosen it because we obtained the best classification performance for this representation and two datasets with emotional sentences [14, 15]. For the constant-Q spectrogram, the frequency distribution is geometric and the ratio of the band's centre frequency to its width is constant. As a result, the resulting frequency scale has a different accuracy reproduction in the low and medium frequency range compared to Mel scale [16]. Although such a representation was initially proposed for reproducing the Western musical scale [17], the resolution spectrum in the frequency range up to 4kHz of several Hz can also be used for speech signals. For this reason, using features of emotional speech built based on the CQT spectrogram leads to higher classification accuracy than using the Mel spectrogram [13].

To describe the representation obtained in the neural network learning process for emotional speech signals, we decided to use prosody information by calculating the variability of fundamental frequency (F_0) for spoken sentences and comparing it with time–frequency information marked by activation maps. In Figure 5 (top panel) a CQT-96 spectrogram is depicted for example emotional sentence where in Figure 5 (bottom panel) the trajectory of fundamental frequency (F_0) is shown The estimation of F_0 values was performed using the algorithm presented in [18].

2.2 Class activation map

Process of determining activation maps for individual samples using the Grad-CAM [19] algorithm can be broken down into four smaller stages. In the beginning, the **A** feature matrix is retrieved from the selected layer of the convolutional model for a given sample and the output vector **y** of the model. We used a features matrix generated based on the last convolution layer of the model because, as the authors of the Grad-CAM method assume, this layer may be the source of the most important information obtained by the model in the training process [19]. The three-dimensional matrix **A** has

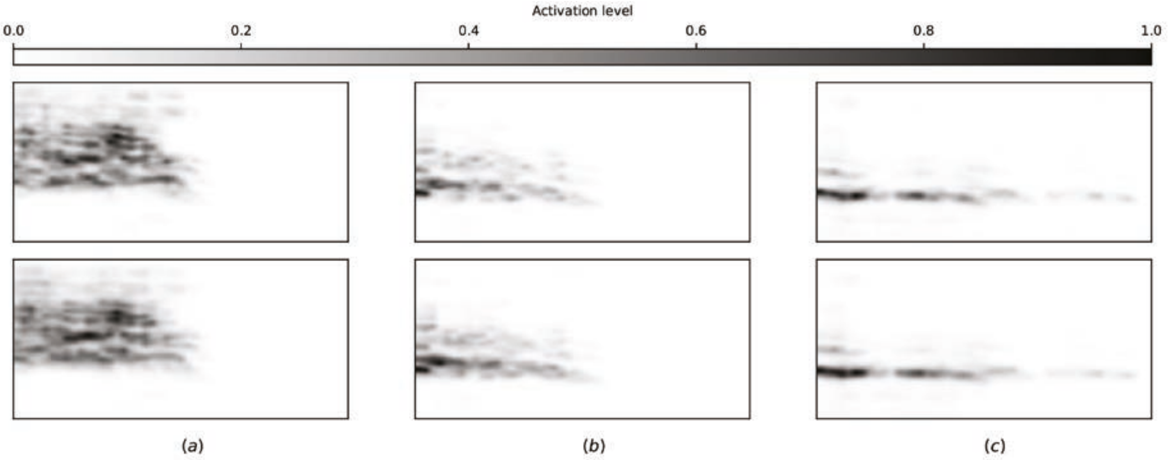


Figure 3. Examples of calculated masks for three emotional states: *anger* (a), *neutral* (b) and *sadness* (c). Top row contains masks for the best model, bottom row the sum of masks for all models.

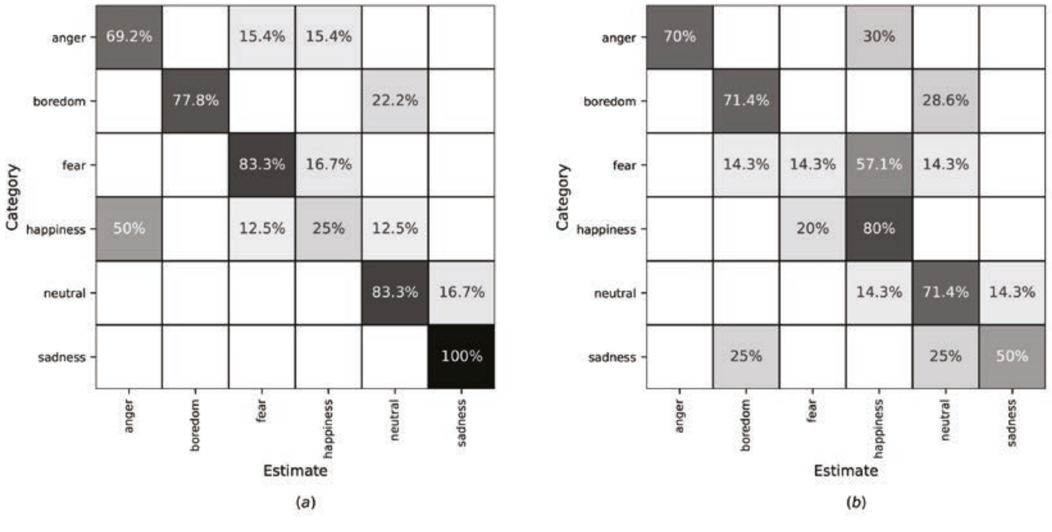


Figure 4. Confusion matrices for the best models: females (a), males (b).

size $H \times W \times K$. The H denotes height, W is width, and K is the number of channels. In contrast, the output of the $\mathbf{y}$ model is a vector of length corresponding to the number of classes. The next step is determining the K gradients for each feature channel relative to the selected c class. In this case, c was the class of individual samples. This operation can be written as follows:

$$\nabla_k = \frac{\delta \mathbf{y}_c}{\delta \mathbf{A}^{(k)}}, \tag{3}$$

where k are channel numbers, $\mathbf{y}_c$ is the model output for the c class, and $\mathbf{A}^{(k)}$ is the feature map from channel k. In our experiments, we performed this operation using the *GradientTape* algorithm from the *TensorFlow* library [20]. The resulting gradients have the same size as feature maps, $H \times W$. Then, the weights for individual channels α_k are determined by averaging the gradient values across rows and columns. As a result, a vector of length K is computed whose values represent the mean gradient intensity for individual channels. The process for determining the coefficients describes the following formula:

$$\alpha_k = \frac{1}{Z} \sum_{i=1}^{H} \sum_{j=1}^{W} \nabla_{i,j,k}, \tag{4}$$

where the value of Z equals the product of H and W. In the last step, the final class activation map $\mathbf{G}$ is determined. It is obtained by applying a weighted sum feature map $\mathbf{A}$ through channels, where coefficients α_k are the weights of each k channel. The result is a matrix with dimensions of $H \times W$, which is computed by applying ReLU [21] operation, which zeroes values less than zero. This step can be described by the formula 5:

$$\mathbf{G} = \mathrm{ReLU}\left(\sum_{k=1}^{K} \alpha_k \cdot \mathbf{A}^{(k)} \right). \tag{5}$$

The matrix $\mathbf{G}$ contains activation levels that define the importance of the input data fragments in the classification process. An example of the result obtained with the Grad-CAM algorithm is presented in Figure 6(a). The diagram shows an example activation map obtained with the Grad-CAM algorithm while Figure 6(b) shows this low -dimension latent space mapped to the time–frequency domain corresponding to CQT representation using linear interpolation. As a result, it enables us to visualize the importance of time–frequency components of CQT representation in the classification process.

3 Experiments

In this section, we presented a description and the results of the performed experiments[1]. Sections 3.1 and 3.2 contain information about the audio database and CNN architecture we chose. In section 3.3, we presented the central part of our experiments, describing the methods we use to analyse the activation maps.

3.1 Audio data

In our study, we used an audio database called Berlin database of emotional speech [14]. We chose it for the following reasons. Firstly, this database is often used in research on emotional speech [22, 23] and is widely available. Secondly, as presented by the authors of [24], this database enables the highest efficiency of classification (88.47% [25]) compared to other databases, such as RAVDESS (87.5% [15, 26]) or IEMOCAP (75.60% [27, 28]). However, unlike the TESS [29] database, which is classified with an accuracy of 99.6% [30] contains longer utterances, not just words with the same prefix. Furthermore, the results of our earlier experiments [13] show that using the selected database, we were able to achieve the best classification efficiency with popular architectures (*DenseNet, ResNet, Inception, MobileNet*).

The chosen database contains 535 sentences in the German language, recorded as monophonic, with a sample rate equal to 16 kHz. The utterances are spoken by ten professional actors, including five males aged 25, 26, 30, 31, and 32 years old, and five females aged 21, 31, 32, 34, and 35 years old. For the sake of the uneven distribution of the samples, we decided to remove the examples of *disgust* emotional state and the samples from the actors with ID equal to 09 and 12. The sentences were removed because they had too few recordings for some emotional states, which prevented them from being evenly divided into training, testing, and validation subsets. As a result, 417 recordings were left in the database. Table 1 shows the distribution of recordings and the number of samples of individual emotions and speakers after reduction.

[1] The complete results are available on `https://github.com/staticvoice/convcam`

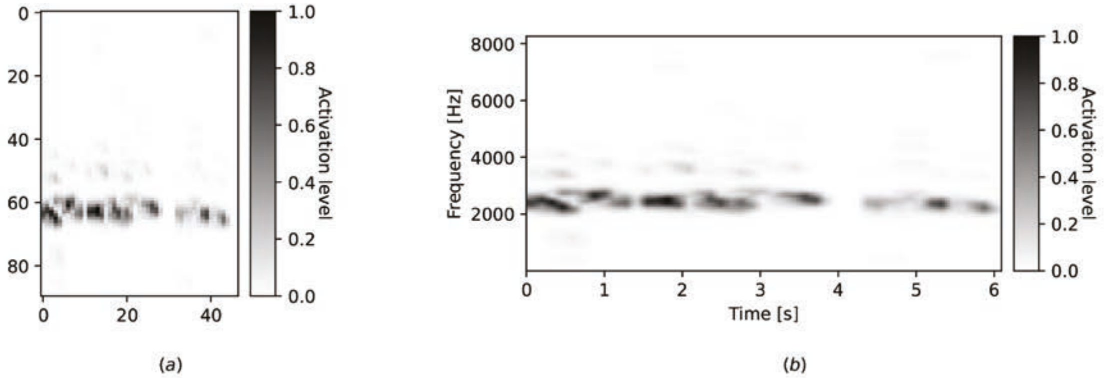


Figure 6. Examples of activation maps: raw activation map (a), scaled version to the size of original CQT-96 representation size (b).

Table 1. Distribution of samples in the balanced data set.

Speaker ID	Sadness	Fear	Happiness	Neutral	Boredom	Anger	Total
08	6	6	11	10	10	12	55
03	7	4	7	11	5	14	48
14	10	12	8	7	8	16	61
11	7	10	8	9	8	11	53
10	3	8	4	4	8	10	37
13	5	7	10	9	10	12	53
16	8	7	11	5	14	14	59
15	4	8	6	11	9	13	51
Total	50	62	65	66	72	102	417

3.2 CNN architecture

Popular convolutional network architectures consist of many convolutional layers, which require many arithmetical operations. It influences the number of processed features and the computational complexity. To reduce the number of calculations, various operations reducing the number of parameters are used, such as *MaxPooling* [31]. However, such reductions cause maps of the last features of the convolutional layers of these architectures to have a tiny size, which affects the size of activation maps generated on their basis. For mapping activation maps, e.g., 3×6 to representation CQTs for which 96×202 were generated, requires to use of interpolation of frequency axis of 3 values to 96, which introduces inaccuracies in their representation. We decided to use a simple convolutional network Figure 7 where we used three *2D Convolutional* layers to relate the results of their operations to a two-dimensional time–frequency plane. These layers contain 64, 128, and 256 filters, with ReLU activation function and kernel size equal to 3.

After each of them, there is a layer of *MaxPooling*, where the first two reduce the size of feature maps in only one dimension that corresponds to the timeline, while the third, on the other hand, reduces the number of features in both axes. The use of such a scheme of reducing the number of features makes it possible to obtain a map activation from the last convolutional layer of size 90×47, which allows for a more accurate reference to the audio representation than would be in the case of the use of popular convolutional networks. Table 2 shows the size of the activation maps of popular convolutional networks compared to the network used in this work. The convolutional and max pooling layers are followed by *Flatten* layer, *Dense* layer with 128 units, ReLU activation function, and a *Dropout* layer with a drop rate of 0.5.

The proposed architecture has the first linear layer with 33914880 parameters compared to the preceding layer (295168 parameters), creating a bottleneck problem. We decided on such a solution because our study concerns the analysis of the decision process of convolutional layers. Adding excessive operations after the last convolutional layer for data reduction could distort the influence of knowledge from the convolution layers in the decision-making process.

The last *Dense* layer has the number of units corresponding to the number of classes and activation function *Softmax*. As an optimizer, we have used *ADAM* algorithm with a learning factor equal to 0.001 and a *cross-entropy* as a loss function.

3.3 Activation map analysis

A set of experiments was carried out to determine the relationships between the obtained activation maps with the physical properties of the speech signal.

In this study, we used features such as the fundamental frequency, time domain envelope and energy distribution in the selected CQT representation to analyze the obtained activation maps. The time-domain envelope of a speech signal is one of the most basic components in the speech signal perception process [36]. The temporal envelope of the signal also reflects prosodic features such as speaking rate or intonation [37]. F0 plays a fundamental role in the prosody of a given utterance [38] and the perception of speech signals [39]. In addition, the frequency ranges of segments containing voiced parts of speech signals convey the speaker's characteristics [40], such as gender, age, and health [41]. Moreover, the variability of these segments and the lengths between voiced and unvoiced parts are one of the sources of the nature of the utterance [42]. Both the envelope of the signal in the time domain and the fundamental frequency variability in the time domain reflects the basic properties of the speech signal and its time–frequency structure. Therefore, these parameters have been used as a reference element when interpreting the layers of convolutional neural networks trained for emotional speech signals. In section 3.3.1, as a result of the importance analysis of speech signals in the classification process, we presented the relations between the obtained activation maps and the time domain envelopes for the considered recordings. Then, in section 3.3.2, to check the association of F_0 with emotional states in the classification process, we compare the energy in F_0 segments of recordings with activation level corresponding to fragments of their activation map on the time axis. In the last part of the experiments (section 3.3.3), to investigate the variability of the energy of the *CQT* representation for individual emotions, we presented an analysis of the distribution of its

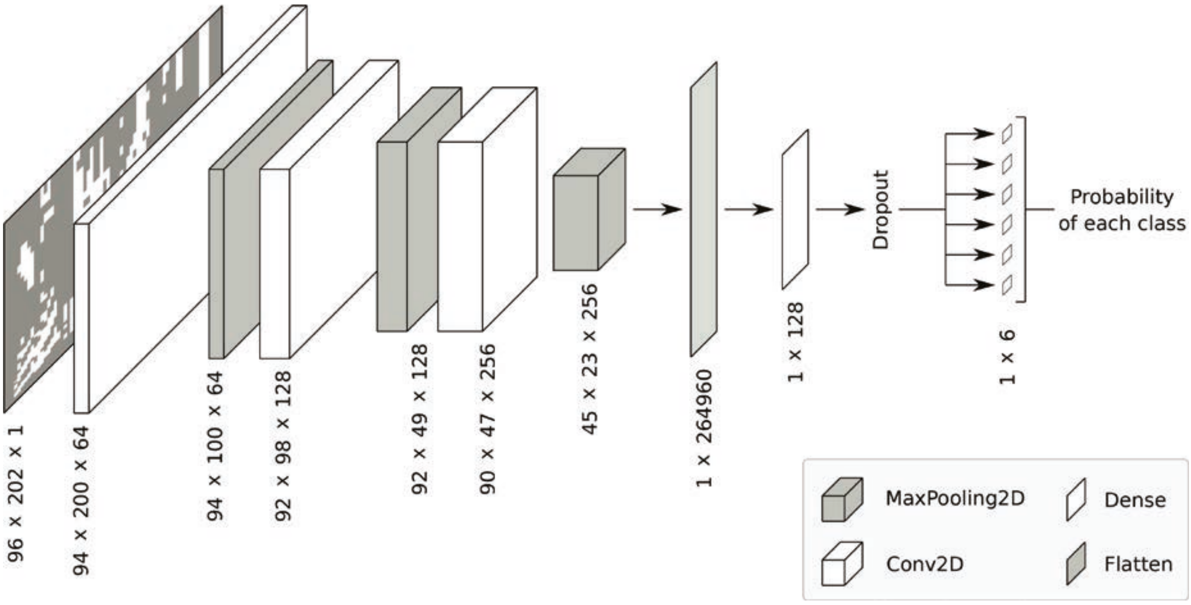


Figure 7. Architecture of proposed *SimpleNet* convolutional neural network.

Table 2. Shapes of the last convolutional layers of selected CNN architectures.

	SimpleNet	**VGG16** [32]	**ResNet50** [33]	**MobileNet** [34]	**InceptionV3** [35]
Shape	90×47	6×12	3×7	3×6	1×4

total energy in relation to the energy of the areas overlapping with the generated masks of individual emotional states.

3.3.1 Time envelope analysis

To analyse the obtained matrices of the activation maps in the time domain, we reduced them to a one-dimensional vector by summing its rows. Finally, for each activation map $H \times W$, the vector $\mathbf{q}$ of length W was obtained.

$$\mathbf{q}_j = \sum_{i=1}^{H} \mathbf{G}_{ij}, \tag{6}$$

where: $j = 1,\ldots,W$ and $i = 1,\ldots,H$ denotes columns and rows of activation matrix $\mathbf{G}$ respectively. Then, to be able to compare the obtained activation map to the source signal for individual recordings, we have determined the energy contour $\mathbf{Q}$ in the time-domain using the following formula:

$$\mathbf{Q}_n = \sum_{k=0}^{K-1} x(k+n\cdot K)^2, \tag{7}$$

where x represents the input speech signal containing N samples, $n = 0,\ldots,W-1$, $K = \left\lfloor \frac{N}{W} \right\rfloor$. The frame length used in computing the energy contour was equal to the size of the frame used in determining the *CQT* representation. The obtained contour vectors $\mathbf{Q}$ consisted of the same number of elements as the corresponding summed activation maps $\mathbf{q}$. In such a case, it is possible to use the RMSE measure to compare them:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{n=0}^{N-1}(x_n - y_n)^2}, \tag{8}$$

where: x_n and y_n are values of compared vectors, and N is the length of both vectors. We decided to use RMSE measure because, for identical vectors, the value is equal to zero. Its value is higher as the differences between elements of vectors are higher, and it measures the vectors in the same units. Before calculate the RMSE, the $\mathbf{q}$ and $\mathbf{Q}$ vectors were normalized. Cases where RMSE was the lowest and the highest are shown in the figure 8. The top row shows female recordings (left) and male (right), for which the RMSE value was the lowest. Their activation contour $\mathbf{q}$ and the $\mathbf{Q}$ envelope were the closest to each other. The bottom row shows the cases with the highest RMSE for which the activa-

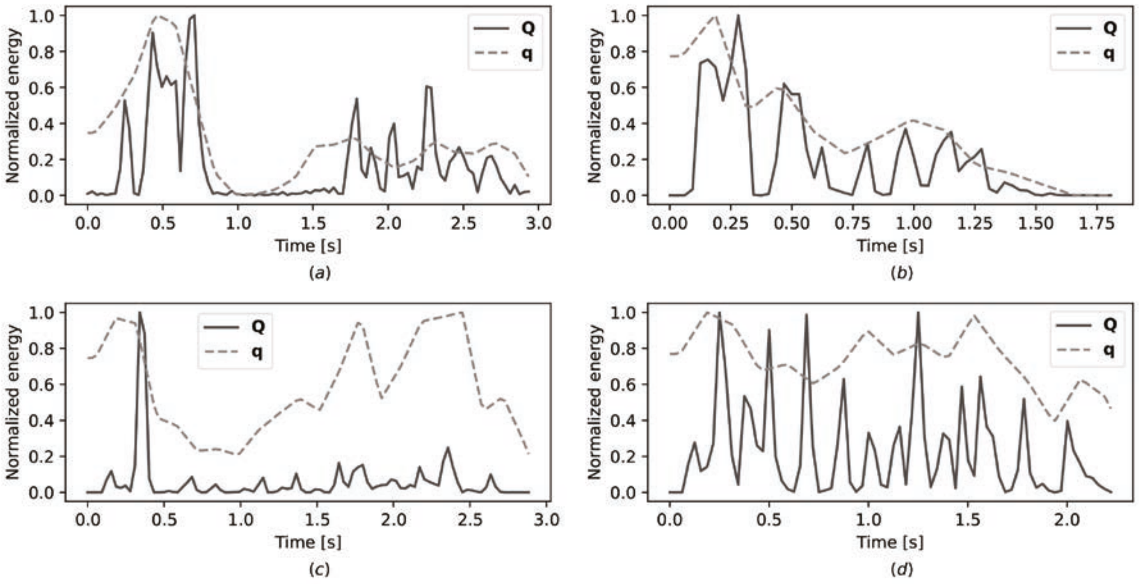


Figure 8. Comparison of the time envelopes and activation trajectories with the lowest RMSE values: females in *sadness* state (a), males in *neutral* state (b) and with the highest RMSE values: females in *anger* state (c), males in *neutral* state (d).

Table 3. Statistical properties of the RMSE values between time envelope **Q** and the activation contour **q** for all emotional states and genders.

		Anger	**Boredom**	**Fear**	**Happiness**	**Neutral**	**Sadness**
Female	**min**	0.086	0.079	0.078	0.108	0.061	0.060
	max	0.347	0.140	0.291	0.181	0.248	0.137
	mean	0.204	0.109	0.189	0.144	0.156	0.104
	range	0.261	0.061	0.212	0.073	0.187	0.076
Male	**min**	0.142	0.101	0.290	0.215	0.084	0.228
	max	0.299	0.194	0.290	0.326	0.333	0.243
	mean	0.205	0.143	0.290	0.271	0.194	0.235
	range	0.157	0.093	0.0	0.110	0.249	0.014

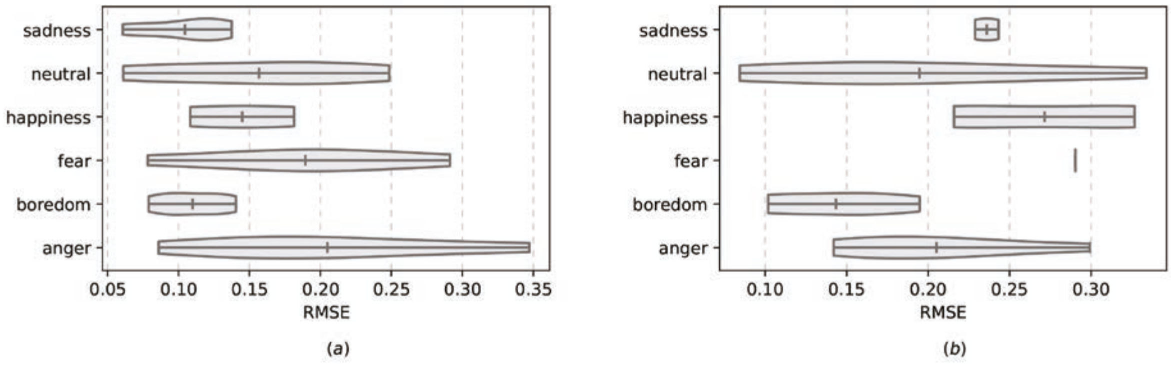


Figure 9. RMSE for all correctly classified samples between envelope **Q** and **q**: females (a), and males (b).

tion contour **q** and the **Q** envelope differed the most. The lowest RMSE values were obtained for recordings representing the *sadness* (Figure 8a) emotional state for females equal to 0.06, and for males in *neutral* (Figure 8b) state the RMSE value was equal to 0.08. On the other hand, the highest RMSE values belonging to female sentence in *anger* state (Figure 8c) and equal to 0.34. In the case of males, the RMSE was equal to 0.33 for *sadness* (Figure 8d). One of the reasons for high RMSE values may be the low time resolution of the representation of activation maps **q**, which is due to the architecture of the used network and the scaling of the activation map to the size of the input CQT representation. An example of this case is illustrated in Figure 8d, which shows a relatively complex envelope of a recording whose large number of peaks negatively affects the RMSE measure. The table 3 shows the full list of RMSE values for all emotional states and both genders. Additionally, Figure 9 illustrates the distribution of the RMSE measure for all analysed recordings. As can be seen in the case of females, the emotional state for which the RMSE value was the lowest globally was *sadness* and *boredom*, which also achieved the lowest RMSE values in male recordings. These results may show that a significant part of the information from the convolutional network with the above-mentioned emotional states occurs in the temporal representation of the speech signal.

3.3.2 Dominant frequency trajectory in the activation map

In the next step, we developed the method presented in Figure 10 for interpretation of the activation maps, which allowed us to analyse them in the frequency domain. This method works based on the fundamental frequency and enables a direct comparison of the activation map of a given recording with its F_0 trajectory. The entire process can be performed in the following five steps:

1. The selection of the columns of the activation matrix Figure 10 (middle panel) corresponding to the successive segments of F_0 and being consistent in time with it as depicted in Figure 10 (top panel).

2. Replace values in selected fragments below half of the maximum value in a given column of the activation matrix with zeros. As a result, this will eliminate the „non-essential" elements of the activation map. An example of the effect of this operation is shown in

3. Figure 10 (middle panel), where the remaining non -zero values are shown in grey.

4. The corresponding frequencies are selected for each column of the activation matrix where they have non-zero values. Consequently, for each column in the selected segments, a vector is created containing as many frequency values as there were in the given column with non-zero values.

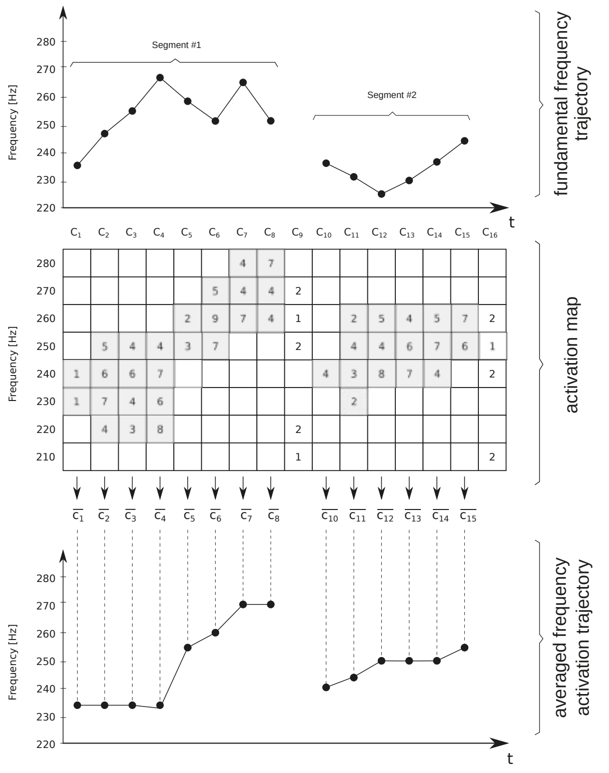
5. For each vector, its average is calculated, which gives the average frequency value for each of the selected columns. This operation is represented by the formula:

$$\overline{\mathbf{c}_j} = \frac{1}{M}\sum_{i=1}^{M} R(\mathbf{G}_{ij}), \qquad (9)$$

   where $j = 1,\ldots,W$, $\overline{\mathbf{c}_j}$ denotes averaged frequency of column j, M is the number of rows containing non-zero values, $\mathbf{G}$ is the activation matrix whereas $R(\cdot)$ is the function which returns the frequency related to the cell of the activation map.

6. In the last step, the frequency trajectory of the activation map is created from the averaged values as depicted in Figure 10 (bottom graph) and which can be compared with the F_0 trajectory.

The resulting frequency trajectory from the activation map has the same lengths contour F_0 and is compared with it by calculating the RMSE measure. The Figure 11 illustrates the determined mean values in consecutive F_0 segments and the activation trajectory obtained using approach shown in Figure 10 for cases with the lowest RMSE value. Moreover, the Tab. 4 shows the lowest RMSE value for individual emotional states grouped by gender. In addition, the distribution of RMSE values is shown in Figure 12 for all analysed recordings.


Segment #1
Segment #2
Frequency [Hz]
280
270
260
250
240
230
220
210
t
fundamental frequency trajectory
activation map
averaged frequency activation trajectory
C_1 C_2 C_3 C_4 C_5 C_6 C_7 C_8 C_9 C_{10} C_{11} C_{12} C_{13} C_{14} C_{15} C_{16}
$\overline{C_1}$ $\overline{C_2}$ $\overline{C_3}$ $\overline{C_4}$ $\overline{C_5}$ $\overline{C_6}$ $\overline{C_7}$ $\overline{C_8}$ $\overline{C_{10}}$ $\overline{C_{11}}$ $\overline{C_{12}}$ $\overline{C_{13}}$ $\overline{C_{14}}$ $\overline{C_{15}}$

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}
280							4	7								
270						5	4	4	2							
260					2	9	7	4	1		2	5	4	5	7	2
250		5	4	4	3	7			2		4	4	6	7	6	1
240	1	6	6	7						4	3	8	7	4		2
230	1	7	4	6							2					
220		4	3	8					2							
210									1							2

Figure 10. Scheme of determining the frequency trajectory based on the activation map.

As depicted in the Figure 11, in the case of *boredom* emotion, the trajectory of F_0, and the trajectory determined from the map activation practically coincide for recordings that have obtained the lowest RMSE for both females and males. According to the table 4, the differences between them changed from 7Hz to 10Hz. The Figure 12 shows that globally, *boredom* state also achieved low values of RMSE. This case may indicate that the fundamental frequency contains much relevant information about this emotional state. The situation is similar for *sadness* and *neutral* states for recordings which have the lowest RMSE value as shown in Figure 11. Except for minimal deviations in individual segments, the trajectories mostly coincide for female and male recordings. In the Tab. 4, one can also see that their differences are much smaller than that of the states *anger*, *fear*, and *happiness*. Whereas on the Figure 12 can be seen that RMSE values are in low ranges. In the case of *happiness* state, the differences are much more significant for both female and male recordings. For best cases presented in Figure 11 trajectories differ quite noticeably from each other, and their differences are vary from 100Hz to 150Hz as shown in Tab. 4. Also, RMSE values for all recordings representing *happiness* emotional state have a pretty large spread, as depicted in Figure 12, which may indicate that the fundamental frequency much fewer influences this emotion. The situation is similar for the female recordings for the *fear* emotional state. On the other hand, the analysed set contained only one recording for males, which cannot be the basis for any conclusions. The most significant differences occurred for the emotional state *anger*. How can be seen in Figure 11, for recordings with the lowest RMSE values, the differences are much greater than in the other emotional states and range from 158Hz to 439Hz as shown in Tab. 4. In the case of males, the compared trajectories are practically opposite. In the global perspective, as shown in Figure 12, the distribution of the coefficients RMSE values for all recordings of *anger* is more varied and is high as 700Hz. It follows that the fundamental frequency is a fairly poor source of information for *anger* emotional state. The estimation errors of fundamental frequency F_0 also have a negative impact on the comparison process. The errors depend on the estimation method, quality of speech recordings, and F_0 estimation method. We observed them in cases with a clear difference confined to individual segments for *fear* and *neutral* states in the case of female speakers, as depicted in Figure 11. An example of such errors is shown in Figure 13, where octave errors were the cause of wrongly estimated F_0 values for several signal frames.

3.3.3 Activation maps energy distribution analysis

In the next stage, we analysed the energies of input CQT representations in relation to their fragments distinguished by the convolutional neural network due to its importance in the classification process. We conducted the analysis using masks of emotional states. In the first step, for individual recordings, we determined the energy E of them *CQT* representation according to the formula:

$$E = \sum_{i=1}^{H} \sum_{j=1}^{W} \mathbf{C}_{i,j}^2, \qquad (10)$$

where $\mathbf{C}$ denotes the CQT representation. Next, we transformed the masks into a binary form $\overline{\mathbf{P}}$ with a threshold equal $\beta = 1/10$ of the maximum value, which allowed the elimination of the least signifi-

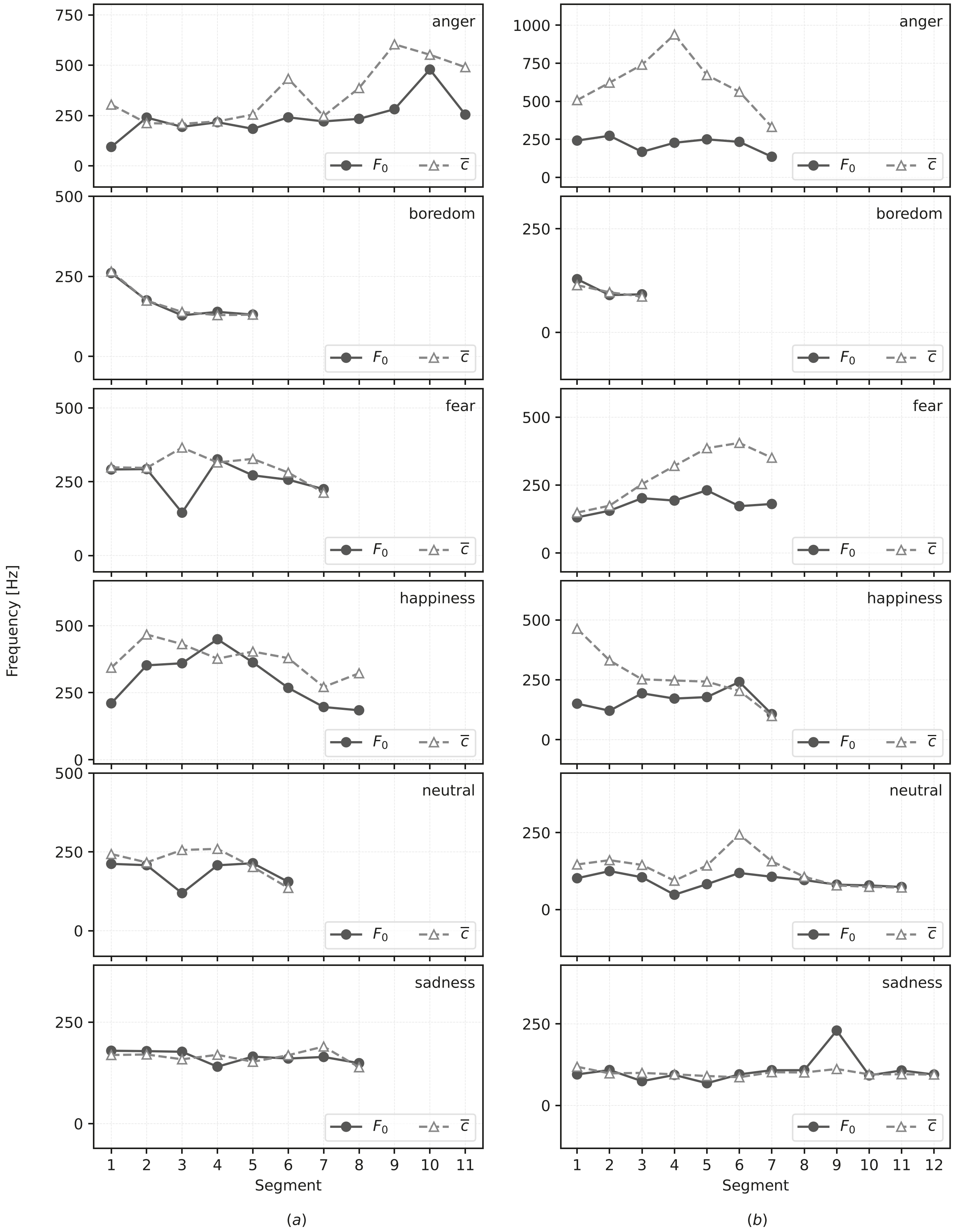


Figure 11. The lowest RMSE between F_0 and $\mathbf{\bar{c}}$ frequency trajectories: females (a), males (b).

Table 4. Statistical properties of the RMSE values between F_0 and $\bar{\mathbf{c}}$ for all emotional states and genders.

		Anger	Boredom	Fear	Happiness	Neutral	Sadness
Female	**min**	158.44	6.98	86.76	99.86	61.94	17.22
	max	1145.22	285.07	654.63	397.00	181.30	171.47
	mean	533.10	91.62	249.22	248.43	116.56	110.43
	range	986.77	278.09	567.87	297.13	119.36	154.25
Male	**min**	439.40	9.78	134.71	149.65	51.31	36.51
	max	820.42	87.75	134.71	382.33	144.16	61.72
	mean	604.01	34.70	134.71	232.55	93.41	49.12
	range	381.01	77.97	0.0	232.67	92.84	25.21

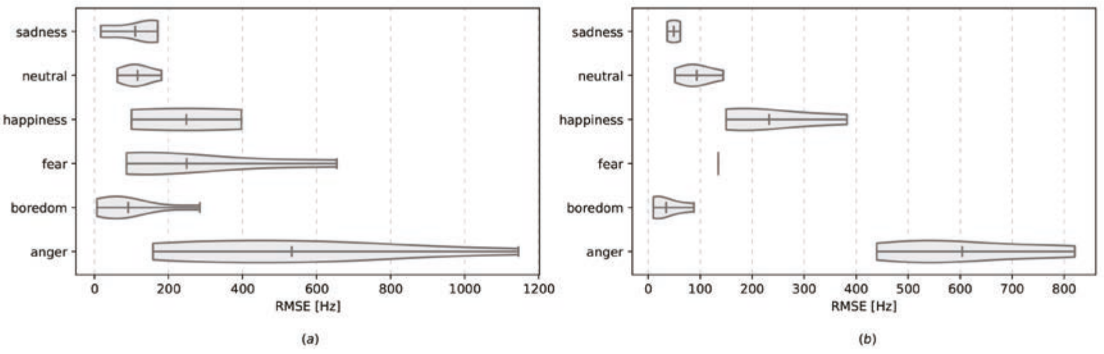


Figure 12. RMSE values for all recordings between F_0 and $\bar{\mathbf{c}}$ in data set: females (a), and males (b).

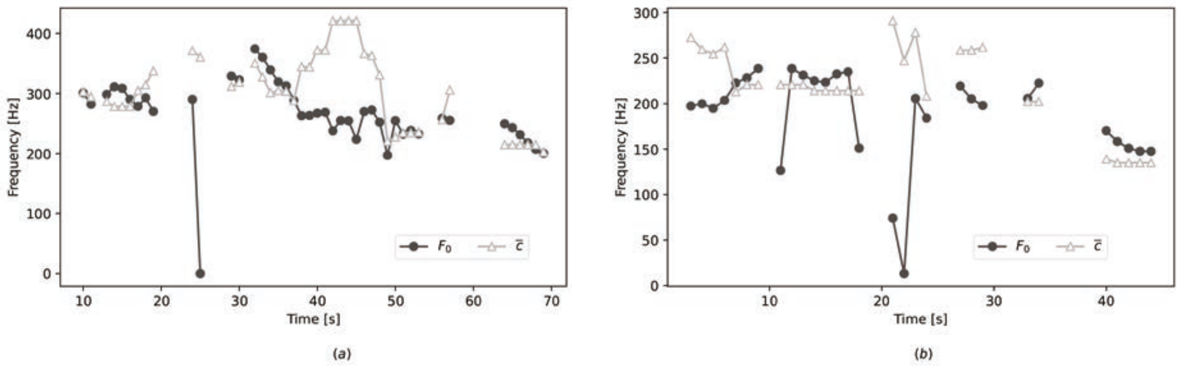


Figure 13. Example F_0 trajectories with errors: *fear* (13b01Ab) (a), *neutral* (13a01Nb) (b), both for females.

Table 5. Statistical properties of the $\hat{E}$ ratio for individual emotional states and genders.

		Anger	Boredom	Fear	Happiness	Neutral	Sadness
Female	**min**	88.9%	81.1%	89.7%	89%	87.8%	81.0%
	max	96.6%	96.5%	96.4%	92.4%	93.7%	94.4%
	mean	93.9%	91%	93.7%	90.7%	91.2	88.5%
	range	7.7%	15.4%	6.7%	3.4%	5.9%	13.4%
Male	**min**	85.1%	76.1%	93%	90.1%	85.7%	90%
	max	96.6%	93.1%	93%	95.5%	94.3%	90.3%
	mean	92.4%	88.9%	93%	93.3%	90.4%	90.2%
	range	11.5%	17%	0%	5.4%	8.6%	0.3%

cant values:

$$\overline{\mathbf{P}_{i,j}} = \begin{cases} 0, & \text{if } \mathbf{R}_{i,j} < \max(\mathbf{R}) \cdot \beta \\ 1, & \text{otherwise} \end{cases}, \quad (11)$$

where $j = 1, \ldots, W$, $i = 1, \ldots, H$, $\mathbf{R}$ is a matrix representing a mask of a selected emotional state. In the next step, the energy $\overline{E}$ of the CQT representation was determined again for individual recordings. But this time for fragments consistent with individual masks, which are obtained from the product of $\overline{\mathbf{P}}$ and $\mathbf{C}$. Next, we calculated the ratio $\hat{E}$ of the obtained energies $\overline{E}$ to the total energy E expressed as a percentage. The obtained results for individual variants are presented in Tab. 5, which shows the minimum values, maximum, average and range of $\hat{E}$ ratios for each individual emotions grouped by gender. As can be seen, based on the average values that are around 90% for all emotional states and genders, the convolutional neural network mainly focused on areas of CQT representations containing the highest energy. As a result, these areas are the primary carrier of information about emotional states. On the other hand, the lowest mean in the case of females was obtained by *sadness*, and in the case of males, it was *boredom*. Moreover, the highest mean of $\hat{E}$ was for *anger* state for females and *happiness* for males.

In the case of the range to which there were recordings of individual emotions, then, for females, the lowest difference between the minimum and maximum value was equal to 5.9% for emotion *neutral*. For males, it was *happiness* state with a value of 5.4%. This case may indicate the stability of the distribution of essential areas within the recordings of these emotional states. In contrast, the highest differences were found for *boredom* state for both females and males, which were equal to 15.4% and 17% respectively. Such a situation may be due to the wide variety of CQT energy groups for recording them. Recordings representing *happiness* for females and *fear* and *sadness* for males were omitted from the analysis due to the low number of correctly classified samples (below three samples). For the *anger* state, for females, the $\hat{E}$ coefficients of the recordings belonging to the speaker with an ID of 16 were significantly different from other recordings. When omitted, the range drops from 7.7% to 4%. For males, also the two samples underestimate the values. However, they belong to actors with IDs 13 and 15, the remaining samples in the upper range. This dispersion may result from the statement's content with the ID b03, which both represent. For the *boredom* state, the two items have a lower $\hat{E}$ value. However, they belong to the speaker with the ID 08. Without them, the range drops from 15.4% to 5.3% for *boredom* state. For males, only one sample is very different from the others and belongs to the actor with an ID equal to 10. Without it, the distribution changed from 17% to 2.2%. For *fear* emotional state in case females, there are no significant differences For males, the data comes from only one recording, so any useful observations cannot be performed. For *happiness* emotion, in the case of females, the analysis was omitted due to the small number of recordings. For males, no significant deviations were observed. The only interesting situation is for the 10 and 03 actors, for which the energy ratios are almost identical for different statements. There are no individual differences for females for *neutral* state. For males, the 15 actor's samples have significantly lower $\hat{E}$ values than the three samples belonging to three different actors, whose range is 2.5%. For *sadness* emotional state, in female's recordings, the two samples belonging to speaker 14 and 16 have much smaller coefficients $\hat{E}$ and represent other statements. Additionally, the remaining recordings of these speakers are in the upper range of values, so there is no rule in this case. Due to the observed and described deviations, in the case of females, the lowest the range of energy ratio changes falls to *anger* emotional state while for males it is *boredom* state.

4 Conclusions

The paper presents the results of determining the representation of individual layers of convolutional networks connected with emotional content. Experiments have been performed to specify the similarities between activation maps and the low-level properties of source speech signals. The database containing utterances with emotional content was divided into two groups according to the gender of the speaker. A few studies were carried out separately due to differences in the vocal tract and, thus, in the fundamental frequency. For every gender type, the database has been further divided into three sets (training, testing, and validation), considering a balanced division into speakers and emotional states. A dedicated convolutional network architecture was proposed, characterised by a minimised number of pooling operations to maintain the adequate resolution of generated activation maps from the last convolutional layer of the network. The emotional states were classified using the proposed network to determine the quality of the generated activation maps. Activation maps were generated for all samples from the test set. Several experiments were carried out using the algorithm proposed in the study concerning the similarity of energy distributions in activation maps with the dynamics of changes in fundamental frequency. As a result of the conducted experiments we showed that the convolutional neural network considers the prosodic features of the emotional utterance in the learning process and energy distribution related to voiced parts of speech contained in the analysed spoken sentences. Additionally, a comparative analysis of the energy distribution in the activation maps with time-domain envelopes were carried out. The results of this analysis showed significant conformity of the energy distribution in the activation maps, that is, the significance of the fragments of the input representation with the recording envelope for emotions such as *boredom* and *sadness* than in other emotions. From this, it can be concluded that a considerable part of the information about these emotional states can be obtained from the temporal structure of the speech signal. At the last stage of the experiments, we analysed the ratio of the total energy of the *CQT* representation to the energy in its fragments, which, from the network's point of view, turned out to be the most important in the classification process. As a result, it turned out that in classifying the emotional states, the neural network mainly uses the areas of the input speech signal representation, which contains clusters of time–frequency units with the highest energy. The resulting activation maps provide information about the regions of the time–frequency representation that were the source of important information about individual emotional states in the classification process. A detailed analysis of these areas can be used to generate a dedicated feature space characterizing individual emotions, which may improve the classification process. Moreover, detailing time–frequency ranges of given emotions can help to understand what physical features of the speech signal determine the emotional properties of the voice. In future work, we plan to use the proposed algorithm for interpreting recursive and transformer neural networks. In addition, similar studies will be performed using others time–frequency representations.

References

[1] A. Karim, A. Mishra, M. H. Newton, and A. Sattar, Machine learning interpretability: A science rather than a tool, vol. abs/1807.06722, 2018.

[2] M. D. Zeiler and R. Fergus, Visualizing and understanding convolutional networks, in Computer Vision – ECCV 2014, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 818–833.

[3] S. Das, N. N. Lønfeldt, A. K. Pagsberg, and L. H. Clemmensen, Towards interpretable and transferable speech emotion recognition: Latent representation based analysis of features, methods and corpora, 2021.

[4] Q. Zhang, Y. N. Wu, and S.-C. Zhu, Interpretable convolutional neural networks, in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2018-06.

[5] K. V. V. Girish, S. Konjeti, and J. Vepa, Interpretabilty of speech emotion recognition modelled using self-supervised speech and text pre-trained embeddings, in Proc. Interspeech 2022, 2022, pp. 4496–4500.

[6] M. Colussi and S. Ntalampiras, Interpreting deep urban sound classification using layer-wise relevance propagation, CoRR, vol. abs/2111.10235, 2021.

[7] E. Jing, Y. Liu, Y. Chai, J. Sun, S. Samtani, Y. Jiang, and Y. Qian, A deep interpretable representation learning method for speech emotion recognition, Information Processing and Management, vol. 60, no. 6, p. 103501, 2023.

[8] G. Beguš and A. Zhou, Interpreting intermediate convolutional layers in unsupervised acoustic word classification, in ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 8207–8211.

[9] G. Begus and A. Zhou, Interpreting intermediate convolutional layers of CNNs trained on raw speech, CoRR, vol. abs/2104.09489, 2021.

[10] T. Nguyen, M. Raghu, and S. Kornblith, Do wide and deep networks learn the same things? uncovering how neural network representations vary with width and depth, in 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021, 2021.

[11] P. Tzirakis, G. Trigeorgis, M. Nicolaou, B. Schuller, and S. Zafeiriou, End-to-end multimodal emotion recognition using deep neural networks, IEEE Journal of Selected Topics in Signal Processing, vol. PP, 04 2017.

[12] G. Begus and A. Zhou, Interpreting intermediate convolutional layers of generative cnns trained on waveforms, IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 30, pp. 3214–3229, 2022.

[13] L. Smietanka and T. Maka, DNN architectures and audio representations comparison for emotional speech classification, in 2021 International Conference on Software, Telecommunications and Computer Networks (SoftCOM). Split, Hvar, Croatia: IEEE, sep 2021.

[14] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, A database of german emotional speech, in in Proceedings of Interspeech, Lissabon, 2005, pp. 1517–1520.

[15] S. R. Livingstone and F. A. Russo, The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north american english, PLOS ONE, 2018.

[16] T. Lidy and A. Schindler, Cqt-based convolutional neural networks for audio scene classification. Budapest, Hungary: DCASE, 09 2016.

[17] J. C. Brown, Calculation of a constant Q spectral transform, The Journal of the Acoustical Society of America, vol. 89, no. 1, pp. 425–434, January 1991.

[18] P. Boersma, Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound, vol. 17, no. 1193, pp. 97–110, 1993.

[19] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618–626.

[20] M. Abadi et al., Tensorflow: Large-scale machine learning on heterogeneous distributed systems, 2016.

[21] A. F. Agarap, Deep learning using rectified linear units (relu), arXiv preprint arXiv:1803.08375, 2018.

[22] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and B. W. Schuller, Multitask learning from augmented auxiliary data for improving speech emotion recognition, IEEE Transactions on Affective Computing, pp. 1–13, 2022.

[23] Y. Liu, H. Sun, W. Guan, Y. Xia, Y. Li, M. Unoki, and Z. Zhao, A discriminative feature representation method based on cascaded attention network with adversarial strategy for speech emotion recognition, IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 1063–1074, 2023.

[24] E. Guizzo, T. Weyde, S. Scardapane, and D. Comminiello, Learning speech emotion representations in the quaternion domain, IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 1200–1212, 2023.

[25] N. T. Pham, D. N. M. Dang, and S. D. Nguyen, Hybrid data augmentation and deep attention-based dilated convolutional-recurrent neural networks for speech emotion recognition, 2021.

[26] Y. L. Bouali, O. B. Ahmed, and S. Mazouzi, Cross-modal learning for audio-visual emotion recognition in acted speech, in 2022 6th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), 2022, pp. 1–6.

[27] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, IEMOCAP: interactive emotional dyadic motion capture database, Language Resources and Evaluation, vol. 42, no. 4, pp. 335–359, Dec. 2008.

[28] S. Kakouros, T. Stafylakis, L. Mošner, and L. Burget, Speech-based emotion recognition with self-supervised models using attentive channel-wise correlations and label smoothing, in ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023, pp. 1–5.

[29] K. Dupuis and M. Kathleen Pichora-Fuller, Recognition of emotional speech for younger and older talkers: Behavioural findings from the toronto emotional speech set, Canadian Acoustics, vol. 39, no. 3, p. 182–183, Sep. 2011.

[30] S. Jothimani and K. Premalatha, Mff-saug: Multi feature fusion with spectrogram augmentation of speech emotion recognition using convolution neural network, Chaos, Solitons & Fractals, vol. 162, p. 112512, 2022.

[31] J. Nagi, F. Ducatelle, G. A. Di Caro, D. Cireşan, U. Meier, A. Giusti, F. Nagi, J. Schmidhuber, and L. M. Gambardella, Max-pooling convolutional neural networks for vision-based hand gesture recognition, in 2011 IEEE International Conference on Signal and Image Processing Applications (ICSIPA), 2011, pp. 342–347.

[32] K. Simonyan and A. Zisserman, Very deep convolutional networks for large-scale image recognition, in International Conference on Learning Representations – ICLR'2015, 2015.

[33] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, in IEEE Conference on Computer Vision and Pattern Recognition – CVPR'2016, 2016, pp. 770–778.

[34] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, Mobilenets: Efficient convolutional neural networks for mobile vision applications, arXiv preprint arXiv:1704.04861, 2017.

[35] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, Rethinking the inception architecture for computer vision, in IEEE Conference on Computer Vision and Pattern Recognition – CVPR'2016, 2016, pp. 2818–2826.

[36] U. Masashi, K. Miho, K. Maori, K. Shunsuke, and A. Masato, How the temporal amplitude envelope of speech contributes to urgency perception, in Proceedings of the 23rd International Congress on Acoustics, ser. Proceedings of the International Congress on Acoustics. Aachen , Germany: International Commission for Acoustics (ICA), 2019, pp. 1739–1744.

[37] P. Ríos-López, M. T. Molnar, M. Lizarazu, and M. Lallier, The role of slow speech amplitude envelope for speech processing and reading development, Frontiers in Psychology, no. 8, 2017.

[38] K. Stevens, Acoustic Phonetics, ser. Current Studies in Linguistics. London: MIT Press, 2000.

[39] N. Hellbernd and D. Sammler, Prosody conveys speaker's intentions: Acoustic cues for speech act perception, Journal of Memory and Language, vol. 88, pp. 70–86, 2016.

[40] S. Pearsell and D. Pape, The effects of different voice qualities on the perceived personality of a speaker, Frontiers in Communication, vol. 7, 2023.

[41] M. Nishio and S. Niimi, Changes in speaking fundamental frequency characteristics with aging, The Japan Journal of Logopedics and Phoniatrics, vol. 46, pp. 136–144, 04 2005.

[42] H. Deng and D. O'Shaughnessy, Voiced-unvoiced-silence speech sound classification based on unsupervised learning, in 2007 IEEE International Conference on Multimedia and Expo, 2007, pp. 176–179.

Łukasz Śmietanka is currently pursuing a Ph.D. at the West Pomeranian University of Technology, Szczecin, Poland. He received an MSc degree in computer science in 2021 at the same university. His research activity over the last years focused on various aspects of speech emotion recognition and machine learning.
https://orcid.org/0000-0002-2038-2136

Tomasz Maka received his Ph.D. degree in computer science from Szczecin University of Technology in 2005. He is currently working as an assistant professor in Faculty of Computer Science and Information Technologies, West Pomeranian University of Technology, Szczecin. His scientific interests include audio analysis, machine hearing and audio feature engineering.
https://orcid.org/0000-0001-5898-2201