



Transport and Telecommunication, 2021, volume 22, no. 4, 471–481
Transport and Telecommunication Institute, Lomonosova 1, Riga, LV-1019, Latvia
DOI 10.2478/ttj-2021-0037

AVIATION PROFILING METHOD BASED ON DEEP LEARNING TECHNOLOGY FOR EMOTION RECOGNITION BY SPEECH SIGNAL

K.T. Koshekov¹, A.A. Savostin², B.K. Seidakhmetov¹, R.K. Anayatova¹, I.O. Fedorov¹

*¹JSC “Civil Aviation Academy”
Almaty, Republic of Kazakhstan, 050039, Akhmetov str., 44
office@agakaz.kz*

*²M. Kozybayev North-Kazakhstan University
Petropavlovsk, Republic of Kazakhstan, 150000, Pushkin str., 86
mail@nkzu.kz*

This paper proposes a method of automatic speaker-independent recognition of human psycho-emotional states by analyzing the speech signal based on Deep Learning technology to solve the problems of aviation profiling. For this purpose, an algorithm to classify seven human psycho-emotional states, including anger, joy, fear, surprise, disgust, sadness, and neutral state was developed. The algorithm is based on the use of Mel-frequency cepstral coefficients and Mel spectrograms as informative features of speech signals audio recordings. These informative features are used to train two deep convolutional neural networks on the generated dataset. The developed classifier testing on a delayed verification dataset showed that the metric for the multiclass fraction of correct answers' accuracy is 0.93. The solution proposed in the paper can be in demand in human-machine interfaces creation, medicine, marketing, and in the field of air transportation.

Keywords: aviation profiling, emotion recognition, speech signal, neural network

1. Introduction

Modern trends in ensuring the safety of air travel require more active study of profiling technologies based on the influence of the psycho-emotional state (PES) of personnel. There is already an urgent need for new effective tools for automatic recognition of a person's PES by his or her voice in the fields of medicine, marketing, safety, worker health control in hazardous industries and transport now.

High-performance PES recognition will raise the quality level of human-machine interfaces used in the aviation industry to a higher level. Machine analysis of voice signals of aviation personnel opens up the possibility of automatic determination of stress and fatigue levels, recognition of depressive states, prevention of fatigue, etc.

The use of automatic PES recognition technology is effective in industries where there is no visual contact between interlocutors, and communication is primarily by voice messaging, in particular. At the same time, PES recognition systems can be implemented in the process of continuous detection and risk factors control to reduce possible sources of danger for human and material values.

Despite the importance and relevance of these problems no general theory has been proposed at this time that allows you to link the PES of speaker with voice signal characteristics. However, high results have been achieved in solving the similar problem of automatic speech recognition. The success achieved in this area has been made possible by the use of modern algorithms for digital signal processing, as well as the development of machine learning (ML) methods (Uday *et al.*, 2019).

This circumstance largely determines the approach proposed in this paper to the development of a new method of automatic classification of human PES by the characteristics of its voice signal.

2. Purpose and objectives of the study

One of the main problems in creating human PES recognition systems is the existing uncertainty and ambiguity in formulating the concept of emotion and, consequently, the theoretical models of their classification. For these reasons, when developing a system for recognizing human PES, it is necessary to focus on a set of seven archetypal emotions: anger, joy, fear, disgust, sadness, surprise and a neutral state

(calmness). Then in the course of their automatic classification the probability of assigning the emotions shown by the speaker to each of the seven listed emotional classes should be calculated. Possessing such quantitative characteristics will allow to give a preliminary estimation of speaker's PES.

It is important to note that the task of automatic classification of human emotions by voice should meet the requirements of the speaker irrespective of gender, age and peculiarities of pronunciation. At the same time, it is necessary to take into account the complex structure of the speech signal and the processes that generate it, the peculiarities of the psychophysical perception of sounds by humans, and the arbitrary sense of the analyzed utterances (Lawrence *et al.*, 2007).

Taking into account the vagueness of the formulated problem of classification and the designated requirements to its implementation, it is advisable to use intellectual methods of analysis based on technology and deep neural networks for the PES detection. With their help it is possible to reveal the hidden patterns in the data, including the existence of some uncertainty from the position of the researcher.

Based on this, the objectives of this work are formulated as follows.

1. Creation of a representative emotional corpus of speech signals for seven archetypal types of emotions in English language speech, distinguished by a variety of used utterances, composition of speakers, variability of the degree of emotion manifestation, etc.

2. Development of a process of speech signal digital preprocessing (preprocessing) for selecting informative features.

3. Development of a method of speech signal classification for automatic speaker PES detection without the necessity to recognize the semantic load of the analyzed utterances.

4. Evaluation of the developed method effectiveness of speaker's automatic classification of PES.

English is chosen according to considerations of its prevalence in the world and the widespread practice of international use in voice communication. The absence of the need to decode the speech of the speaker has a number of significant advantages. Among them, the reduction of the complexity of the method of analysis, the possibility of classification by short fragments of speech, the work only by voice intonation in the conditions of speech illegibility, as well as the uniformity in the analysis of other languages. The latter circumstance is confirmed by the existence of a clear similarity in the manifestation of emotions in conversation in arbitrary speakers, as evidenced by studies of evolutionary biology (Cornelius, 1996).

Based on the above, the aim of this study is to develop a method for automatic recognition of speaker emotion from his speech signal by means of data mining.

3. Data collection and preprocessing

3.1. Formation of an emotional corpus of speech signals

To create a model of the PES classifier based on the ML algorithms, it is required to form a representative training base of sound recordings of speech signals with different emotional coloring (emotional corpus). In order to reduce the time of the research and to comply with ethical issues, during the formation of the training emotional corpus there was applied the approach generally accepted in this kind of tasks on the use of sound recordings of model bases, formed with the participation of professional actors.

The analysis of emotional corporuses being on open access allowed us to select the following bases of records, satisfying the objectives of the study:

- The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) (Livingstone *et al.*, 2018);
- Surrey Audio-Visual Expressed Emotion (SAVEE) (Haq *et al.*, 2009);
- Toronto Emotional Speech Set (TESS) (Pichora-Fuller *et al.*, 2020);
- Emotional speech database (ESD) (Kun *et al.*, 2021).

The RAVDESS corpus is an audio-visual database of emotional speech and songs. From the entire corpus, audio recordings of emotional speech were selected from 24 professional actors (12 women and 12 men) who voiced two lexically selected utterances in a neutral North American accent. Speech included happiness, fear, anger, sadness, disgust, surprise, calm, and neutral state.

Each speech pattern is uttered at two levels of emotional intensity: normal and strong.

The SAVEE database consists of entries from 4 male actors for 7 emotions: happiness, fear, anger, sadness, disgust, surprise, and neutral state. A total of 480 utterances in British English. The sentences were selected from the standard TIMIT corpus (Garofolo *et al.*, 1993) and phonetically balanced for each emotion.

In the TESS recordings database, the information is presented as a set of 200 target words uttered in the carrier phrase “Say the word ...” by two actresses (aged 26 and 64) in the seven emotional states mentioned above. The total number of recordings is 2,800. Both actresses speak English.

From the ESD corpus, recordings for 10 native English speakers uttering 350 different utterances were selected. There are 5 types of emotions in the base: neutral state, happiness, anger, sadness and surprise. A training dataset was generated from the listed databases, the structure of which is shown in Table 1 and Figure 1.

Table 1 Audio recordings distribution by emotion class in the training dataset

Database	RAVDESS		SAVEE		TESS		ESD		Total
Emotion type	male	female	male	female	male	female	male	female	
angry	96	96	60	0	0	400	1750	1750	4152
disgusted	96	96	60	0	0	400	0	0	652
fearful	96	96	60	0	0	400	0	0	652
happy	96	96	60	0	0	400	1750	1750	4152
neutral	48	48	60	0	0	400	1750	1750	4056
sad	96	96	60	0	0	400	1750	1750	4152
surprised	96	96	60	0	0	400	1750	1750	4152
Overall	624	624	420	0	0	2800	8750	8750	21968
	1248		420		2800		17500		

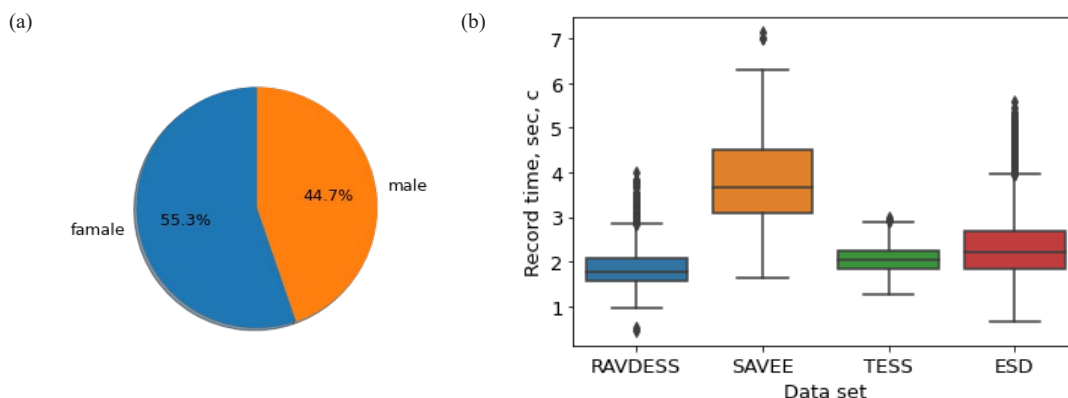


Figure 1. Formed emotional corpus characteristics: (a) ratio of male and female speakers; (b) record duration statistics by dataset

The prepared data set contains recordings of seven types of PES of English-speaking speakers. The audio recordings differ significantly in the form of emotion expression, speaker composition and pronunciation variants, providing the required representativeness. This structure of the emotion corpus will facilitate a more varied representation of samples for training the classifier model.

3.2. Development of a digital speech signal preprocessing process

The main task of the preprocessing stage of automatic emotion recognition is to select informative features of the objects under study. The informative features are used as input parameters for the classification model. However, before extracting the informative features over the speech signal data, a number of important transformations must be performed. These transformations must ensure that the speech signal with the lowest noise component is extracted from the audio recording and that long pauses in the analyzed utterances are removed.

Figure 2 shows a generalized structure of the proposed preprocessing process.

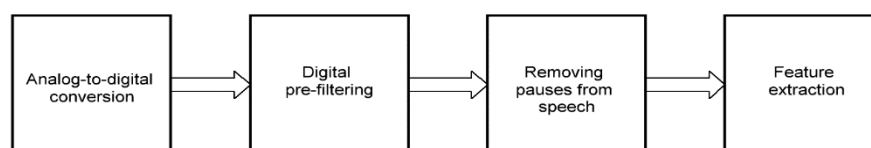


Figure 2. Structure of the information conversion process at the preprocessing stage

In accordance with Figure 2 to unify the process of obtaining digital recordings of speech signals at the stage of analog-to-digital conversion is set sampling rate widespread format AudioCD with a value of $f_s = 22050$ Hz and the quantization bit rate of 16 bits. On the electronic media the data are stored without compression in *.wav format.

The preliminary digital filtering stage is intended to remove noise components from the signals. In this case, the type and characteristics of the applied digital filters depend on the level and nature of the interference affecting the signal. Since in this work specially prepared quality sound recordings are used, the digital filtering of signals at this stage was not applied.

In the spontaneous speech signal obtained for the task of automatic emotion recognition, there will always be a large number of pauses of various lengths. Long pauses in the speaker's speech will not contain useful information about the emotional coloring of the utterance. Possessing the information about the moment of the beginning and the end of the phrase will allow reducing significantly the number of arithmetic operations over the counts of the studied signal. Therefore, at the preprocessing stage (Figure 2) it is necessary to remove pauses of duration about 1 second and more from speech records. This will allow you to preserve the integrity of the sequence of speech beats in phrases and sentences, but at the same time remove long empty sections in the sound recording.

The low level of background noise in the signals from the training set allows us to use a simplified procedure to remove pauses from the speech of speakers. In particular, a simple threshold procedure is used, where the absolute value $|x(n)|$ of the speech signal $x(n)$ is compared with a given threshold h , on the basis of which the decision is made whether the current signal value belongs to either a pause in speech, or the utterance.

To stabilize the threshold procedure better, we use as the analyzed signal in our work the signal $\hat{x}(n) = h(k) \bullet |x(n)|$, where $h(k)$ is the impulse response of the moving average filter with the window size $W=0,1f_s$ and the transfer characteristic of the form:

$$H_{MA}(z) = \frac{1}{W} \sum_{k=1}^W z^{k-\frac{W}{2}}. \quad (1)$$

As a result of applying the threshold procedure, we will obtain at the output of the conversion the desired signal s , the samples of which will correspond to the samples $\hat{x}(n)$ taken when the condition $\hat{x}(n) > h$ is met, as shown in Figure 3.

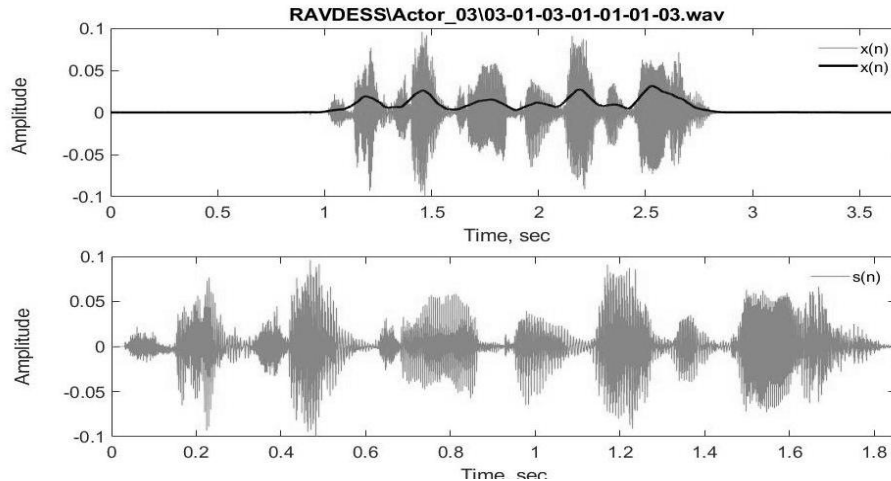


Figure 3. The process of extracting fragments of speech from the audio signal

The described procedure was applied to all available samples from the generated training dataset.

3.3. Feature extraction

An important feature of a speech signal is its statistical quasi-stationarity for short time intervals. For this reason, when working with speech signals it is advantageous to use short-term analysis (Lawrence *et al.*, 2007), the meaning of which consists in the study of separate short sections of a sound recording, called frames. Frame size is determined by the size of the window, which is sequentially

shifted along the signal under study at equal time intervals. During the formation of frames, an overlap between neighboring sections is organized to compensate distortions at the edges of the processed fragments, as shown in Figure 4.

In this work we used frames of length 512 samples (about 23 ms) with an overlap of 35 % of this length.

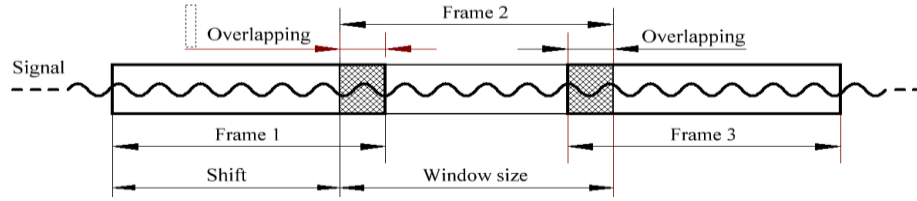


Figure 4. The process of forming frames for the investigated speech signal

In the short-term analysis frames of the speech signal can be considered as a sequence of separate sounds with different properties. The result of the processing of each frame is the selection of a set of features for the sound contained in it. As a result, the short-term analysis calculates certain attributes for each frame, which can act as some time-dependent informative characteristic of the speech signal as a whole.

The most common variant of the short-term analysis in digital signal processing is the windowed discrete Fourier transform (FWDFT) (Jont, 1977). However, for the study of speech signals and emotional state recognition it is ineffective to use information about the spectral composition of the signal obtained by the FWDFT. It is necessary to take into account the subjective nature of the psychophysical features of the perception of sounds by the human hearing organs.

Taking this into account, as a result of the analysis the following informative signs for classification of PES were allocated.

1. *Mel spectrogram*. It is known that the frequency of a sound signal is related to a subjective characteristic known as sound pitch. Comparison of frequency values with the sound height value allows to construct a spectral characteristic of a sound signal, where a chalk-scale will be used instead of the frequency axis (Zwicker *et al.*, 1990). This scale was obtained on the basis of statistical processing of a large amount of data on subjective perception of sound frequencies by humans and can be approximated by an equation of the form:

$$B(f) = 1125 \ln(1 + f / 700). \quad (2)$$

An estimate of the power spectrum of a sound signal based on the mel scale is called a Mel spectrogram (*melspec*) and can be calculated according to the formula:

$$P_i(m) = \sum_{k=0}^{N-1} (S_i(k))^2 H_m(k), \quad 0 \leq m < M, \quad (3)$$

where $S_i(k)$ – fast discrete Fourier transform (FDFT) for speech signal s for of the i -th frame, obtained in the short-term analysis; $N = 2^j$, $j \in \mathbb{N}$ – the number of analyzed signal frame samples; $H_m(k)$ – data from $M = 128$ special triangular.

In equation (3) $S_i(k)$ is calculated according to the formula

$$S_i(k) = \sum_{n=0}^{N-1} s_i(n) w(n) e^{-j2\pi kn/N}, \quad (4)$$

where $w(n)$ is Hamming Window function used to reduce the FDFT leakage at the finite duration interval; k – FDFT frequency domain index.

2. *Mel-frequency cepstral coefficients*. Informative signs based on Mel-frequency Cepstral Coefficient (MFCC) are also based on the use of the Mel-frequency scale. In contrast to the Mel spectrogram, MFCCs are less sensitive to the peculiarities of the speaker's speech path and significantly reduce the space of individual characteristics of the speech signal:

$$c_i(l) = \sum_{m=0}^{M-1} 10 \log P_i(m) \cdot \cos \left(\left(\pi l \left(m + \frac{1}{2} \right) / M \right) \right), \quad 0 \leq l < M. \quad (5)$$

After calculating $MFCC_i(l)$, at $M = 40$, the coefficient $ci(0)$ is discarded, since it does not carry information about the speaker's speech and sets a constant offset.

The triangular filter bank used to calculate the Mel spectrogram and MFCC is defined as follows:

$$H_m(k) = \begin{cases} 0, & k < f(m-1); \\ \frac{k - f(m-1)}{f(m) - f(m-1)}, & f(m-1) \leq k \leq f(m); \\ \frac{f(m+1) - k}{f(m+1) - f(m)}, & f(m) \leq k \leq f(m+1); \\ 0, & k > f(m+1). \end{cases} \quad (6)$$

where $m = \overline{1, M}$, an equation (6) satisfies $\sum_{m=0}^{M-1} H_m(k) = 1$.

The use of short-term analysis in calculating the *melspec* and *MFCC* makes it possible to estimate the change in pitch over time, and hence to track the change in the speaker's intonation.

The selected informative attributes *melspec* and *MFCC* for a single sample of the speech signal represent a two-dimensional data structure in the form of an array of size $M \times N \times I$, where M is the number of coefficients used in the calculation of the analyzed attribute, and N is the number of signal frames.

As an example, Figure 5 shows the characteristics of the speech signal obtained by short-term analysis. To match the timeline, the short-term analysis was performed without overlapping the neighboring frames.

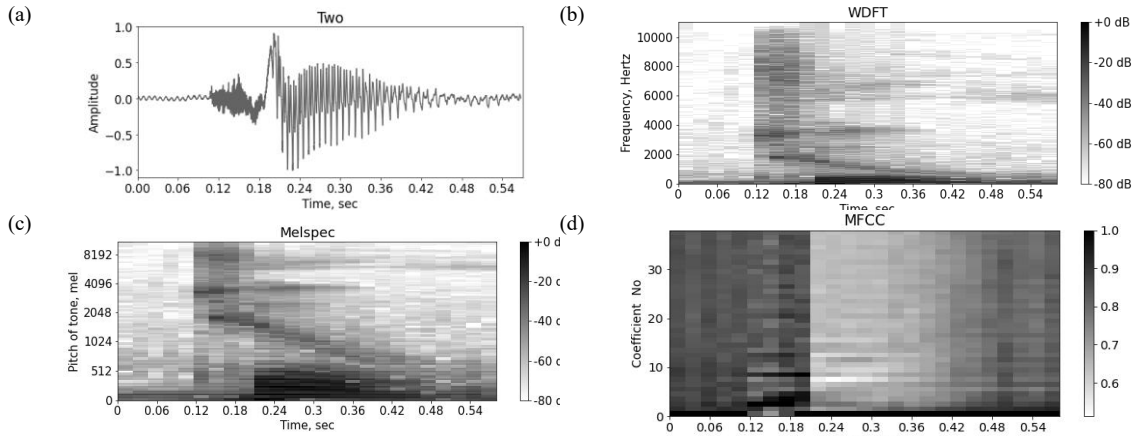


Figure 5. Characteristics of the speech signal when pronouncing the word "two":

(a) temporal characteristic; (b) amplitude spectrum of WDFT in decibels; (c) Mel spectrogram in decibels; (d) Mel-frequency cepstral coefficients normalized in range [0, 1]

As follows from Figures 5 (b) and 5 (c), for the Mel spectrogram the signal representation is more accentuated in the lower frequency range, and with increasing frequency the informativeness decreases. Figure 5 (d) shows, for example, that the *MFCC* features allow us to clearly separate the unvoiced sound $|t|$ (manifested as a noise signal from 0.12 seconds in Figure 5 (a)) and the vocalized sound $|u:|$ (in the form of a quasiperiodic process from 0.2 seconds).

4. Development of a method for classifying speaker emotions

4.1. Development of a classifier model based on deep convolutional neural networks

Ordered set of elements obtained $X_{M \times N}$ values for *melspec* (Figure 5 (b)) and *MFCC* (Figure 5 (d)) can be regarded as a monochrome image (having one channel). Then, each pixel of such an image will take the value of the informative feature for the current frame, in our case. As it is known (Goodfellow *et al.*, 2016), the highest performance in image classification is obtained by convolutional neural networks (CNN). Since the selected informative features of PES have a similar structure, the construction of the classifier in this paper was carried out on the basis of CNN.

The search for the optimal configuration of the neural network was performed by dividing the prepared training data set (Table 1) into three subsamples as training, testing, and validation. To do this,

using a random stratified partitioning all data were divided in the following proportion: 70% is for training the network, 10% is for testing, and 20% is for the final testing. The test subsample allowed us to compare different variants of network architectures, as well as to select the optimal hyperparameters.

As a result, the architecture of the deep convolutional neural network (DCNN), shown in Figure 6 (the right side of the blocks shows the dimensionality of the data, the left side - the size of the convolutional kernel) was developed as part of the problem to be solved. The notation “div” in Figure 6 is used to describe the mathematical operation of division without considering the remainder. When using informative features, *melspec* $M = 128$, $N = 27$: for *MFCC*, $M = 39$, $N = 27$.

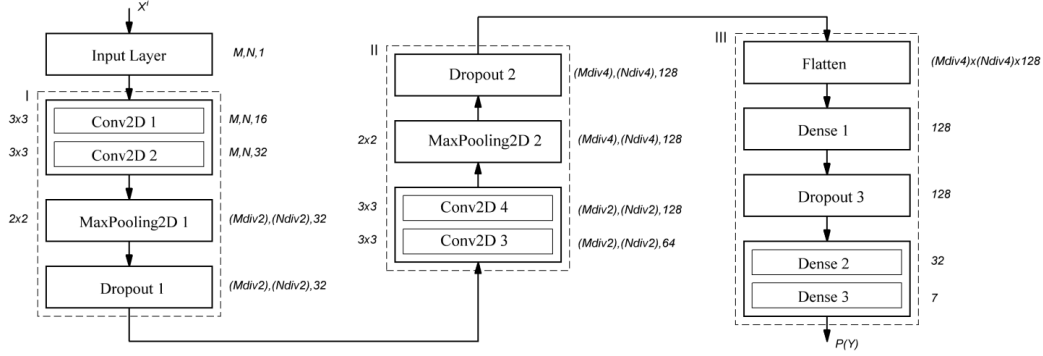


Figure 6. The architecture of the Deep CNNs

According to Figure 6, the network architecture has three main segments: two convolutional and one fully connected (network segments in Figure 6 are highlighted by a dotted line and denoted by Roman numerals). For all convolutional layers (Conv2D), the kernel size is 3×3 elements, and it is shifted by one element for each dimension. Also, for each convolutional layer, the activation function in the form of nonlinearity ReLU (rectified linear units) is used. The MaxPooling2D 1 and MaxPooling2D 2 layers in the network structure in Figure 6 are responsible for the subsampling operation.

Dropout layers are used for regularization in the learning process (Figure 6). On Dropout 1 and Dropout 2, every fourth block of input data is discarded, and on Dropout 3, every second block is discarded. In the subsequent layers Flatten, Dense 1 and Dense 2, compaction is performed. In other words, this segment implements a fully coupled neural network.

The output layer Dense 3 is a SoftMax-classifier with 7 output neurons for $C = 7$ types of classified PES.

The DCNNs training was performed separately for each of the selected informative *melspec* (3) features and *MFCC* (5). In other words, two trained networks were obtained for *melspec* and *MFCC*, respectively.

When training the neural network, an augmentation procedure was performed, i.e., data obtained as a result of calculation of appropriate informative attributes for a randomly taken time interval within the analyzed audio recording was fed to its input. From each speech signal sample s , $n = 30 - 40$ two-dimensional arrays of input data X^i were extracted. The number n depends on the duration of the analyzed sound recording sample. The duration of each randomly extracted i -th time interval s^i was taken to be 400 ms. The obtained data X^i were subjected to the normalization operation in the $[0, 1]$ range.

The learning algorithm implements an error back propagation method. The optimization problem is solved by stochastic gradient descent over 256-element mini-batches using the Adam algorithm (Kingma *et al.*, 2014).

The categorical cross-entropy is used as an error function to perform multiclass PES classification:

$$\text{loss} = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C I[y_i \in C_c] \log p[y_i \in C_c], \quad (7)$$

where N is sample size, $p[y_i \in C_c]$ is probability of the i -th sample belonging to c class; $I[y_i \in C_c]$ is indicator function.

The multiclass fraction of correct answers is used as a quality metric of *accuracy*:

$$\text{acc} = \frac{\sum_{c=1}^C \frac{tp_c + tn_c}{tp_c + tn_c + fp_c + fn_c}}{C}, \quad (8)$$

where $C = 7$ is the number of classified emotions; tp_c (*true positive*) are correctly classified samples as belonging to a c class; tn_c (*true negative*) are correctly classified samples as not belonging to the c class; fp_c (*false positive*) are specimens not belonging to c class, but mistakenly classified as belonging to it; fn_c (*false negative*) are specimens belonging to c class, but mistakenly classified as not belonging to it.

The construction and training of neural networks was performed in Python 3.7 using the tensorflow-gpu 2.1.0 library (Abadi *et al.*, 2016). Figure 7 shows the graphs of the DCNNs training process.

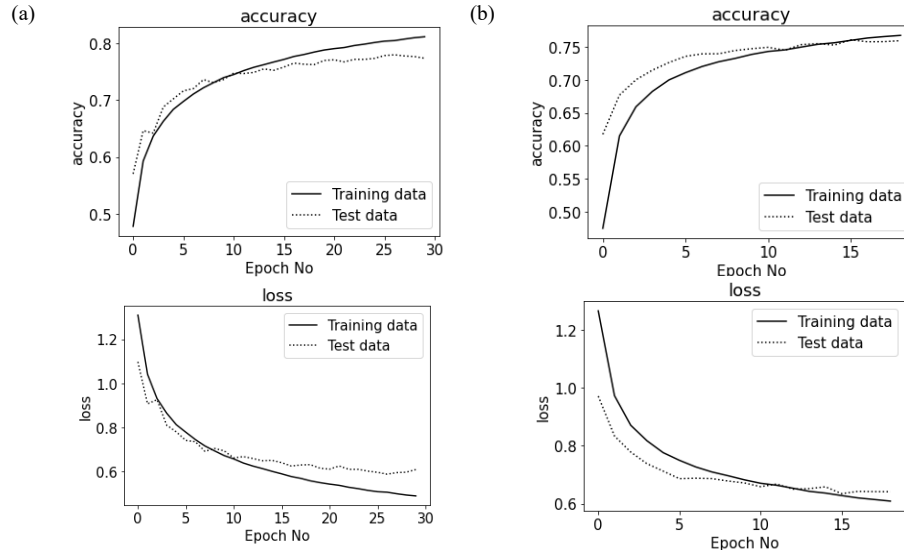


Figure 7. Neural network training graphs:
(a) DCNNs, trained on *melspec* signs; (b) DCNNs, trained on *MFCC* signs

The scheme in Figure 8 explains the mechanism of using two trained DCNNs in the classification process. The output of each neural network generates a vector of probabilities ($P(Y)_1, P(Y)_2$ in Figure 8) of a sample s belonging to each of $C = 7$ classes.

In this case, the strategy of the final determination of class by the data $P(Y)_1, P(Y)_2$ is calculated as their arithmetic average.

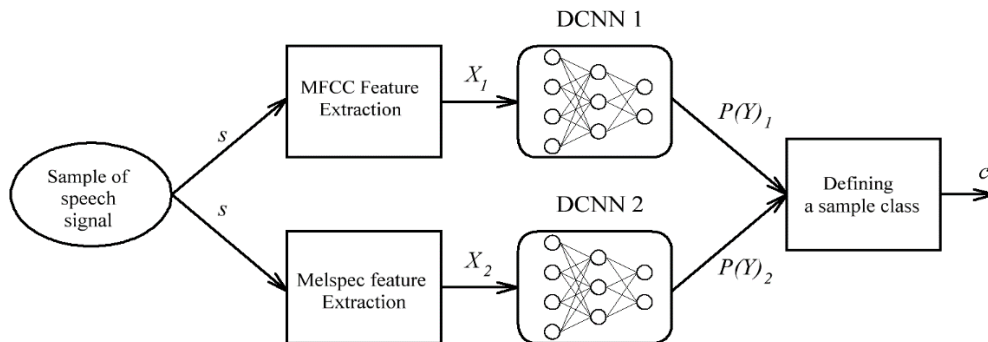


Figure 8. Using two trained DCNNs for the final classifier model

In the classification process, the analyzed sample is divided into time intervals of 400 ms. For each interval, the informative *melspec* and *MFCC* features are calculated, and the arithmetic mean of the calculated features for all intervals is fed to the inputs of the DCNNs.

5. Evaluation of the developed method effectiveness

5.1. Classification Quality metrics

Figure 9 shows the confusion matrix for the final human PES classifier based on the speech signal, obtained for the verification subsample. This classifier is based on the use of two DCNNs models trained on the *MFCC* and *melspec* features, respectively.

Table 2 and Table 3 present the classification quality metrics calculated by emotion classes for the resulting classifier model.

True class	neutral	795	8	14	1	0	1	1
	happy	2	755	7	11	3	4	16
	sad	20	13	783	2	5	6	3
	angry	5	24	0	811	5	9	6
	fearful	0	8	2	0	91	0	7
	disgusted	1	2	3	4	4	113	1
	surprised	1	60	4	22	9	2	762
		neutral	happy	sad	angry	fearful	disgusted	surprised
		Predicted class						

Figure 9. Confusion matrix of the final classifier model

The results shown testify to the effectiveness of the approach for constructing and selecting informative features. The proposed method for classifying emotions by speech signal avoids the need to recognize spoken phrases in the analysis process, which greatly simplifies the classification procedure. The model uses only acoustic data for prediction. The output of the model generates probabilities that the sample belongs to each of the seven classes. Based on this, it is possible to build fuzzy logic for the operation of automatic human condition monitoring systems.

Table 2. Classification quality metrics by sample class

Emotion type	Precision, <i>pre</i>	Recall, <i>rec</i>	<i>F1</i> -score	The number of samples in the test subsample
neutral	0,9695	0,9648	0,9672	824
happy	0,9461	0,8678	0,9053	870
sad	0,9411	0,9631	0,9520	813
angry	0,9430	0,9530	0,9480	851
fearful	0,8426	0,7778	0,8089	117
disgusted	0,8828	0,8370	0,8593	135
surprised	0,8860	0,9573	0,9203	796

The data in Table 2 show that due to unbalanced sampling, the classification accuracy for the emotions fearful and disgusted is lower than for the other classes. However, we can assume with a fair degree of certainty that the quality of classification can be improved by increasing the training sample for these classes.

Table 3. Average classification Quality metrics

Metric type	Value
Multiclass correct answers percentage <i>acc</i>	0,9319
Average for classes <i>pre</i>	0,9162
Weighted average <i>pre</i>	0,9323
Average for classes <i>rec</i>	0,9044
Weighted average <i>rec</i>	0,9319
Average for classes <i>F1</i>	0,9097
Weighted average <i>F1</i>	0,9316

5.2. Proposed classification method comparison with other studies in the field

Table 4 presents the results of the analysis comparing the classification quality obtained in this work and in the studies with the closest characteristics of the data used and requirements for the results.

The data in Table 4 show that the proposed classification method surpasses most of the known models of emotion detection based on the speech signal. At the same time, in (Razuri *et al.*, 2015) the share of correct answers is 96.97%, which is 3.78% higher than the results of this study. However, it

should be noted that the classification according to the base (Martin *et al.*, 2006) in the work (Razuri *et al.*, 2015) was made only for 6 types of emotions without determining the neutral state. Also, in (Razuri *et al.*, 2015), 264 samples of audio signals extracted from video recordings were used for the study, with one statement for each emotion. For these reasons, it can be assumed that the generalizing ability of the classification algorithms proposed in the study (Razuri *et al.*, 2015) is insufficient.

Table 4. Comparative analysis of emotional state classification Quality by the speech signal obtained in this research and similar studies

References	Classification methods	Database	Quality metrics in %
(Schuller <i>et al.</i> , 2004)	Gaussian Mix Model (GMM), k-Means, Support Vector Machine (SVM), Multilayer Perceptron (MLP)	(Schuller, 2002)	Correct answers percentage $acc = 74,2\%$
(Lee <i>et al.</i> , 2005)	Emotional weight evaluation	Private database	Classification error $err = 1 - acc$. $err = 17,85\% - 25,45\%$ for men $err = 12,04\% - 24,25\%$ for women.
(Go <i>et al.</i> , 2003)	Wavelet analysis, linear discriminant analysis	Private database	Accuracy $pre = 57\% - 93,3\%$ for men; $pre = 68\% - 93,3\%$ for women.
(El Ayadi <i>et al.</i> , 2007)	Gaussian mixture of vector autoregressive model (GMVAR)	(Burkhardt <i>et al.</i> , 2005)	$acc = 76\%$
(Schuller <i>et al.</i> , 2003)	Hidden Markov Model (HMM)	Private database	$acc = 86,8\%$
(Razuri <i>et al.</i> , 2015)	MLP, decision trees	(Martin <i>et al.</i> , 2006)	$acc = 96,97\%$
This work	Two DCNNs	RAVDESS (Livingstone <i>et al.</i> , 2018); SAVEE (Haq <i>et al.</i> , 2009); TESS (Pichora-Fuller <i>et al.</i> , 2020); ESD (Kun <i>et al.</i> , 2021)	$acc = 93,19\%$, $pre = 91,62\%$

6. Conclusions

To solve the problem of automatic classification of human emotions based on a speech signal, we propose a method that combines the results of classification from two DCNNs trained on different types of informative features: Mel spectrograms (*melspec*) and Mel-frequency cepstral coefficients (*MFCC*). As a result, the classification result is formed as the average value of probabilities of belonging of the studied sample to each of the seven classes of emotions, predicted by each neural network. With this approach to solving the PES classification problem, it is possible to achieve a multiclass fraction of correct answers equal to 0.9319 on the saved verification subsample.

The classification results are speaker-independent, as data from four different emotional corpuses are used to train neural networks. This approach can be the basis for classification of automatic systems control creation of the person's psycho-emotional condition on his spontaneous speech.

Acknowledgements

The work "AP09258659 Research and implementation of profiling technology in the system of aviation security in the Republic of Kazakhstan with the use of artificial intelligence" was carried out within the framework of Kazakhstan Committee of Science of the Ministry of Education and Science.

References

1. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J. *et al.* (2016) TensorFlow: A system for large-scale machine learning. In: *Proceedings of 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, USENIX Association, 2016, 265-283.
2. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B. (2005) A database of German emotional speech. In: *Proceedings of the Interspeech 2005*, Lissabon, Portugal, 2005, 1517-1520.
3. Cornelius, R. R. (1996) *The science of emotion: Research and tradition in the psychology of emotions*. Prentice-Hall, Upper Saddle River, NJ.
4. Ekman, P. (1971). Universals and cultural differences in facial expressions of emotion. *Nebraska Symposium on Motivation*, 19, 207-283.

5. El Ayadi, M.M.H., Kamel, M.S., Karray, F. (2007) Speech emotion recognition using Gaussian mixture vector autoregressive models. In: *Proceedings of ICASSP 2007*, 4, 957–960.
6. Garofolo, John S. *et al.* (1993) TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1. Web Download. In: *Proceedings of the Philadelphia: Linguistic Data Consortium*, 1993. DOI: 10.35111/17gk-bn40.
7. Go, H., Kwak, K., Lee, D., Chun, M. (2003) Emotion recognition from the facial image and speech signal. In: *Proceedings of the IEEE SICE 2003*, 3, 2890–2895.
8. Goodfellow, I., Bengio, Y., Courville, A. (2016) *Deep Learning*. Cambridge, MA: MIT Press.
9. Haq, S., Jackson, P.J.B. (2009) Speaker-Dependent Audio-Visual Emotion Recognition, In: *Proceedings of International Conference on Auditory-Visual Speech Processing*, 53-58.
10. Jont, B. Allen (1977) Short Time Spectral Analysis, Synthesis, and Modification by Discrete Fourier Transform. In: *Proceedings of IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-25(3), 235–238.
11. Kingma, D., Ba, J. (2014) Adam: A Method for Stochastic Optimization. In: *Proceedings of 3rd International Conference for Learning Representations, San Diego, 2015*, 1-15.
12. Kun, Z., Berrak, S., Rui, L., Haizhou, L. (2021) Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset. In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 6-11 June 2021. Toronto, Ontario, Canada. DOI: 10.1109/ICASSP39728.2021.9413391.
13. Lawrence, R., Ronald, S. (2007) Introduction to Digital Speech Processing, *Foundations and Trends in Signal Processing*, 1(1–2), 1-194. DOI: 10.1561/20000000001.
14. Lee, C., Narayanan, S. (2005) Toward detecting emotions in spoken dialogs. In: *Proceedings of IEEE Trans. Speech Audio Process*, 13(2), 293–303.
15. Livingstone, S., Russo, F. (2018) 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*, 13(5). DOI: 10.1371/journal.pone.0196391.
16. Martin, O., Kotsia, I., Macq, B., Pitas, I. (2006) The INTERFACE'05 Audio-Visual Emotion Database. In: *Proceedings of Data Engineering Workshops, Proceedings. 22nd International Conference*. DOI: 10.1109/ICDEW.2006.145.
17. Pichora-Fuller, M. K., Dupuis, K. (2020) Toronto emotional speech set (TESS). *Scholars Portal Dataverse*, V1. DOI: 10.5683/SP2/E8H2MF.
18. Razuri, J. G., Sundgren, D., Rahmani, R., Larsson, A., Moran, A. C., Bonet, I. (2015) Speech emotion recognition in emotional feedback for Human-Robot. *Interaction. International Journal of Advanced Research in Artificial Intelligence*, 4(2). DOI: 10.14569/IJARAI.2015.040204.
19. Schuller, B. (2002) Towards intuitive speech interaction by the integration of emotional aspects. In: *Proceedings of IEEE International Conference on Systems, Man and Cybernetics*, 6. DOI: 10.1109/ICSMC.2002.1175635.
20. Schuller, B., Rigoll, G., Lang, M. (2003) Hidden Markov model-based speech emotion recognition. In: *Proceedings of the International Conference on Multimedia and Expo (ICME)*, 1, 401–404.
21. Schuller, B., Rigoll, G., Lang, M. (2004) Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In: *Proceedings of the ICASSP 2004*, 1, 577–580.
22. Uday, K., John, L., Whitaker, J. (2019) *Deep Learning for NLP and Speech Recognition*. Springer Nature Switzerland AG.
23. Zwicker, E., Fastl, H. (1990) *Psycho-acoustics*. Springer-Verlag, 2nd Edition.