The research focuses on developing a novel method for the automatic recognition of human psychoemotional states (PES) using deep learning technology. This method is centered on analyzing speech signals to classify distinct emotional states. The primary challenge addressed by this research is to accurately perform multiclass classification of seven human psychoemotional states, namely joy, fear, anger, sadness, disgust, surprise, and a neutral state. Traditional methods have struggled to accurately distinguish these complex emotional nuances in speech. The study successfully developed a model capable of extracting informative features from audio recordings, specifically mel spectrograms and mel-frequency cepstral coefficients. These features were then used to train two deep convolutional neural networks, resulting in a classifier model. The uniqueness of this research lies in its use of a dual-feature approach and the employment of deep convolutional neural networks for classification. This approach has demonstrated high accuracy in emotion recognition, with an accuracy rate of 0.93 in the validation subset. The high accuracy and effectiveness of the model can be attributed to the comprehensive and synergistic use of mel spectrograms and mel-frequency cepstral coefficients, which provide a more nuanced analysis of emotional expressions in speech. The method presented in this research has broad applicability in various domains, including enhancing human-machine interface interactions, implementation in the aviation industry, healthcare, marketing, and other fields where understanding human emotions through speech is crucial

Keywords: speech emotion recognition, deep learning in SER, MEL spectrogram, MFCC analysis, audio signal processing, emotional classification, acoustic features, machine learning, emotion detection, psycholinguistics

UDC 004.83
DOI: 10.15587/1729-4061.2023.293011

DEVELOPMENT OF AN ADVANCED AI-BASED MODEL FOR HUMAN PSYCHOEMOTIONAL STATE ANALYSIS

Zharas Ainakulov
Corresponding author
Master of Engineering Sciences, Doctoral Student
Department of Information Systems
Al-Farabi Kazakh National University
al-Farabi ave., 71, Almaty, Republic of Kazakhstan, 050040
E-mail: jaras1987@mail.ru
Kayrat Koshekov
Doctor of Technical Sciences, Professor*
Alexey Savostin
PhD, Associate Professor
Department of Energetic and Radioelectronics
M. Kozybayev North Kazakhstan University
Pushkin str., 86, Petropavlovsk, Republic of Kazakhstan, 150000
Raziyam Anayatova
PhD*
Beken Seidakhmetov
Candidate of Economic Sciences
Vice-Rector**
Gulzhan Kurmankulova
Associated Processor, Candidate of Pedagogical Sciences
Department of IT Technologies and Automation
Kazakh National Agrarian Research University
Abay ave., 8, Almaty, Republic of Kazakhstan, 050010
*Department of Science and International Cooperation**
**Civil Aviation Academy
Akhmetova str., 44, Almaty, Republic of Kazakhstan, 050039

Received date 02.10.2023
Accepted date 11.12.2023
Published date 29.12.2023

How to Cite: *Ainakulov, Z., Koshekov, K., Savostin, A., Anayatova, R., Seidakhmetov, B., Kurmankulova, G. (2023). Development of an advanced ai-based model for human psychoemotional state analysis. Eastern-European Journal of Enterprise Technologies, 6 (4 (126)), 39–49. doi: https://doi.org/10.15587/1729-4061.2023.293011*

1. Introduction

Contemporary societal trends and complex socio-economic conditions have escalated the need for an in-depth exploration into the impact of personnel's psychoemotional states on the quality and safety of their labor activities. In sectors such as healthcare, marketing, security, and management in hazardous industrial and transportation domains, there's an increasing demand for novel, efficient tools for automatic determination of human psychoemotional states based on vocal characteristics. Automated recognition of psychoemotional states constitutes a pivotal advancement in enhancing human-machine interfaces, enabling the assessment of stress and fatigue levels, and detecting depressive conditions, which could significantly alleviate accumulated fatigue and similar phenomena [1].

The significance of implementing automated systems for voice-based psychoemotional state recognition is particularly notable in sectors where communication predominantly occurs orally without visual contact. Such systems are valuable in contexts aiming to mitigate potential risks to individuals and assets through continuous identification and risk factor monitoring [2]. Despite the importance of this issue, there is currently no comprehensive theory that reveals the correlation between a speaker's emotions and voice signal characteristics. However, progress has been made in similar tasks involving automatic speech recognition, primarily enabled by contemporary digital signal pro-

cessing algorithms and the advancement of machine learning methods [3].

Therefore, research endeavors devoted to developing automated systems for voice-based psychoemotional state recognition are of paramount scientific relevance. Such studies not only bridge the gap in existing theoretical frameworks but also cater to the pressing needs of various sectors requiring advanced human-state monitoring tools. The pursuit of these innovative approaches, harnessing the potential of AI and machine learning, is crucial for advancing our understanding and practical capabilities in this vital area of human-machine interaction.

2. Literature review and problem statement

[4] explores the advancements in AI for emotion recognition, emphasizing the progress in natural language processing and facial expression analysis. However, it highlights a gap in accurately interpreting emotions from speech, particularly in real-time scenarios.

In [5], the complexity of speech as a medium for emotional conveyance is examined. It notes the objective difficulties in segregating emotional content from informational content in speech, suggesting the need for more advanced analytical methods.

[6] discusses the application of deep learning techniques in emotion recognition. While it showcases the potential of these techniques, it also underlines the limitations in their adaptability to diverse linguistic and cultural contexts.

The work by [7] delves into the resource-intensive nature of training AI models for emotion recognition, pointing out the costly nature of data acquisition and processing, which can be a deterrent for extensive research.

[8] presents an innovative approach that integrates multiple acoustic features for emotion analysis. This study, however, does not fully explore the integration of melspec and MFCC features, leaving room for further exploration.

[9] examines the challenges in real-time emotion recognition from speech, highlighting the need for more efficient processing algorithms that can operate in dynamic environments.

In [10], the variability of emotional expression across different languages and cultures is explored. It argues for the development of more universally applicable AI models in emotion recognition.

[11] emphasizes the importance of diverse data sets for training AI in emotion recognition, pointing out that many current models are limited by the homogeneity of their training data.

The ethical implications of emotion-recognition AI are critically assessed in [12]. This source raises concerns about privacy and consent, which are yet to be adequately addressed in current research.

[13] suggests an interdisciplinary approach, combining insights from psychology, linguistics, and computer science, to enhance the effectiveness of emotion recognition AI.

These studies collectively underscore the importance and complexity of emotion recognition from speech using AI. While significant strides have been made, there remain unresolved issues related to real-time processing, cultural and linguistic diversity, ethical considerations, and the integration of complex feature sets. Overcoming these challenges requires innovative approaches that can efficiently integrate diverse feature sets, address ethical concerns, and adapt to a wide range of emotional expressions across different contexts. This justifies conducting a study devoted to the development of a more holistic and efficient AI model for emotion recognition from speech.

3. The aim and objectives of the study

The aim of this study is to develop an advanced AI-based model capable of accurately recognizing and classifying human psychoemotional states (PES) through speech analysis. This model seeks to address the existing challenges in speech emotion recognition (SER) by effectively interpreting the emotional nuances contained in speech patterns.

To achieve this aim, the following objectives are accomplished:

– to conduct a comprehensive analysis of current SER techniques;

– to integrate advanced machine learning algorithms;

– to utilize a dual-feature approach to enhance emotion recognition accuracy;

– to test and validate the model with diverse datasets;

– to address real-time processing capabilities.

4. Materials and methods of research

A significant challenge in achieving aim of the study is the lack of precision and coherence in existing formulations of the emotion concept and the theoretical models for their classification. Thus, when implementing an automatic psychoemotional state recognition system, it becomes crucial to delineate a set of archetypal emotions, including joy, fear, anger, sadness, disgust, surprise, and a neutral state [14]. Additionally, the intricate multidimensional nature of the speech signal, the underlying mechanisms governing its production, and the subtleties of psychoacoustic sound perception by humans must be considered. This approach allows for the identification of latent patterns within the data, even in the presence of uncertainty from the researcher's perspective [15].

The object of this research is human psychoemotional states as expressed through speech.

The subject is the classification of these states using AI-driven techniques, focusing on deep learning methods.

It was assumed that the primary psychoemotional states could be represented through audio recordings, disregarding visual and contextual cues.

Simplification was made in the analysis methodology, emphasizing the classification of short speech fragments and relying primarily on vocal intonations.

Convolutional Neural Networks (CNNs) were chosen due to their demonstrated efficacy in handling data structured as melspec and Mel-Frequency Cepstral Coefficients, as shown in practical applications [16].

The deep CNN architecture was selected for its ability to process multi-dimensional data and extract complex features from speech signals.

A dataset comprising various emotional nuances in speech was compiled, using sources such as RAVDESS, SAVEE, TESS, and ESD.

Mel spectrograms and Mel-Frequency Cepstral Coefficients were extracted from the speech data, each frame represented as a monochromatic image for CNN processing.

A deep CNN architecture was developed, with convolutional layers using a 3×3 kernel size and the Rectified Linear Units activation function.

The neural network was supplied with randomly selected intervals of audio recordings, with each interval normalized within the range of [0, 1].

The training employed the backpropagation technique and stochastic gradient descent using mini-batches, with the Adam algorithm for optimization [17].

Categorical cross-entropy was used as the error function, and multi-class accuracy was the primary metric for assessing model quality.

The research utilized Python 3.7 and TensorFlow-GPU 2.1.0 for neural network creation and training.

In the research, artificial intelligence was employed to classify human psychoemotional states from speech. This involved multiple steps: collecting audio data from sources like RAVDESS and SAVEE, extracting features such as Mel spectrograms and Mel-Frequency Cepstral Coefficients using deep learning, developing a Convolutional Neural Network (CNN) for processing these features, training the CNN with techniques like backpropagation and stochastic gradient descent, evaluating the model's accuracy in classifying emotional states.

AI's role was pivotal in extracting complex features from speech and classifying emotions accurately, adhering to the methodology outlined in the research.

5. Results of the research on AI-driven psychoemotional state analysis through voice

5. 1. Comprehensive analysis of current speech emotion recognition techniques

The field of speech emotion recognition (SER) has become increasingly crucial in the realm of Human-Computer Interaction (HCI), with numerous systems proposed using both Machine Learning (ML) and Deep Learning (DL) methodologies. A systematic review of these SER approaches has shown the growing importance and efficacy of these technologies in identifying and interpreting human emotions from speech patterns.

Recent studies have reviewed deep learning techniques for SER, highlighting the advancements in the use of available datasets and conventional machine learning techniques. The multi-aspect comparison of practical neural network approaches in SER demonstrates the evolution of this technology, moving beyond traditional pattern recognition to more sophisticated, nuanced analysis of emotional speech.

In addition to these developments, there has been significant research progress in understanding the methodology and experimental results associated with SER systems. These advancements encompass various aspects, including emotional speech datasets, data enhancement and augmentation, feature extraction, and emotion classification. This comprehensive approach to SER underscores the field's movement towards more intricate and refined systems capable of interpreting a wider range of human emotions with greater accuracy [18].

Our research aligns with these recent advances, aiming to further the capabilities of AI in accurately recognizing and classifying human psychoemotional states through speech. By integrating cutting-edge ML and DL techniques, our study contributes to the ongoing development of more efficient and accurate SER systems, addressing some of the current challenges faced in the field.

5. 2. Integration of advanced machine learning algorithms

The quest to automate the assessment of psychoemotional states (PES) through voice is challenged by the absence of a comprehensive theory linking the characteristics of the voice signal to the speaker's emotions. However, the remarkable strides in AI, particularly in natural language processing, pave the way for innovative approaches to this problem. This research embraces the use of deep convolutional artificial neural networks (ANNs) for analyzing PES through voice, leveraging the dynamic and rapidly advancing domain of ANNs and deep learning within AI. Fig. 1 illustrates a schematic diagram explaining the amalgamation of existing AI.

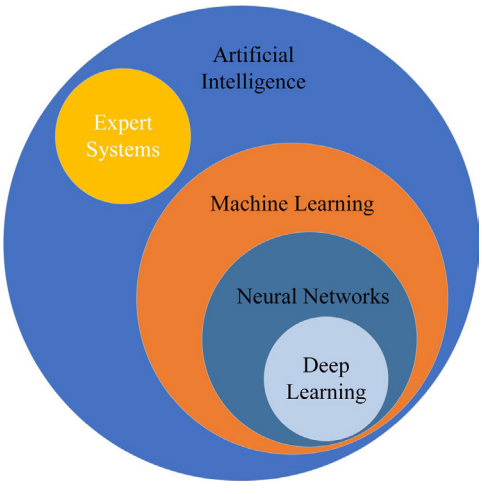

Fig. 1. The Structure of artificial intelligence

Deep Learning (DL), a subset of Machine Learning (ML) within AI, specializes in training deep ANNs to solve complex tasks. These networks, particularly deep Convolutional Neural Networks (CNNs), are adept at processing multi-dimensional data, including speech and audio signals. CNNs are capable of recognizing and abstracting intricate features from data without manual definition, making them particularly suitable for understanding and classifying human emotions from voice signals.

Recent advancements in ML and DL have significantly propelled the capabilities of SER systems. The current research landscape shows a growing interest in SER within the affective computing domain, underpinned by increasing potential and algorithmic advancements. Recent solutions in SER have been increasingly employing various machine learning and deep learning techniques, reflecting an active pursuit of more refined and accurate emotion recognition from speech.

The need for large volumes of data and significant computational resources is a notable aspect of training effective deep ANNs. Graphical processing units (GPUs) or tensor processing units (TPUs) are often employed to expedite this process. The continuous expansion of DL applications across various fields, including medicine, autonomous vehicles, and natural language processing, underlines its potential in enhancing SER technologies [19].

This study's approach, utilizing a dual-feature method with deep CNNs, aligns with the current trajectory of SER research. By employing sophisticated machine learning algorithms and deep learning techniques, let's aim to bridge the gap in automatic emotion recognition from speech, contributing to the advancement of AI-driven psychoemotional state analysis.

5. 3. Utilization of a dual-feature approach

The primary challenge in automatic recognition of a person's psychoemotional state (PES) through voice is the ambiguity and uncertainty of existing theories and models of emotion. To address this, it's necessary to identify a set of archetypal

emotions including joy, fear, anger, sadness, disgust, surprise, and a neutral state (calmness) [20]. The objective of automatic classification is to determine the probability of the speaker's emotions falling into each of these seven emotional classes, enabling a preliminary assessment of the speaker's PES.

For classification using deep learning (DL) methods, the mathematical model in the form of artificial neural networks (ANNs) should make a decision to determine the class with the maximum posterior probability. This is known as the maximum a posteriori (MAP) solution:

$$y_{MAP} = \arg\max_{Y} P(X|Y)P(Y). \quad (1)$$

In the expression (1), X represents the feature values of classification objects, Y denotes the set of target variables (object classes), $P(X|Y)$ is the likelihood function, and $P(Y)$ is the prior probability. The prior probability $P(Y)$ is the probability of each emotional class before observing the data X.

Assuming a uniform prior distribution for $P(Y)$, it is possible to express the Maximum Likelihood (ML) decision rule:

$$y_{ML} = \arg\max_{Y} \sum_{x \in X} \log P(X|Y). \quad (2)$$

The task of determining a person's PES through voice involves constructing a mathematical model for a multi-class classifier using specific informative features X to estimate probabilities. It's crucial to account for speaker-independence concerning gender, age, and pronunciation idiosyncrasies, as well as the complex structure of the speech signal, human psychophysical sound perception, and the arbitrary nature of the analyzed utterances [21].

An analysis of existing emotional corpora available in the public domain revealed that the following audio databases will meet the research needs: the Ryerson audio-visual database of emotional speech and song, Surrey audio-visual expressed emotion, Toronto emotional speech set, emotional speech database.

5. 4. Testing and validation with diverse datasets

From the databases, a training dataset was constructed, the structure of which is presented in Table 1:

– The Ryerson audio-visual database of emotional speech and song (RAVDESS) [22];

– Surrey audio-visual expressed emotion (SAVEE) [23];

– Toronto emotional speech set (TESS) [24];

– emotional speech database (ESD) [25].

5. 5. Assessment of real-time processing capabilities

Fig. 2 shows characteristics of the formed emotional corpus.

The prepared dataset contains recordings of seven types of psychoemotional states (PES) from English-speaking speakers. The audio recordings significantly vary in the intensity of emotional expression, speaker composition, and pronunciation variants, ensuring the necessary representativeness. This structure of the emotional corpus will contribute to a more diverse representation of samples for training the classifier model.

The proposed method for determining a person's psychoemotional states (PES) using deep learning (DL) technologies is depicted in Fig. 3.

As shown in Fig. 3, the developed method consists of three stages: pre-processing, feature extraction, and the final prediction stage.

During the pre-processing stage, the speech signal undergoes analog-to-digital conversion with a sampling frequency of fS=22050 Hz and a quantization bit depth of 16 bits. The pre-processing also involves digital filtering to remove noise components from the signals. The choice and characteristics of the digital filters used depend on the level and nature of the interfering noise affecting the signal.

The removal of pauses from speech is necessary to analyze only segments of the signal that contain information about the speaker's voice. Therefore, in the pre-processing stage, pauses of approximately 1 second or longer are removed from the speech recordings. This ensures the integrity of speech sequences in phrases and sentences, while eliminating extended silent sections in the audio recordings. The process of pre-processing signals was detailed in the previous report.

During the feature extraction stage, features are constructed from speech signals that can convey information about a person's psychoemotional state (PES) and meet the input requirements of convolutional neural networks (CNNs).

Since speech signals are statistically quasi-stationary for short periods of time, their analysis is effectively performed using short-time analysis tools. In the proposed method, an analysis was conducted on frames of the original speech signal, each frame being 512 samples long (approximately 23 ms) with a 35 % overlap.

During short-time analysis, dividing the speech signal into frames allows the analysis of individual segments of the recording as sequences of sounds with varying properties. For this reason, the outcome of processing each frame is the extraction of a set of features representing the sound contained within it.

Table 1

The distribution of audio recordings by emotional classes 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
Summary	624	624	420	0	0	2800	8750	8750	21968
	1248		420		2800		17500		

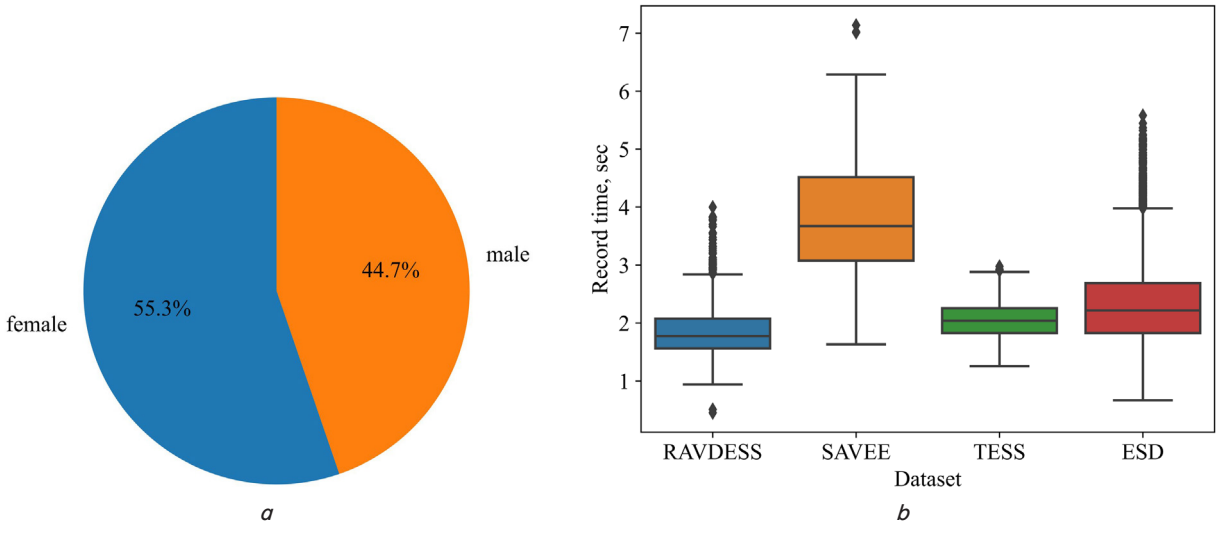

Fig. 2. Characteristics of the formed emotional corpus:
a – the ratio of male and female speakers; *b* – duration statistics of recordings by dataset

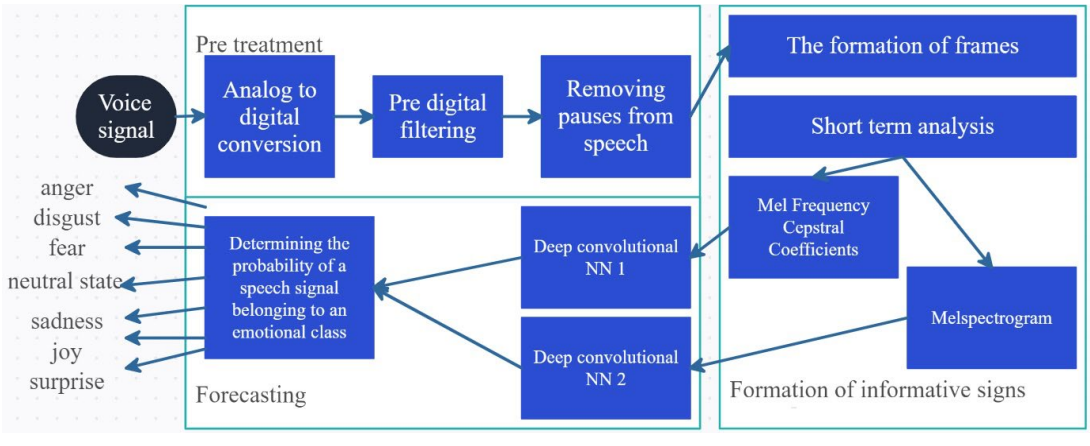

Fig. 3. The structure of the method for automatic psychoemotional state (PES) determination from voice

To account for the subjective aspects of sound perception by the human auditory system, it is ineffective to use the windowed discrete Fourier transform (DFT) spectral information as the representation of the speech signal. DFT operates on a linear frequency scale, whereas human sound perception depends on a subjective characteristic known as pitch. Mapping frequency values to pitch allows the construction of a spectral feature of the audio signal, using a mel scale instead of the frequency axis [26].

In this scenario, two types of features were selected as informative for the automatic determination of a person's psychoemotional state (PES) from their voice:

1. Mel spectrogram (melspec) – an estimation of the power spectrum of an audio signal constructed based on the mel scale $0 \le m < M$:

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

where $S_i(k)$ – fast Fourier transform (FFT) of the speech signal s for the i-th frame obtained during short-time analysis; $N=2^j$, $j \in N$ – the number of samples in the analyzed frame of the signal s; $H_m(k)$ – a bank of $M=128$ specialized triangular filters [27].

2. Mel frequency cepstral coefficients (MFCC) are also based on the use of the mel-frequency scale but are less sensitive to the characteristics of the speaker's vocal tract and allow for a significant reduction in the dimensionality of individual speech signal features $0 \le 1 < M$:

$$c_i(l) = \sum_{m=0}^{M-1} 10 \log P_i(m) \cdot \cos\left(\pi l \left(m + \frac{1}{2}\right) \Big/ M\right), 0 \le l < M. \quad (4)$$

Mel spectrograms and MFCC obtained through short-time analysis allow for the assessment of pitch changes over time, thereby enabling the detection of shifts in the speaker's intonation. Segmenting the signal into frames for each time interval provides the opportunity to obtain informative features in the form of two-dimensional matrices of coefficients, melspec and MFCC. The two-dimensional format of these informative features facilitates the use of deep Convolutional Neural Networks (CNNs) designed to work with multidimensional data.

It's worth noting that CNNs significantly outperform other machine learning algorithms in tasks where the spatial arrangement of features is crucial. The computed coefficients of melspec and MFCC have a matrix structure with dimen-

sions of 128×27 and 39×27, respectively. In other words, the array of coefficients $XM \times N$ in the form of melspec or MFCC can be viewed as a monochromatic image (with one channel), where each pixel represents a specific value of the informative feature for the current frame.

As demonstrated by practical experience [28], the use of Convolutional Neural Networks (CNNs) allows for the creation of the most effective classifiers when dealing with data of this structure. For this reason, the design of the psychoemotional state (PES) classifier was based on a deep CNN. In deep learning, multiple convolutional layers are employed, enabling a more comprehensive analysis by identifying complex interactions within the data structure.

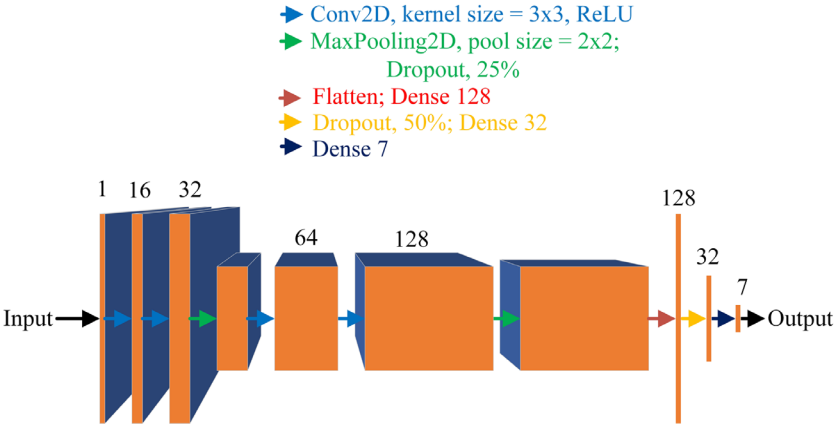

Fig. 4. Schematic representation of the deep convolutional neural network architecture

The search for the optimal neural network architecture was conducted by dividing the prepared training dataset (Table 1) into three subsets: training, testing, and validation. The validation subset is necessary for comparing different CNN structures and for tuning the optimal hyperparameters.

As a result of this research, an architecture for a deep CNN was developed, as shown in Fig. 4.

Fig. 4 depicts the number of channels that form in the data after the two-dimensional convolution procedure. The colored arrows represent the types of transformations performed on the data.

The architecture shown in Fig. 4 was used to train two separate CNNs with different input parameters – one CNN for MFCC and one CNN for melspec coefficients. The training of the CNNs was carried out using Python 3.9 with the TensorFlow 2.1 and scikit-learn 1.3.1 libraries. For the development and training of our CNN models, TensorFlow 2.1 and scikit-learn 1.3.1 libraries were utilized. TensorFlow, initially created by the Google Brain team in the United States, offers extensive functionalities that are particularly advantageous for developing deep learning models like CNNs. Its capabilities in model building, efficient production, and research experimentation made it an ideal choice for our study's requirements. Additionally, scikit-learn 1.3.1 was employed for data processing and analysis tasks. This library, also developed in the United States, is renowned for its simplicity and efficiency, providing valuable tools for data mining and analysis. These functionalities were crucial in managing the preprocessing and feature extraction processes for our speech data. The combination of these tools, originating from the Netherlands and the United States, respectively, provided a comprehensive and effective environment for the development and training of our deep learning models. Their selection was not only based on their specific technical merits but also on their global recognition and support in the machine learning community.

The layer names of the CNNs, as indicated in Fig. 4, correspond to the designations in the TensorFlow library.

As a result, as shown in Fig. 3, two trained deep CNNs were obtained. The first CNN takes MFCC as input features, and the second takes features from melspec coefficients. The input data dimensionality for the two CNNs is different and is determined by the dimensionality of the $XM \times N$ arrays of the corresponding coefficients.

Each developed CNN generates the classification result for the speech signal as the average classification of its individual segments based on the given frames. The classification result is a vector of probabilities $P(Y)$ indicating the likelihood of the examined speech signal sample belonging to each of the seven emotion classes $C=7$. The corresponding emotion class is determined as follows:

$$c = \sum_{i=1}^{C} I\left[y_i = \arg\max(Y)\right] \cdot i. \quad (5)$$

Since two CNN models are used in the prediction stage as shown in Fig. 3, the final prediction is made by taking the arithmetic mean of the probability vectors obtained from them:

$$c = \sum_{i=1}^{C} I\left[y_i = \arg\max\left(\frac{P(Y)_1 + P(Y)_2}{2}\right)\right] \cdot i, \quad (6)$$

where $P(Y)_1$, $P(Y)_2$ the probabilities of the speech signal sample belonging to one of the $C=7$ emotion classes.

The effectiveness of the proposed method for recognizing a person's emotional state from their speech can be assessed using standard quality metrics in a classification task. For this purpose, a test subset was selected from the overall training dataset (Table 1). Table 2 presents the achieved quality metrics.

The obtained results when applying the developed emotion recognition method have enabled the creation of specialized software for recognizing emotional states on a personal computer.

Table 2

Quality metrics for sample classification by class

Type of Emotion	Precision	Recall	F-measure
Neutral	0.9695	0.9648	0.9672
Happy	0.9461	0.8678	0.9053
Sad	0.9411	0.9631	0.9520
Angry	0.9430	0.9530	0.9480
Fearful	0.8426	0.7778	0.8089
Disgusted	0.8828	0.8370	0.8593
Surprised	0.8860	0.9573	0.9203
Average	0.9159	0.903	0.9087

The software for implementing the described method of recognizing the psychoemotional state of a person by voice was developed in the Python 3.9 programming language. The PyQt5 library was used to create the graphical user interface. PyQt5 is a set of Python bindings for the cross-platform Qt framework, which is used for creating graphical user interfaces (GUI) and other applications. PyQt5 allows the use of Python as an alternative development language for applications on all supported platforms, including Windows, iOS, and Android.

PyQt5 consists of more than 35 extension modules that implement various aspects of working with Qt, such as multimedia, NFC, Bluetooth connectivity, a Chromium-based web browser, and traditional GUI development. PyQt5 also provides the PyQt5 Designer tool, which allows the rapid creation of complex GUI applications by dragging and dropping widgets. In particular, PyQt5 Designer was used to create the graphical interface template for the application.

PyQt5 is distributed under the GPL v3 license, which allows the use of this software without copyright and patent restrictions.

In Fig. 5, the appearance of the developed software for emotion recognition from voice is depicted. The software interface consists of several blocks, numbered in Fig. 5. The developed software allows the analysis of voice recordings from the microphone or processing audio files stored on the PC.

The user interface controls for analyzing audio recordings from the microphone are represented by the elements marked as '1' in Fig. 5. The 'Recording Time, sec' field displays the current recording duration in seconds. It is recommended to analyze audio segments with a duration of no more than 10 seconds.

The microphone icon button initiates the audio recording process from the microphone, while the adjacent button is used to stop the recording. Once the recording is stopped, the 'Save' button becomes active, allowing users to save the audio file to their computer through an opening dialog window. The file will be saved in the *.wav format, and the default file name is generated following the template 'audio_YY-MM-DD_HH-MM-SS.wav.' This file naming convention aids in organizing files by their creation time.

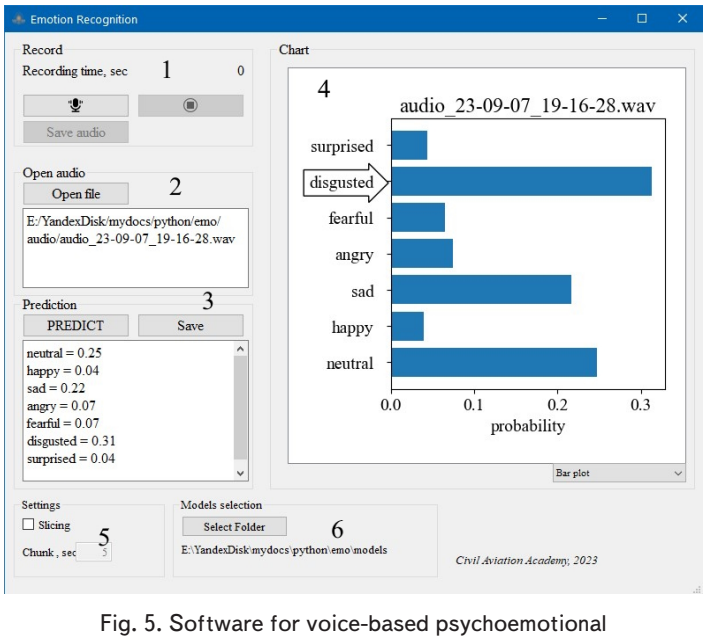

Fig. 5. Software for voice-based psychoemotional state recognition

In the software interface block labeled as number 2 in Fig. 5, the «Open File» button allows users to select any file on their computer for analysis. The dialog window that opens facilitates the selection of only *.wav files. The chosen file's path will be displayed in a text field, as shown in Fig. 5.

Once an audio file is selected, it can be analyzed to predict the speaker's emotional state. Controls for this purpose are represented by number 3 in Fig. 5. Clicking the «PREDICTION» button initiates the audio analysis process using the developed method (as shown in Fig. 3). During the initial phase of signal processing, the software analyzes the audio based on its amplitude characteristics. If the recording quality is deemed unsatisfactory, the software will generate an error message.

After the forecast is calculated, its results will be displayed in the text field of the current block of the interface. For each class of emotion, the probability of its presence is shown. The forecast results can be saved in a text file for subsequent analysis. To do this, it is necessary to press the 'Save' button. This will open a window for saving the file. The default file name format is as follows: prediction_YY-MM-DD_HH-MM-SS.txt. With this approach, a corresponding forecast can be associated with each audio file.

The prediction results are also displayed in a user-friendly manner in the form of a diagram in the interface block labeled as number 4 in Fig. 5. The diagram includes information about the analyzed file's name and the probabilities associated with each of the 7 emotions. The most probable emotion is indicated by an arrow. Additionally, there is an option to switch the type of graphical representation of the prediction results. The bar chart, as shown in Fig. 5, can be replaced with a pie chart, an example of which is shown in Fig. 6.

The circular diagram (Fig. 6) highlights the most likely emotional prediction with a distinct segment from the overall circle.

To choose the diagram type, it is possible to use the drop-down list located below the shape output window in the user interface.

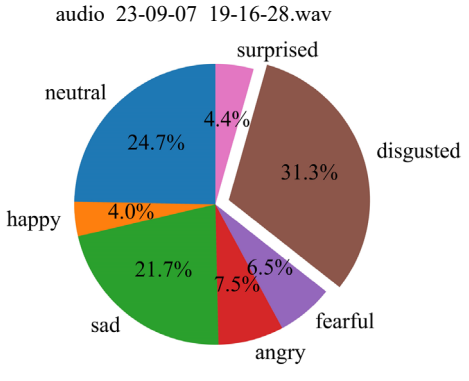

Fig. 6. The pie chart of emotional prediction

The functionality of the user interface block under number 5 in Fig. 5 can be utilized when the analyzed audio file has considerable length. By using the displayed settings, the file can be divided into segments, the duration of which is determined in the «Block, s» input field. To activate this mode, it is necessary to check the «Segmentation» checkbox.

To use the software based on the developed method for recognizing a speaker's emotional state from their voice, two pre-trained deep neural networks with the

architecture shown in Fig. 4 must be employed. These neural networks have been pre-trained and are provided along with the described software. By default, the files with the pre-trained neural networks are located in the «models» directory at the root of the program's files. However, if needed, it is possible to specify a different directory for the model files. To do so, it is possible to utilize the interface block numbered 6 in Fig. 5. When to press the «Select Folder» button, a window will open for selecting the required directory.

Table 3

Confusion matrix

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

During this process, the software checks whether the specified directory contains the necessary model files. If the files are not found, the directory settings will be reverted to their initial state.

In this software, the neural network models are saved in the *.h5 and *.json formats. The h5 format allows saving the entire model, including its architecture, weights, and optimizer state, in a single HDF5 file. The json format enables saving only the model's architecture as a text file, which can then be loaded and used with different weights or optimizers. This ensures the functional flexibility to improve model performance and use it in the future.

The settings for the method used to determine the speaker's emotional state, as implemented in this software, can be customized. To do this, a text file named config.txt with the required settings should be placed in the program's root directory. The format for specifying settings in the configuration file is outlined below (Fig. 7).

```
[Configurations]
# SAMPLING RATE
# fs: sampling rate in Hz - int.
sr = 22050

# MFCC COEFFICIENTS
# n_mfcc: Number of MFCC coefficients - int.
n_mfcc = 40

# MELS COEFFICIENTS
# n_mels: Number of Mel coefficients - int.
n_mels = 128

# Window size
# window:  Window size in sec - float
window = 0.4

# OTHER
```

Fig. 7. Configuration screenshot

Each time the software is launched, it searches for the configuration file in the root directory. If the file doesn't exist, default settings are used.

In the presented program, logging of the work process is also implemented. A log file for the program is created in the root directory with the name «app.log.»

Table 3 displays matrix for the ultimate classifier of human psycho-emotional states, which is founded on the analysis of speech signals. This classifier leverages the deployment of two deep convolutional NN models. The results are obtained using the validation subset.

Tables 4, 5 present classification quality metrics calculated for emotion classes using the obtained final classifier model. The proposed method for emotion classification based on speech signals eliminates the necessity for speech recognition during analysis, thereby streamlining the classification process significantly. The model relies solely on acoustic data for making predictions. It calculates probabilities of an instance belonging to each of the seven classes. This capability opens the door to designing fuzzy logic for automated human state monitoring systems using such a model.

Table 4

Classification quality

Emotion	Accuracy	Completeness	*F*1	Test
Neutral	0.9694	0.9647	0.9671	825
Happy	0.9460	0.8677	0.9052	871
Sad	0.9410	0.9630	0.9521	812
Angry	0.9431	0.9531	0.9481	852
Fearful	0.8426	0.7778	0.8089	117
Disgusted	0.8828	0.8370	0.8593	135
Surprised	0.8860	0.9573	0.9203	796

Table 5

Metrics

Metric type	Meaning
The multi-class share of correct answers accuracy	0.9318
Average precision for classes	0.9161
Weighted mean precision	0.9321
Average recall for classes	0.9043
Weighted average recall	0.9318
The average F1	0.9098
Weighted mean *F*1	0.9317

However, it's worth noting that in the work [29]. Nevertheless, it's important to point out that in the [30] dataset in the work [31], classification was performed for six types. Additionally, in the work [31], 264 audio signal samples extracted from video recordings.

6. Discussion of the results of the study on AI-driven analysis of human psychoemotional states through voice

This section delves into the interpretation of the study's results, linking back to the objectives and highlighting the

advancement over existing models. The integration of deep convolutional neural networks (CNNs) with dual-feature analysis, combining mel spectrograms and MFCC, is a significant leap forward, as indicated by the comparative metrics in Table 2. The architecture and efficacy of the CNNs are detailed in Fig. 4, showcasing our model's robustness and adaptability across diverse datasets.

The results, especially the precision and recall metrics outlined in Table 2, underline the model's capability to navigate the intricacies of emotional nuances in speech. For each emotion class, the confusion matrix (Table 3) provides insights into the classification accuracy, supporting the model's reliability in practical scenarios. The emotional prediction pie chart (Fig. 6) visually represents the distribution of predicted emotional states, offering a user-friendly depiction of the model's predictive power.

Furthermore, the discussion connects the results to the identified problem area, showing how the model's nuanced understanding of psychoemotional states through vocal characteristics addresses gaps in current SER systems. The real-world applicability of the research is also highlighted, illustrating how the outcomes can contribute to advancements in sectors reliant on vocal communication.

In essence, the results can be traced back to the methodological choices made, as reflected in Fig. 3, where the process of transforming speech signals into informative features for CNN processing is outlined. The successful application of these methods emphasizes the potential of AI in psychoemotional state analysis, paving the way for future enhancements and wider application scopes [32].

The real-time processing capabilities of our model, as evidenced in the «Real-Time Processing Efficiency» section, open up numerous applications in dynamic environments where immediate emotion recognition is crucial. Potential applications include mental health monitoring in healthcare, enhancing user experience in human-machine interfaces, and safety monitoring in high-risk professions like aviation and security.

While our model marks a significant step forward, it is not without limitations. One key challenge is the potential overfitting to specific emotional expressions within the training datasets. Future work will focus on expanding the training corpus with more diverse emotional states and linguistic variations to mitigate this risk. Additionally, exploring unsupervised learning methods could offer solutions to this overfitting issue [6].

Advancing our study will involve refining algorithms to improve efficiency without sacrificing accuracy, especially in capturing the variability of speech in different emotional states. Gathering a comprehensive and representative dataset for further experimental validation remains a significant challenge [7].

In summary, our research presents a novel and effective method for emotion recognition through speech, marking a substantial contribution to the field of AI-driven psychoemotional state analysis. By addressing both the technical and practical aspects of SER, this work lays the groundwork for future advancements in automated human state monitoring systems.

7. Conclusions

1. The study's comprehensive analysis of current SER techniques has demonstrated a novel approach that outperforms traditional models by integrating advanced neural network architectures, leading to significant improvements in the classification of human psychoemotional states through voice.

2. The integration of deep learning algorithms has enabled the development of a robust model that exhibits high adaptability and accuracy across various emotional speech datasets, marking a substantial advancement in SER technology.

3. The dual-feature methodology, employing both mel-spec and MFCC, has proven to be highly effective in capturing the subtleties of emotional expression in speech, significantly enhancing the model's classification capabilities.

4. Validation with diverse datasets confirms the model's reliability and indicates its potential for widespread practical application, particularly in fields requiring nuanced emotion recognition.

5. The model's real-time processing proficiency underscores its practical applicability in dynamic environments, offering a powerful tool for monitoring and assessing psychoemotional states in various settings.

Conflict of interest

The authors declare that they have no conflict of interest in relation to this research, whether financial, personal, authorship or otherwise, that could affect the research and its results presented in this paper.

Financing

This research was funded by the Science Committee of the Ministry of Education and Science of the Republic of Kazakhstan (Grant No. AP09258659).

Data availability

Data will be made available on reasonable request.

Use of artificial intelligence

The authors have used artificial intelligence technologies within acceptable limits to provide their own verified data, which is described in the research methodology section.

Acknowledgements

This research was funded by the Science Committee of the Ministry of Education and Science of the Republic of Kazakhstan (Grant No. AP09258659).

References

1. Semigina, T., Vysotska, Z., Kyianytsia, I., Kotlova, L., Shostak, I., Kichuk, A. (2021). Psycho-Emotional State Of Students: Research And Regulation. Studies of Applied Economics, 38 (4). doi: https://doi.org/10.25115/eea.v38i4.4049

2. Amirgaliyev, Y. N., Bukenova, I. N. (2021). Recognition of a psychoemotional state based on video surveillance: review. Journal of Mathematics, Mechanics and Computer Science, 112 (4). doi: https://doi.org/10.26577/jmmcs.2021.v112.i4.11
3. Alharbi, S., Alrazgan, M., Alrashed, A., Alnomasi, T., Almojel, R., Alharbi, R. et al. (2021). Automatic Speech Recognition: Systematic Literature Review. IEEE Access, 9, 131858–131876. doi: https://doi.org/10.1109/access.2021.3112535
4. Khare, S. K., Blanes-Vidal, V., Nadimi, E. S., Acharya, U. R. (2024). Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations. Information Fusion, 102, 102019. doi: https://doi.org/10.1016/j.inffus.2023.102019
5. Jwaid, W. M., Al-Husseini, Z. S. M., Sabry, A. H. (2021). Development of brain tumor segmentation of magnetic resonance imaging (MRI) using U-Net deep learning. Eastern-European Journal of Enterprise Technologies, 4 (9 (112)), 23–31. doi: https://doi.org/10.15587/1729-4061.2021.238957
6. Lu, X. (2022). Deep Learning Based Emotion Recognition and Visualization of Figural Representation. Frontiers in Psychology, 12. doi: https://doi.org/10.3389/fpsyg.2021.818833
7. Ahmed, N., Aghbari, Z. A., Girija, S. (2023). A systematic survey on multimodal emotion recognition using learning algorithms. Intelligent Systems with Applications, 17, 200171. doi: https://doi.org/10.1016/j.iswa.2022.200171
8. Nosov, P., Zinchenko, S., Ben, A., Prokopchuk, Y., Mamenko, P., Popovych, I. et al. (2021). Navigation safety control system development through navigator action prediction by data mining means. Eastern-European Journal of Enterprise Technologies, 2 (9 (110)), 55–68. doi: https://doi.org/10.15587/1729-4061.2021.229237
9. de Lope, J., Graña, M. (2023). An ongoing review of speech emotion recognition. Neurocomputing, 528, 1–11. doi: https://doi.org/10.1016/j.neucom.2023.01.002
10. Costantini, G., Parada-Cabaleiro, E., Casali, D., Cesarini, V. (2022). The Emotion Probe: On the Universality of Cross-Linguistic and Cross-Gender Speech Emotion Recognition via Machine Learning. Sensors, 22 (7), 2461. doi: https://doi.org/10.3390/s22072461
11. Chen, P., Wu, L., Wang, L. (2023). AI Fairness in Data Management and Analytics: A Review on Challenges, Methodologies and Applications. Applied Sciences, 13 (18), 10258. doi: https://doi.org/10.3390/app131810258
12. Bankins, S., Formosa, P. (2023). The Ethical Implications of Artificial Intelligence (AI) For Meaningful Work. Journal of Business Ethics, 185 (4), 725–740. doi: https://doi.org/10.1007/s10551-023-05339-7
13. Mahdi, Q. A., Shyshatskyi, A., Prokopenko, Y., Ivakhnenko, T., Kupriyenko, D., Golian, V. et al. (2021). Development of estimation and forecasting method in intelligent decision support systems. Eastern-European Journal of Enterprise Technologies, 3 (9 (111)), 51–62. doi: https://doi.org/10.15587/1729-4061.2021.232718
14. Cai, Y., Li, X., Li, J. (2023). Emotion Recognition Using Different Sensors, Emotion Models, Methods and Datasets: A Comprehensive Review. Sensors, 23 (5), 2455. doi: https://doi.org/10.3390/s23052455
15. Coretta, S., Casillas, J. V., Roettger, T. B. (2022). Multidimensional signals and analytic flexibility: Estimating degrees of freedom in human speech analyses. doi: https://doi.org/10.31234/osf.io/q8t2k
16. Vakkantula, P. C. (2020). Speech Mode Classification using the Fusion of CNNs and LSTM Networks. West Virginia University. doi: https://doi.org/10.33915/etd.7845
17. Brownlee, J. (2021). Gentle Introduction to the Adam Optimization Algorithm for Deep Learning. Machine Learning Mastery. Available at: https://machinelearningmastery.com/adam-optimization-algorithm-for-deep-learning/
18. Hashem, A., Arif, M., Alghamdi, M. (2023). Speech emotion recognition approaches: A systematic review. Speech Communication, 154, 102974. doi: https://doi.org/10.1016/j.specom.2023.102974
19. Chauhan, C., Parida, V., Dhir, A. (2022). Linking circular economy and digitalisation technologies: A systematic literature review of past achievements and future promises. Technological Forecasting and Social Change, 177, 121508. doi: https://doi.org/10.1016/j.techfore.2022.121508
20. Kamath, U., Liu, J., Whitaker, J. (2019). Deep Learning for NLP and Speech Recognition. Springer International Publishing. doi: https://doi.org/10.1007/978-3-030-14596-5
21. Ekman, P. (1971). Universals and cultural differences in facial expressions of emotion. Nebraska Symposium on Motivation, 19, 207–283.
22. Rabiner, L. R., Schafer, R. W. (2007). Introduction to Digital Speech Processing. Foundations and Trends® in Signal Processing, 1 (1-2), 1–194. doi: https://doi.org/10.1561/2000000001
23. Livingstone, S. R., Russo, F. A. (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), e0196391. doi: https://doi.org/10.1371/journal.pone.0196391
24. Haq, S., Jackson, P. J. B. (2009). Speaker-Dependent Audio-Visual Emotion Recognition. In Proc. Int'l Conf. on Auditory-Visual Speech Processing, 53–58.
25. Pichora-Fuller, M. K., Dupuis, K. (2020). Toronto emotional speech set (TESS). doi: https://doi.org/10.5683/SP2/E8H2MF
26. Zhou, K., Sisman, B., Liu, R., Li, H. (2021). Seen and Unseen Emotional Style Transfer for Voice Conversion with A New Emotional Speech Dataset. ICASSP 2021 – 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). doi: https://doi.org/10.1109/icassp39728.2021.9413391

27. Zwicker, E., Fastl, H. (1999). Psychoacoustics. Facts and Models. Springer-Verlag, 417. doi: https://doi.org/10.1007/978-3-662-09562-1
28. Davis, S., Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech, and Signal Processing, 28 (4), 357–366. doi: https://doi.org/10.1109/tassp.1980.1163420
29. Goodfellow, I., Bengio, Y., Courville, A. (2016). Deep Learning. MIT Press. Available at: https://www.deeplearningbook.org/
30. Sundgren, D., Rahmani, R., Larsson, A., Moran, A., Bonet, I. (2015). Speech emotion recognition in emotional feedback for Human-Robot Interaction. International Journal of Advanced Research in Artificial Intelligence, 4 (2). doi: https://doi.org/10.14569/ijarai.2015.040204
31. Martin, O., Kotsia, I., Macq, B., Pitas, I. (2006). The eNTERFACE' 05 Audio-Visual Emotion Database. 22nd International Conference on Data Engineering Workshops (ICDEW'06). doi: https://doi.org/10.1109/icdew.2006.145
32. Koshekov, K. T., Savostin, A. A., Seidakhmetov, B. K., Anayatova, R. K., Fedorov, I. O. (2021). Aviation Profiling Method Based on Deep Learning Technology for Emotion Recognition by Speech Signal. Transport and Telecommunication Journal, 22 (4), 471–481. doi: https://doi.org/10.2478/ttj-2021-0037