**The procedure of adapting the design parameters of the convolutional neural network during the speaker’s emotions recognition.**

|  |  |  |
| --- | --- | --- |
| Ihor Tereikovskyi National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute" Kyiv, Ukraine [0000-0003-4621-9668] | Shynar Mussiraliyeva Al-Farabi Kazakh National University Almaty, Kazakhstan [0000-0001-5794-3649] | Liudmyla Tereikovska Kyiv National University of Construction and Architecture Kyiv, Ukraine [0000-0002-8830-0790] |
|  |  |  |
| Denys Chernyshev Kyiv National University of Construction and Architecture Kyiv, Ukraine [0000-0002-1946-9242] | Adlet Nyussupov Al-Farabi Kazakh National University Almaty, Kazakhstan [0000-0003-3254-0901] | Yerulan Abaiuly Al-Farabi Kazakh National University Almaty, Kazakhstan [0000-0003-2248-3819] |

**Abstract.** Solving the challenge of improving the recognition system of the speaker's emotions due to the implementation of modern neural networks that are based on convolutional neural networks is the subject of this article. Difficulties of such implementation are identified and there is connection with the need of adapting the convolutional neural network to conditions of given recognition problem. The functional network parameters of adaptation procedure are provided, that implies determination of input and output field parameters, type, values of structural settings and mathematical basics of the convolutional neural networks. Procedure is based on projection of a mel-frequency cepstral coefficient for every quasi-stationary fragment of voice signals with fixed size in the form of mono-chrome square picture. The conducted research experiments showed that the provided procedure allows to build the convolutional neural networks that has speaker’s emotion recognition accuracy at the level of better modern recognition systems. Further researches will be connected to the development of the method that is based on emotional neural networks in free texts voicing cases. This will help to identify and prevent possible threats that come from fake audio and video information. The level of overall information security can be increased using aforementioned technologies.

**Keywords:** recognition of emotions, speech signal mel-frequency cepstral coefficient, speaker, neural network model, convolutional neural network, adaptation procedure, information security.

1. **Introduction**

In this article we will consider the introduction of a speech native analysis tool designed to recognize the speaker's personality and emotions to improve modern information systems. We presume that using these tools in conjunction will improve the efficiency of a fairly wide range of information systems and improve methods of information security. For instance, using instruments to extract certain features from the speaker’s voice will presumably gradually improve converging of the network to the state, where it becomes efficient. Overall, such system could help to recognize emotional state of people who are responsible for very critical tasks and allow us to prevent some catastrophes that are tend to happen because of their bad emotional state. In addition, such an assessment is relevant in the systems security, in the banking sector, advertising business, as well as in assessing the quality of call centers. Although quite a lot of studies have been devoted to the development of technologies for recognizing emotions by voice [11, 13, 16, 26], but practical experience and results of scientific and applied works [1, 9, 18] indicate to the need of significant modernization. Emotions recognition is crucial for the information security field and it will help to stop cyber crimes within the internet. Therefore we can see, that the problem of emotion recognition using analysis of speech features is relevant for us.

1. **Literature and Sources Analysis in the Research Field**

In accordance with [2, 27], a speech signal should be understood as a composite audio signal which is retrieved from the speaker’s vocal apparatus. The difficulty of analyzing audio signals is in both the high variability between the speakers and within the speakers, also it becomes more difficult because of variance in the duration of their speech. Because of this emotionally colored speech processing technologies built to use various numerical characteristics that are stable to recognize emotional state of the speaker.

Emotional coloring of speech is analyzed using features that are derived from frequency and time. A relationship was found between pitch of the voice, speed of speech, volume of speech, and emotional state. For instance, emotions of anger, joy and fear are expressed by speakers using loud, fast speech with wide range of pitch, while, fatigue, apathy and sadness are expressed using slow speech with low timbre and slurred speech.

Following stages can be observed when dealing with audio signal recognition technology: receiving audio signal in analog form, sampling the signal, noise reduction, isolation of quasi-stationary fragments, picking out the features and recognition[6, 10]. Techniques for the first 5 stages are well developed and do not cause difficulties. Mostly derived quasi-stationary fragments have the range of duration of 0.01-0.02 s, while the frequency of samples are 8000 Hz and more. Those techniques are accepted to be the most effective for high-quality speech signal recognition [13, 14, 17], as it is discovered that analysis spectrum should be in the range from 50 to 4000 Hz. Mel-frequency cepstral coefficients (MFCC) calculated from the results of spectral analysis for each quasi-stationary fragments (QF) and usually used as informative signs of a speech signal [10, 22, 28]. Recognition quality also depends on training data being used as testing data, so we must not use the same data to test our solution . Some of the well-known speech corpuses are: for English speech - eNTERFACE, TESS, SAVEE, for Russian - RUSLANA, REC, for German - EmoDB and for Turkish speech - BUEmoDB.

These corpuses consist of recordings of speech of many speakers, which provide us with voice signals with a range of emotional coloring. Number of emotional states vary from 6 (anger, fear, disgust, sadness, happiness, surprise) to 22. At the same time, literature data indicate that the functioning of the mentioned recognition modules is based on mathematical models based on the Bayesian approach [2], the theory of deterministic chaos [9], clustering [1], hidden Markov processes. [18], support vector machines [3, 12], methods of the theory of fuzzy logic [20], “Divide and conquer” algorithms [27], as well as the neural network theory[10, 13]. At the same time, both the data of modern literary sources [5, 28] and the results of the study of the commercial technologies, such as Siri and VoiceNavigator, suggest that analyzing voice signals using neural networks has broad prospects.

Speech recognition feature using an LSTM-type neural network, which allows processing dynamic data series, is considered in [10]. Some of standard neural network methods are considered in the articles [2, 23]. These articles concluded that types of the neural network considered in the work are not suitable for effective speech recognition dur no low accuracy. It is also assumed, that constructing and training an LSTM network will meet too much difficulties due to the complexity of forming the training sample. At the same time a convolutional neural network (CNN) may succeed in processing audio signals, since they can be transformed to the suitable format for CNN. It is assumed to achieve the speaker recognition accuracy at the level of 96-97%, which is declared in [5, 15, 21]. Similar results were also obtained in [16], in which CNNs were analyzing speaker’s emotions from the voice signals. At the same time, the achieved recognition accuracy is at the level of 70-80%. However, since characteristics and features obtained for CNN can vary dramatically, we assume that it is possible to achieve more accurate results. Since CNNs are usually used for computer vision [4, 7, 24] and require image-like data, there are many ways to construct such data from an audio signal, which we will be using to achieve needed results. Therefore, there is an interest to use modern types of CNN for voice signal analysis. It is worth to note that SqueezeNet, which is one of the most popular and tested types of modern CNN, has some characteristics that make it suitable for our tasks, those characteristics being high recognition accuracy and low resource intensity, which in some cases has decisive importance for monitoring systems of the speaker's emotional state.

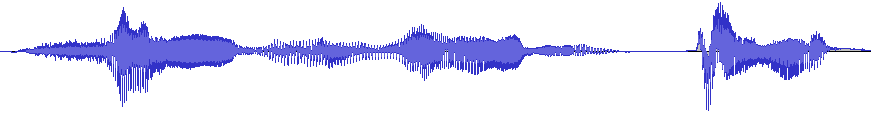
Convolutional neural network is a further development of the multilayer perceptron that implements the hierarchical concept of image classification.

It should be noted that the generally accepted technique for analyzing a speech signal, in addition to cleaning and normalizing noise, provides for its division into quasi-stationary frames and primary processing in order to reduce the volume and increase the information content of the analyzed parameters. The result of this processing is a set of mel-frequency cepstral coefficients.

Therefore, the main goal of this study is to develop a procedure for adapting the parameters of a convolutional neural network to the conditions of the problem of recognizing the speaker's emotions at fixed sections of the speech signal.

1. **Adaptation procedure**

The adaptation procedure can be divided into five stages, designed to determine the parameters of the output signals, the parameters of the input field, CNN type, parameters structural values, CNN math basis, in accordance with the results of studies [7, 24] devoted to the construction of neural network models intended for the analysis of biometric parameters. These parameters of the neural network are highly dependent of parameters of training samples and number of training samples. Also, the expediency of using the TESS database was determined in data [20, 28] as a source of training examples, which can be downloaded from the link https://doi.org/10.5683/SP2/E8H2MF. This database contains records of two female voices. Each individual voice recording reflects one of seven basic emotions (happiness, surprise, sadness, neutrality, anger, disgust, fear). Single recording duration 1.2 - 2 s, sampling frequency 24414 Hz, mono channel, sampling depth 32 bit. The recordings are stored in separate wav files, the names of which correspond to the emotional coloring of speech, the speaker's identifier and the content of the spoken text. Speakers are professional actors. One of them was a 26-year-old woman, and the other 64, who voiced overlapping English texts. In all cases, the phrase "Say the word" was repeated. After that phrase she says one of 200 predefined words (boat, cab, dab,…). For example, in Fig. 1 shows a sonogram of a speech signal with a "happiness" emotional coloring when the second speaker pronounces the text "Say the word boat". A fig. 2 shows the corresponding sonogram with a neutral emotional coloring.



**Fig. 1.** Sonogram of the voice text "Say the word boat" with a happy emotional coloring



**Fig. 2.** Sonogram of the dubbed text "Say the word boat" with a neutral emotional coloring.

Since there are two speakers, the number of examples corresponding to one emotion is 400. The total size of the TESS database is 2800 examples.

Based on the listed functions of the TESS database, it is identified that CNN output neurons can be associated with recognizable basic emotions. Therefore, the number of output neurons Ny = 7. The positions that ensure the interpretability of the classification results in the CNN output layer and the softmax activation function are used by an expression in the form:

, (1)

where *yi* - output of the i-th neuron within the output layer, *sk* - common input signal for the k-th neuron of the output layer.

A feature of neural network models such as CNN is the need to represent the input field as a square image with one or more color channels. The known options for representing the input field in the form of a black-and-white, gray or color image depend on the characteristics of the recognition problem. Since the possibility of effectively describing a speech signal using mel-frequency cepstral coefficients is considered to be proven, one point of the CNN input field can be associated with a separate mel-frequency cepstral coefficient. The number of such points is calculated using an expression of the form:

, (2)

where *Z* - the number of mel-frequency cepstral coefficients characterizing the analyzed speech signal, *K*qt – the number of analyzed quasi-stationary fragments (frames) of the speech signal, *Kmel* – the number of mel-frequency cepstral coefficients characterizing one quasi-stationary fragment.

In turn, in the absence of overlap of quasi-stationary fragments, their number on the analyzed fragment of the speech signal can be calculated as follows:

, (3)

where T - the duration of the analyzed fragment of the speech signal, *tqt* – the duration of the quasi-stationary fragment, Round – the rounding function to the smallest integer.

Note that the admissible duration of a quasi-stationary fragment varies from 10 ms to 40 ms. In addition, during duration determination, one should take into account the limitation associated with the implementation of the fast Fourier transform, which means that the number of recorded points of the speech signal should be close to 2*r*. An analytically specified constraint can be displayed as follows:

, (4)

where *kqt*  - the number of registrations of the speech signal on one quasi-stationary fragment, *f* – the speech signal sampling frequency, *r* – a positive integer.

Since the recommended number of mel-frequency coefficients for one quasi-stationary fragment should be equal to *Kmel*=10, and also taking into account the fact that for examples from the TESS database *f*=24414 Hz, T is in the range from 1.2 s to 2 s, *tqt* is in in the range from 10 ms to 40 ms and using expressions (1-4) it was obtained that the number of mel-frequency cepstral coefficients characterizing the analyzed speech signal is Z = 1140. Thus, the analyzed speech signal can be correlated with a vector, 1140 components of which correspond to the mel-frequency cepstral coefficients of the given speech signal.

Let us consider the transformation of the obtained vector of mel-frequency cepstral coefficients to a form suitable for submission to CNN, i.e. into a square matrix with the same number of elements. The specified matrix can be represented as a single-channel square image in a Cartesian coordinate system. It is proposed to correlate the y-axis coordinate with the number of the mel-frequency cepstral coefficient within one quasi-stationary fragment, and correlate x-axis coordinate with the number of the quasi-stationary fragment of the speech signal. The color of an individual point of the image with coordinates (*xn,ym*) will correspond to the value of the m-th mel-frequency cepstral coefficient for the n-th quasi-stationary fragment. Because of the fact that the count of quasi-stationary fragments of the audio signal is more than number of the mel-frequency cepstral coefficients, 10 lines at the top of the image are filled with zeros to preserve the square shape. As a result, the CNN input field is associated with a 1140x1140 matrix, the elements of which correspond to the values ​​of the mel-frequency cepstral coefficients of the speech signal. For example, the result of the formation of the CNN input field for speech signals when the phrase “Say the word boat” is said with the emotional coloring happy, the sonogram of that audio signal is represented in Fig. 1 and is fragmentarily shown on Fig. 3.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Y5 | -13,95 | -18,62 | 9,62 | 8,09 | 11,08 |
| Y4 | -14,13 | -16,11 | 17,04 | 8,18 | 13,24 |
| Y3 | -11,45 | -22,95 | -2,28 | 26,07 | 1,34 |
| Y2 | -9,04 | -27,60 | -2,66 | 28,85 | -10,41 |
| Y1 | -7,43 | -34,40 | -3,78 | 25,34 | -15,41 |
|  | X1 | X2 | X3 | X4 | X5 |

**Fig. 3.** Fragment of the CNN input field when dubbing the text "Say the word boat" with the emotional coloring happy

Note that the proposed stage of determining the CNN input field was developed on the basis of the methods proposed in [4, 7] for coding CNN input data used for computer viruses recognition and neural network analysis of the user's face image.

Determining the parameters of the input and output fields made it possible to proceed to the next stages of adapting the CNN parameters to the conditions of the problem of recognizing the speaker's emotions. In the basic version, it is customary to use the most proven LeNet type, which includes LeNet-5, AlexNet, VGG-16, VGG-16. The main design parameters of this type of CNN are: number of convolution layers (*LCNN*), the number of convolution maps in each i-th convolutional layer (*Lh,i*), convolution kernel size for each i-th convolutional layer (*bi*), shift of the convolution kernel for each i-th convolutional layer (*ri*), number of downsampling layers(*LS*), scale factor for each j-th downsampling layer (*mj*), number of fully connected layers (*LFC*), the number of neurons in each kth fully connected layer (*Lf,k*). The variable parameters of the software of such a CNN are the type and parameters of the activation function for each layer of neurons. To calculate the rest of the structural parameters, the expressions substantiated in [1, 9] can be used:

, (5)

, (6)

P, (7)

(8)

where *HCNN,i*  - the size of the feature map for the i-th convolutional layer, *ri* is shift of the convolution kernel for the i-th convolutional layer, *HS,j* - size of the j-th downsampling layer, *Lh,end* - number of feature maps in the last convolutional layer, *Lf,s* is the total number of neurons in all fully connected layers, *P* - number of training examples.

By analogy with [7, 24], it is proposed to optimize the design parameters of CNN based on the fact that the process of recognizing an input image by a neural network should be as similar as possible to how a person recognizes a similar image on a computer monitor. Integration of this approach with the concept of CNN functioning made it possible to formulate the following adaptation principles:

Principle 1. The size of the kernel and the value of the shift of the convolution kernel for the n-th convolutional layer must correspond to the size of the recognizable features at the corresponding hierarchical level.

Principle 2. Number of layers in CNN and number of levels of human recognition of a two-dimensional image on a computer screen must be similar.

Principle 3. The feature maps count in the n-th convolutional layer must be the same as the number of features at the n-th level of recognition.

Based on the data [25], it was determined that a person is able to distinguish sound impulses with a duration of about 100 ms, which, under the conditions of the set task of recognizing emotions, corresponds to 9 quasi-stationary fragments. Taking into account principle 1, it is determined that, in the first approximation, the size of the convolution kernel for all convolutional layers is equal to *b*=9, and the shift for the first convolutional layer is *r1*=3. For the rest of the convolutional layers, the shift is *ri*=1.

Using principle 2, on the basis of the data [25, 27] and the results of expert assessment, it was determined that the number of levels of human recognition of an image with a size of 1140x1140 is 6. Based on the expediency of using one fully connected layer of neurons in CNN (*LFC*=1) it is calculated that the number of CNN convolutional layers is *LCNN*=5. Number of downsampling layers *LFC*=5. Scale factor for all layers is *m*=2.

Also, as a result of the analysis of the black-and-white version of the images corresponding to the CNN input field, it was determined that in this case, at the first level of the recognition hierarchy, it is possible to detect six elementary features with a size of 9 × 9 points. Four features are segments oriented in different ways in space (vertical, horizontal, ±45°). Two more signs correspond to a completely black or white square. Since the task of recognition implies the need to take into account color shades, the number of such elementary features has been increased up to 36. Therefore, in the first layer of CNN, the count of feature maps is *Lh,1*=36. In other convolutional layers, by analogy with AlexNet, the number of feature maps is taken to be equal to 256.

Structural parameters of CNN adapted to the task of recognizing the speaker's emotions at fixed sections of speech signals from the TESS database are presented in Table. 1.

**Table 1.** Structural parameters of CNN

|  |  |
| --- | --- |
| Parameter name | Parameter value |
| Input field size | *HCNN,0*  =1410 |
| Number of convolution layers | *LCNN* =5 |
| Number of convolution cards | *Lh,1* = 36,  *Lh,2* = *Lh,3* = *Lh,4* = *Lh,4* =256 |
| Convolution map size | *HCNN,1*  =378, *HCNN,2*  =182, *HCNN,3*  =84, *HCNN,4*  =34, *HCNN,5*  =10 |
| Convolution kernel size | *b* =9 |
| Shifting the convolution kernel | *r1* =9,  *r2* = *r3* = *r4* = *r5* =1 |
| Number of downsampling layers | *LS*=5 |
| Scale factor of sub-sampling | *mj*=2 |
| Number of fully connected layers | *LFC*=1 |
| Number of neurons in a fully connected layer | *Lf,k* = 1000 |
| Number of output neurons | *N*y=7 |

At the last stage of the adaptation procedure, based on the data [7], it was determined that in the convolutional layers and in the fully connected CNN layer, that the most effective activation function ReLU should be used given by the expression:

, (9)

where *xs,i* – sum of inputs to the neuron i.

When calculating the output signal of neurons in the subsampling layer, the max pooling mechanism is used. It is characterized by the following expression:

(10)

The developed CNN model was implemented programmatically using one of the most popular libraries for this task, which is Tensorflow on Python language, which made it possible to carry out computer experiments aimed at verifying the proposed adaptation procedure. A personal computer with the following characteristics was used for the experiments: AMD Ryzen 7 3700X (3.6 - 4.4 GHz) / 16 GB of RAM / 1 TB HDD + SSD 512 GB / nVidia GeForce GTX 1660 Ti, 6 GB / Windows 10. The experimental plan included determining the accuracy of emotion recognition when changing the design parameters of CNN. For example, Fig. 4 shows the dependences of the emotion recognition accuracy on the number of learning epochs for a different count of feature maps in first convolutional layer. 256 feature maps are located in other layers.

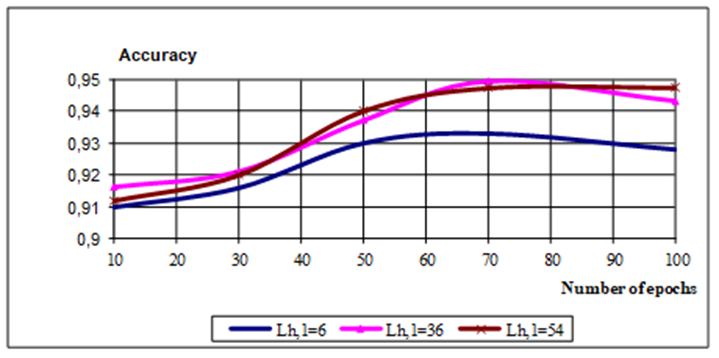
Based on the results of the research, it is uncovered that speaker’s emotion recognition network should be able to achieve goals such as:

* discovering multiple features that characterize speaker’s voice,
* recording and storing data about the features of speaker’s voice,
* transforming arrays of features into matrixes and trying to adapt it to feed to other layers of the network,
* transforming matrices into image-like structures, so that CNN can work with them more effectively,
* using discovered features network needs to effectively determine which emotion is present on audio.

It may be needed that some of the steps are performed by the deterministic system to guarantee results, rather than using the neural network, as it will guarantee 100% correct result. But it may just be more time-efficient in-use if another neural network used instead, as it will not require that many computations.

We would also need to add functions of forming image-like structure from the matrix and teaching neural network using data to the system, but those will be purely deterministic.

The maximum accuracy of recognition of the speaker's emotions, equal to about 0.95, was achieved with the values ​​of the CNN design parameters given in Table 1. The results obtained generally correspond to the best indicators of similar systems [5, 10, 15, 18, 19] and allow us to assert about the successful verification of the proposed adaptation procedure. It can also be stated that further advancements could be made on a neural network method for recognizing the speaker's emotions when dubbing free texts.



**Fig. 4.** Dependence of recognition accuracy on the number of learning epochs for different values *Lh,1*

1. **Conclusions**

The article is devoted to solving the problem of improving the systems for recognizing the emotions of the speaker through the introduction of modern neural network solutions based on CNN. It was established that the implementation of such an introduction is associated with the need to adapt the design parameters of the CNN to the conditions of the recognition problem. A corresponding adaptation procedure has been formed, which provides for the determination of the parameters of the input and output fields, the type, the values of the structural parameters and the CNN software.

The effectiveness of the proposed adaptation procedure was confirmed by computer experiments, which showed the possibility of achieving the accuracy of recognition of the speaker's emotions at the level of 0.95 on a fairly limited training sample with 70 epochs of training, which is commensurate with the results of the best modern systems for a similar purpose.

It is also shown that the ways for further research are associated with the development of a neural network method for recognizing the speaker's emotions when dubbing free texts.

This work was supported by the grant "Development of models, algorithms for semantic analysis to identify extremist content in web resources and creation the tool for cyber forensics" funded by the Ministry of Digital Development, Innovations and Aerospace industry of the Republic of Kazakhstan. Grant No. IRN AP06851248.

**References**

1. M. M. Kabir, M. F. Mridha, J. Shin, I. Jahan and A. Q. Ohi, “A Survey of Speaker Recognition: Fundamental Theories, Recognition Methods and Opportunities,” IEEE Access, vol. 9, pp. 79236-79263, 2021.

2. Rashid Jahangir, Ying Wah Teh, Henry Friday Nweke, Ghulam Mujtaba, Mohammed Ali Al-Garadi, Ihsan Ali, “Speaker identification through artificial intelligence techniques: A comprehensive review and research challenges,” Expert Systems with Applications, vol. 171, 2021.

3. Y. Banaras, A. Javed and F. Hassan, “Automatic Speaker Verification and Replay Attack Detection System using novel Glottal Flow Cepstrum Coefficients,” 2021 International Conference on Frontiers of Information Technology (FIT), pp. 149-153, 2021.

4. I. Dychka, D. Chernyshev, I. Tereikovskyi, L. Tereikovska, V. Pogorelov, “Malware Detection Using Artificial Neural Networks,” Advances in Computer Science for Engineering and Education II, vol. 938, pp. 3-12, 2020.

5. Kong Aik Lee, Ville Vestman, Tomi Kinnunen, “ASVtorch toolkit: Speaker verification with deep neural networks,” SoftwareX, vol. 14, 2021.

6. X. Liu, M. Sahidullah and T. Kinnunen, “Learnable MFCCs for Speaker Verification,” 2021 IEEE International Symposium on Circuits and Systems (ISCAS), pp. 1-5, 2021.

7. R. Salih Kuzu, E. Maiorana and P. Campisi, “Loss Functions for CNN-based Biometric Vein Recognition,” 2020 28th European Signal Processing Conference (EUSIPCO), pp. 750-754, 2021.

8. B. Qiang et al., “SqueezeNet and Fusion Network-Based Accurate Fast Fully Convolutional Network for Hand Detection and Gesture Recognition,” IEEE Access, vol. 9, pp. 77661-77674, 2021.

9. B. J. Abbaschian, D. Sierra-Sosa, and A. Elmaghraby, “Deep Learning Techniques for Speech Emotion Recognition, from Databases to Models,” Sensors, vol. 21, p. 1249, Feb. 2021.

10. T. Kumar, K. O. Villalba-Condori, D. Arias-Chavez, K. Rajesh, M. K. Chakravarthi, S. S. Rajest, “An Evaluation on Speech Recognition Technology based on Machine Learning,” Webology, vol. 19, pp. 646-663, Jan. 2022.

11. P. Laukka, H. A. Elfenbein, “Cross-cultural emotion recognition and in-group advantage in vocal expression: A meta-analysis,” Emotion Review, vol. 13, pp. 3-11, Jan. 2021.

12. S. Nainan, V. Kulkarni, “Enhancement in speaker recognition for optimized speech features using GMM, SVM and 1-D CNN,” International Journal of Speech Technology, vol. 24, pp. 809-822, 2020.

13. H. Dahrouj et al., “An Overview of Machine Learning-Based Techniques for Solving Optimization Problems in Communications and Signal Processing,” IEEE Access, vol. 9, pp. 74908-74938, 2021.

14. V. Makarova, V. A. Petrushin, “RUSLANA: a database of Russian emotional utterances,” Conference of Spoken Language Processing, pp. 2041-2044, 2002.

15. S. H. Kim, Y. H. Park, “Adaptive convolutional neural network for text-independent speaker recognition,” INTERSPEECH 2021, pp. 641-645, 2021.

16. M. Swain, A. Routray, P. Kabisatpathy, “Databases, features and classifiers for speech emotion recognition: a review,” International Journal of Speech Technology, vol. 21, pp. 93-120, 2018.

17. R. K. Srivastava, R. Shree, A. K. Shukla, R. P. Pandey, V. Shukla, D. Pandey, “A Feature Based Classification and Analysis of Hidden Markov Model in Speech Recognition,” Cyber Intelligence and Information Retrieval, vol. 291, pp. 365-379, 2021.

18. L. Schoneveld, A. Othmani, H. Abdelkawy, “Leveraging recent advances in deep learning for audio-visual emotion recognition,” Pattern Recognition Letters, vol. 146, pp. 1-7, 2021.

19. R. Li, J. Y. Jiang, J. L. Li, C. C. Hsieh, W. Wang, “Automatic speaker recognition with limited data,” Proceedings of the 13th International Conference on Web Search and Data Mining, pp. 340-348, 2020.

20. K. K. Sahoo, I. Dutta, M. F. Ijaz, M. Woźniak and P. K. Singh, "TLEFuzzyNet: Fuzzy Rank-Based Ensemble of Transfer Learning Models for Emotion Recognition From Human Speeches," IEEE Access, vol. 9, pp. 166518-166530, 2021.

21. L. Yu, J. Zeng, S. Wang, Y. Zhang, “Phonetic Encoding Contributes to the Processing of Linguistic Prosody at the Word Level: Cross-Linguistic Evidence From Event-Related Potentials,” Journal of Speech, Language, and Hearing Research, vol. 64, pp. 4791-4801, 2021.

22. V. V. Savchenko, “The Principle of the Information-Divergence Minimum in the Problem of Spectral Analysis of the Random Time Series Under the Condition of Small Observation Samples,” Radiophysics and Quantum Electronics, vol. 58, pp. 373-379, 2015.

23. V. V. Savchenko V.V, “Enhancement of the Noise Immunity of a Voice-Activated Robotics Control System Based on Phonetic Word Decoding Method,” Journal of Communications Technology and Electronics, vol. 61, pp.1374-1379, 2016.

24. M. Rakhimov, T. Boburkhon and T. Khurshid, "Speaker Separation: Use Neural Networks," 2021 International Conference on Information Science and Communications Technologies (ICISCT), pp. 01-03, 2021.

25. S. Feng, J. Wei, D. Wang, X. Yang, Z. Yang, Y. Zhang, G. Yu, “SINN: A speaker influence aware neural network model for emotion detection in conversations,” World Wide Web, vol. 24, pp. 2019-2048, 2021.

26. M. K. Yadav, V. Bhateja, M. Singh, “Exploration of Wavelets for Pre-processing of Speech Signals,” Computer Communication, Networking and IoT, vol. 197, pp. 475-483, 2021.

27 O. Yudin, S. Toliupa, O. Korchenko, L. Tereikovska, I. Tereikovskyi and O. Tereikovskyi, “Determination of Signs of Information and Psychological Influence in the Tone of Sound Sequences,” 2020 IEEE 2nd International Conference on Advanced Trends in Information Theory (ATIT), pp. 276-280, 2020.