Journal of Theoretical and Applied Information Technology
15th March 2021. Vol.99. No 5

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

RECOGNITION OF SPEAKER'S EMOTION BY SQUEEZENET CONVOLUTIONAL NEURAL NETWORK

LIUDMYLA TEREIKOVSKA[1], IHOR TEREIKOVSKYI[2],

AIMAN BEKETOVA[3] , GABIT KARAMAN[4] AND NADIIA MAKOVETSKA[5]

[1] Kyiv National University of Construction and Architecture, Kiev, Ukraine
[2] National Technical University of Ukraine"Igor Sikorsky Kyiv Polytechnic Institute", Kiev, Ukraine
[3]Al-Farabi Kazakh National University, Almaty, Kazakhstan
[4]Al-Farabi Kazakh National University, Almaty, Kazakhstan
[5] National Technical University of Ukraine"Igor Sikorsky Kyiv Polytechnic Institute", Kiev, Ukraine

e-mail: [1] tereikovskal@ukr.net , [2]terejkowski@ukr.net,
[3]aiman.beketova@gmail.com ,[4]gabyt.kaz@gmail.com

ABSTRACT

The article deals with the development of neural network means for analyzing a voice signal to recognize the speaker's emotions. We have established the possibility of improving these means through the use of a convolutional neural network of the SqueezeNet type, which determines the necessity to assess the effectiveness of such use. We have also determined that it is possible to assess the efficiency of using the neural network model experimentally by means of indicators of recognition accuracy and duration of training. A software implementation of SqueezeNet has been developed, with a training sample formed, using the publicly available TESS database, consisting of samples of voice signals with 7 emotions for 2 users. Mel-frequency cepstral coefficients are used as the parameters characterizing a voice signal. Using computer experiments, we have found that after 80 periods of training on a fairly limited training sample, SqueezeNet enables using validation examples to achieve speaker recognition accuracy of about 0.95, which is proportionate to the results of the best modern systems of the similar purpose and confirms the possibility of effective use of this type of network for analyzing a voice signal. We have shown the necessity for further research related to the adjustment of neural network solutions to the recognition of the speaker's emotions under a variety of noise interference. We have also determined the feasibility of developing a method for adjusting SqueezeNet architectural parameters to the conditions of the task to analyze a voice signal for simultaneous recognition of the speaker's personality and emotions.

Keywords: *Voice Signal, Emotion Recognition, Neural Network Model, Convolutional Neural Network, SqueezeNet*

1. INTRODUCTION

One of the most important trends in the development of information systems for various purposes is the widespread introduction of voice signal analysis means. Such means have long been successfully used to authenticate users both while entering the information system and during its operation. A peculiar feature of the modern period is an increased interest in voice analysis means designed to recognize the speaker's emotions. This interest is based on the well-known fact that information transmitted by people through words amounts to only 5-10% of the total volume of data in the process of interpersonal communication [1, 2]. Non-verbal signals, such as facial expressions, poses, gestures, touches, smells, account for more than half of all transmitted information, with the voice paralinguistic component being significant. Therefore, it is assumed that the integrated use of the speaker's personality and emotions recognition means will increase the effectiveness of information systems used in various spheres of activity. For example, using voice analysis means enables to provide effective recognition of the quality of educational materials in distance learning systems.

This is a key element in developing means for automated delivery of educational materials, adjusted to the abilities of an individual student. There exist well-known examples of the introduction of voice recognition of emotions to assess customer impressions of advertised products and the quality of call centres operation. In addition, it seems important to determine the psycho-emotional state of operators of critical infrastructure objects based on their voices.

At the same time, practical experience and the results of scientific and experimental works [2-4] indicate the necessity to significantly improve voice recognition modules for reducing resource intensity, increasing recognition accuracy, expanding the range of recognizable emotions, reducing the development time and increasing the level of adjustment to other conditions of application. Thus, one can explain the relevance of the scientific and experimental problem of improving the technology of the voice signal analysis to recognize the speaker's emotions.

2. LITERATURE REVIEW

In accordance with [5, 6], the concept of a voice signal shall mean a complex acoustic signal, the source of which is the human vocal apparatus. Complex analysis of voice signals is caused by high interspeaker and intraspeaker variability, as well as the difference in their duration. Therefore, technologies for processing emotionally coloured speech are designed to recognize the emotional state of the user through numerical analysis of the stable features of voice signals.

To analyze the emotional colour of speech, features are used which can be divided into a frequency and temporal ones. The relationships between pitch, speed, voice volume and emotion are established. For example, speech in a state of fear, anger or joy becomes fast and loud, with a wide range of pitch. In turn, fatigue, apathy or sadness is characterized by slow, low timbre and slurred speech.

A common typical technology for recognizing a voice signal consists of the following stages: registration of an analogue signal, it's sampling, noise removal, separation of quasistationary fragments, determination of informative features and recognition itself [7, 8]. Today implementation of the first five stages is believed to cause no special difficulties. In this case, the duration of a quasistationary fragment is usually within 0.01-0.02s, with the sampling frequency of at least 8,000Hz. This is due to the generally accepted thesis [9-11] that it is necessary to analyze the spectrum within the range of 50-4,000Hz for high-quality recognition of a voice signal.

As informative features of a voice signal, Mel-frequency cepstral coefficients (MFCC) are usually used, calculated on the basis of the results of the spectral analysis for each of the quasistationary fragments (QF) [8, 12, 13]. Also, there is a close dependence of the quality of recognition on the speech corpora (databases of voice signals) used to build recognition modules. The most famous speech corpora are: for English speech – eNTERFACE, TESS, SAVEE, for Russian – RUSLANA (RUSsian LANguage Affective speech), REC (Russian Emotional Corpus) [10], for German – EmoDB and for Turkish one – BUEmoDB [4].

We should note that these corpora include audio recordings of voice signals of various speakers, which reflect from 6 (anger, disgust, fear, sadness, happiness, surprise) to 22 types of emotional states [14-15]. Also, the data indicate that the above recognition modules function on the basis of mathematical models grounded on the Bayesian approach [5], the theory of determined chaos [4], hidden Markovian processes [2], support vector machines [16, 17], dynamic programming methods [6], as well as the theory of neural networks[8, 9]. Moreover, modern literary sources [18] and the results of studying the most well-known means of a similar purpose (Google+, Microsoft Office, VoiceNavigator, Siri) suggest that the neural network methodology for analyzing a voice signal is very promising.

The work [8] considers the possibility of recognizing a voice signal by means of a neural network of the LSTM type, which enables processing dynamic data series. The use of classical types of neural network models is considered in [5, 19]. However, the classical types of neural network models do not allow achieving acceptable recognition accuracy, with the construction of the LSTM network associated with the complexity of the training sample. Moreover, in [18, 20, 21] a convolutional neural network (CNN) is used to recognize a voice signal, due to which speaker recognition accuracy of 96–97% is achieved. Similar results are also obtained in [22], where CNN is used to analyze a voice signal in order to recognize the speaker's emotions, with the achieved recognition accuracy of 70-80%.

We should note that the characteristics of different types of CNN differ greatly, as they are adjusted to solve quite different tasks, related to computer vision [23-25]. Therefore, it is of interest

to study possibilities of analyzing a voice signal by means of modern types of CNN.

One of the most tested modern types of CNN is SqueezeNet, with relatively low resource intensity and sufficient recognition accuracy as the characteristic features, which in some cases is crucial for systems for monitoring the speaker's emotional state [26]. Figure 1 shows the SqueezeNet network structure implemented using MATLAB R2018b.

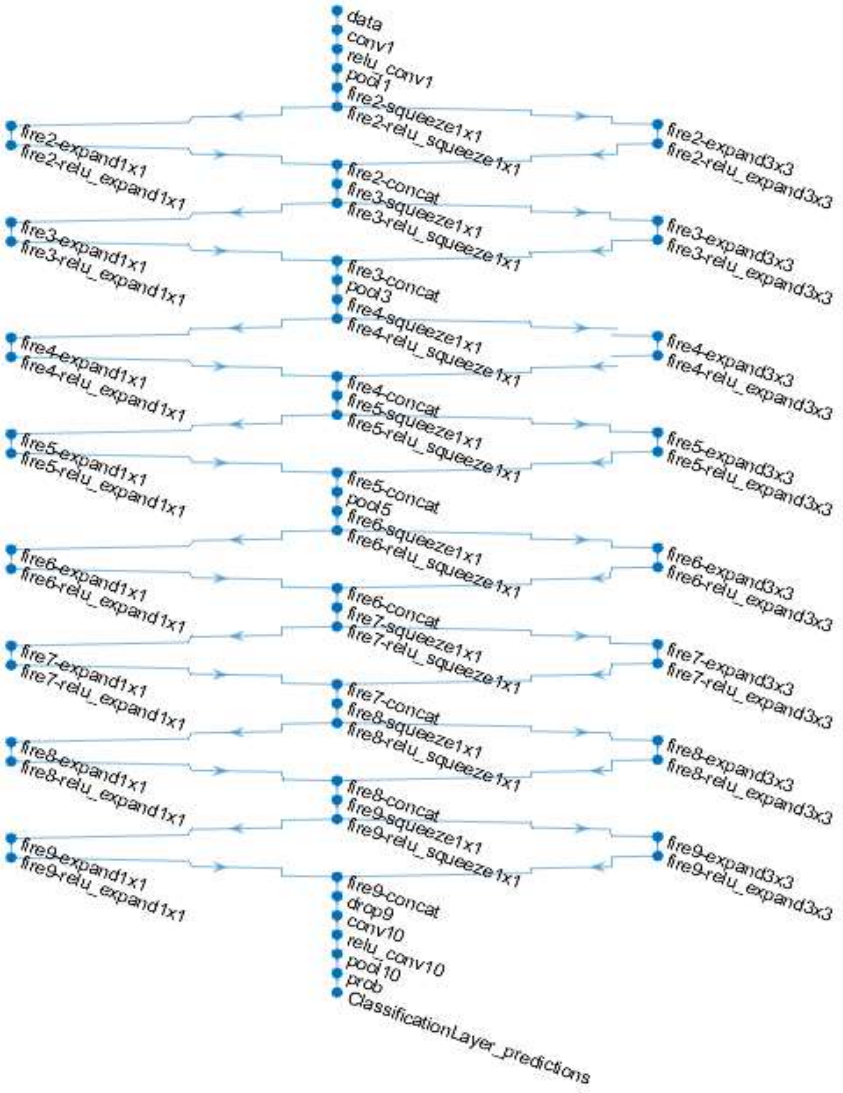

Figure 1. The Structure Of A Neural Network Of The Squeezenet Type

Thus, the purpose of this article is to assess the possibilities of using a neural network model such as SqueezeNet for recognizing the speaker's emotions.

3. DEVELOPMENT AND RESEARCH OF THE NEURAL NETWORK MODEL

According to the results of [4, 7, 24], the first stage in the development of a neural network model designed to analyze the parameters of technical systems is associated with the procedure for determining the input field of a neural network

model. It should be noted that a peculiar feature of CNN is the necessity to present the input information as a square image. The basic version uses a black and white image. More sophisticated options provide for using a multidimensional colour image.

This peculiar feature significantly limits the use of CNN – an ability to analyze a voice signal of a fixed duration. In this case, the general formulation of recognizing emotions of a user of an information system requires analyzing both a voice signal of a predetermined duration and a signal of unspecified duration. The first case can be correlated with monitoring the speaker's emotional state while voicing password data. The second case relates to monitoring voice messages of unspecified content during the interaction of the speaker with the information system.

According to the data of [21, 12, 13], such monitoring can be implemented, even using relatively simplified statistical models, by dividing an unspecified voice signal into fixed sections with a predetermined duration, followed by an analysis of each section.

Thus, the above limitation associated with the duration of the voice signal does not adversely affect the performance of CNN. In the first case, the duration of a voice message when scoring password data is a fixed value. In the second case, the voice message is divided into sections of a predetermined duration.

We are going to consider the procedure for encoding the parameters of a voice signal into a single-channel color square image under the assumption that the MFCC of each of the QFs of this signal are used as the indicated parameters. It is suggested correlating measurement of the abscissa axis with the MFCC number. We should note that the voice signal is usually characterized using 24 such coefficients. The measurement of the ordinate axis is proposed to be correlated with the QF number of the voice signal. Thus, a single QF will correspond to one separate point of the image. On the ordinate axis, the coordinate of the encoded fragment corresponds to the number of this fragment in the voice signal. The coordinate on the abscissa axis corresponds to the MFCC number. The colour of a dot will correspond to the MFCC value.

In the case when the number of QF of the voice signal is greater than the number of MFCC coefficients, to preserve the square shape, the figure below the abscissa axis is supplemented with lines filled with zeros. If the quantity of QF is less than the quantity of MFCC, to save the form, the figure on the right will be supplemented with columns filled with zeros.

Figures 2-4 show the above coding procedure. Figure 2 shows a sonogram of the scored text 'Say the Word Back', with Figure 3 containing diagrams of the values of the third, fourth and fifth MFCC in each of the QF. We should note that the duration of the spoken phrase is 1.452s, which corresponds to 121 QF,provided the sampling frequency of the voice signal is 24,414 Hz, with the signal duration of 0.012s. Voice signal, the sonogram of which is shown in Fig. 3 pronounced with the "angry" emotion. The sonogram shown in fig. 4 corresponds to the "neutral" emotion.

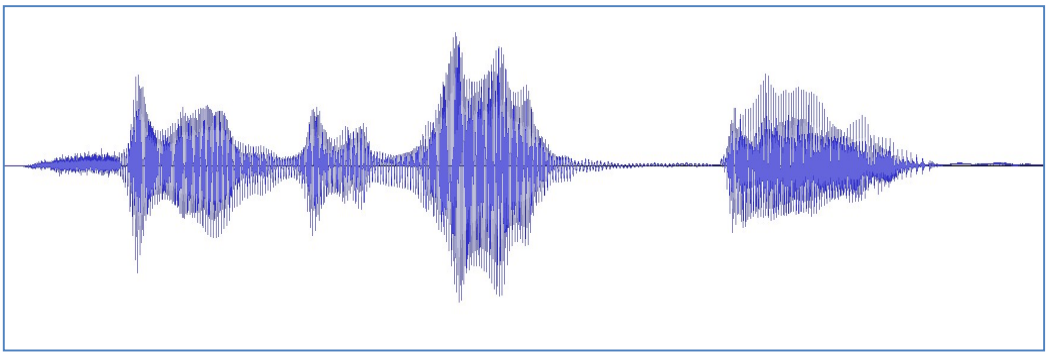

Figure 2. A Sonogram Of The Scored Text 'Say The Word Back'

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

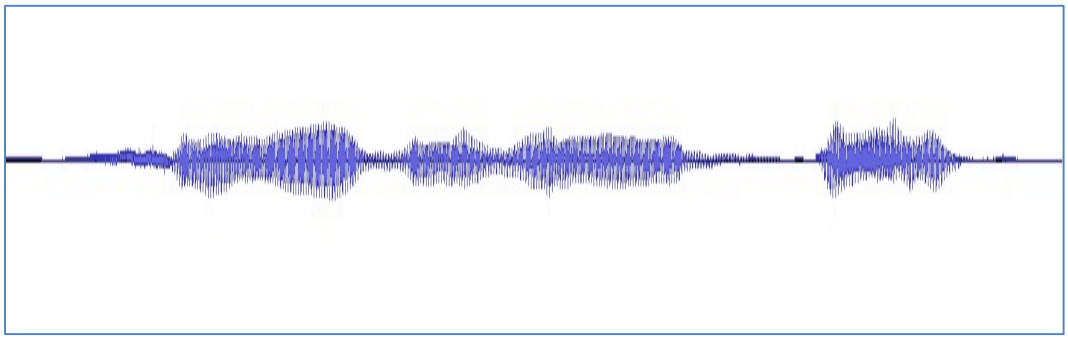

Figure 3. A Sonogram Of The Scored Text 'Say The Word Back'

Figures 4 and 5 show graphs of MFCC versus QF for each of two voice signals. As follows from the analysis of Fig. 4, 5, the dependences of MFCC on QF for the considered voice signals clearly differ from each other, which confirms the possibility of using such dependencies for the analysis of the voice signal.

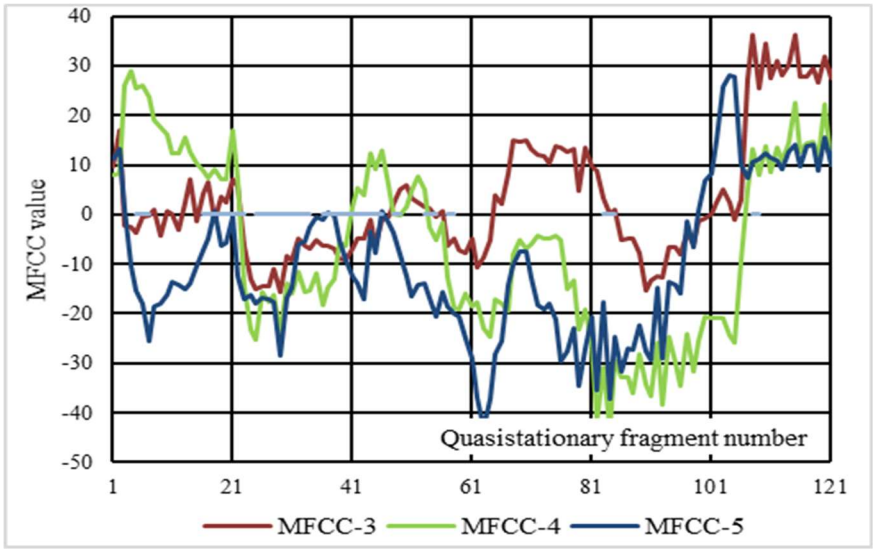

Fig. 4. Graphs Of MFCC Values when Scoring The Text "Say The Word Back"

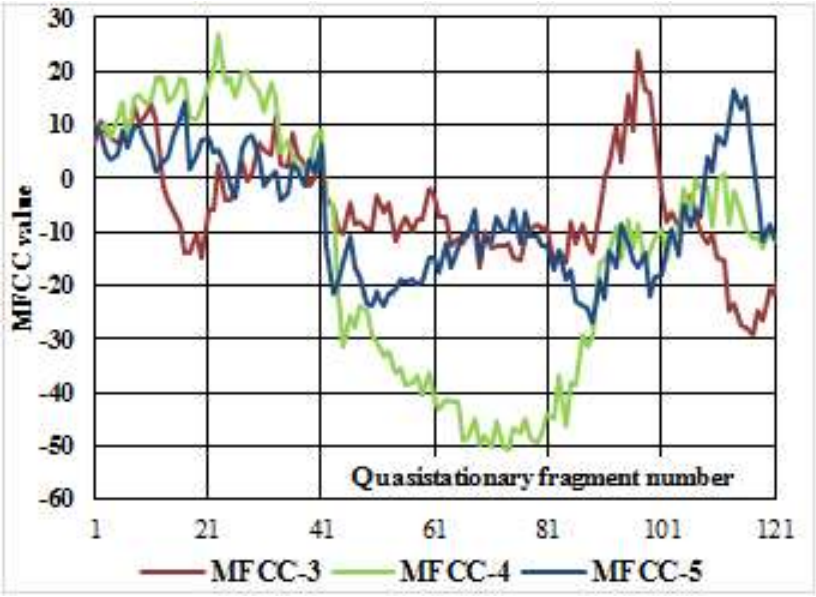

Fig. 5. Graphs Of MFCC Values when Scoring The Text "Say The Word Back"

Figure 6 fragmentarily shows a two-dimensional single-channel image of an encoded voice signal when scoring the text "Say the word back". The image in fig. 7 corresponds to the encoded voice signal for the text "Say the word bite".

		1	2	3	4	5
		MFCC-1	MFCC-2	MFCC-3	MFCC-4	MFCC-5
1	QF-1	-13,95	-18,62	9,62	8,09	11,08
2	QF-2	-14,13	-16,11	17,04	8,18	13,24
3	QF-3	-11,45	-22,95	-2,28	26,07	1,34
4	QF-4	-9,04	-27,60	-2,66	28,85	-10,41
5	QF-5	-7,43	-34,40	-3,78	25,34	-15,41

Fig. 6. Fragmented Display Of The Encoded Voice Signal When Scoring The Text 'Say The Word Back'

		1	2	3	4	5
		MFCC-1	MFCC-2	MFCC-3	MFCC-4	MFCC-5
1	QF-1	15,41	15,39	6,31	10,24	10,12
2	QF-2	14,57	18,32	10,59	8,98	8,22
3	QF-3	13,24	15,83	9,05	10,13	5,06
4	QF-4	11,48	16,30	7,65	7,41	3,65
5	QF-5	13,64	19,28	6,77	11,14	4,43

Fig. 7. Fragmented Display Of The Encoded Voice Signal When Scoring The Text 'Say The Word Back'

In fig. 6, 7 auxiliary fragments are highlighted in gray. The horizontal fragments display the MFCC numbers and the pixel numbers of the CNN input field along the abscissa axis. Vertical auxiliary image fragments display the QF and pixel numbers of the CNN input field along the ordinate axis. For visual clarity, the individual points of the input field are separated by straight lines. For example, an input field point with coordinates x = 2, y = 3 and the value of -22.95 correspond to QF-3 with the MFCC value of -22.95.

The above encoding procedure enables to proceed to the development of a neural network model of the SqueezeNet type for analyzing a voice signal in order to recognize the speaker's emotions. Based on the results of [9, 27], the TESS database is used as a data source for training and testing the neural network, available for download at https://doi.org/10.5683/SP2/E8H2MF. The above database contains 2,800 records of voice signals in English, recorded by two female speakers aged 26 and 64 years. Each voice signal reflects one of 7 emotions (anger, disgust, fear, happiness, surprise, sadness and neutrality). The two-part text was scored. The first part of the text 'Say the Word' is the same for all texts.

The second part of the text is variable and is one of 200 definite words (back, bar, cause, ...).

The duration of each spoken phrase is about 1.5-2s. Each voice signal recording is presented in a separate wav-file (sampling frequency of 24,414 Hz, mono channel). The name of the file reflects the name of the speaker, the second part of the text and the name of the emotion.

Since the duration of the voice signal of wav-files in the TESS database varies from 1.5 to 2s, taking into account possibility of changing the QF duration from 0.01s to 0.02s, the feasibility of half-overlapping QF, peculiar features of the implementation of the fast Fourier transform, specific features of SqueezeNet, it is accepted that the dimensions of the input field of the neural network model is 227x227 pixels. Since the number of MFCCs is 24, starting with 25, the columns of the input field of the network are filled with zeros. We should note that in the initial version, the SqueezeNet network is configured to classify images into 1,000 categories, with the task of analyzing the voice signal, which provides for recognition of 7 emotions, the final 5 layers of the network are adjusted to this condition.

The neural network model is implemented using the Python programming language, TensorFlow and Keras libraries. Also, when creating the program, the following libraries are used: Numpy – for processing arrays, Soundfile – for processing audio files and Matplotlib – for visualizing the results.

The neural network model and the corresponding software enable to move on to the next stage of research related to determining the effectiveness of SqueezeNet in recognizing emotions of a speaker.

In accordance with the recommendations of [2, 25], such parameters are used as recognition accuracy (A) and loss (L) in the training, validation and test samples, to evaluate the effectiveness [24, 22, 27-28].

To calculate the values of these indicators, the following formulae are used:

$$A = N_{right} / N \ , \qquad (1)$$

$$L = N^{-1} \sum_{t=1}^{N} e^{T}(t,Q) W(\theta) e(t,Q), \qquad (2)$$

where N_{right} is the number of correctly recognized examples, N is the total number of examples, $e(t, Q)$ is n_y-by-1 error vector at a given time t, parameterized by the parameter vector Q, n_y is the number of outputs of the neural network, $W(Q)$ is the weighting matrix, specified as a positive semidefinite matrix.

The experiments are carried out on a personal computer with an Intel Core i7-8700 processor (3.2-4.6 GHz), 16 GB of RAM, an nVidia GeForce GTX 1660Ti graphics card, run by Microsoft Windows 10. Figures 8 and 9 show the diagrams of recognition parameter accuracy and losses versus a number of training periods.

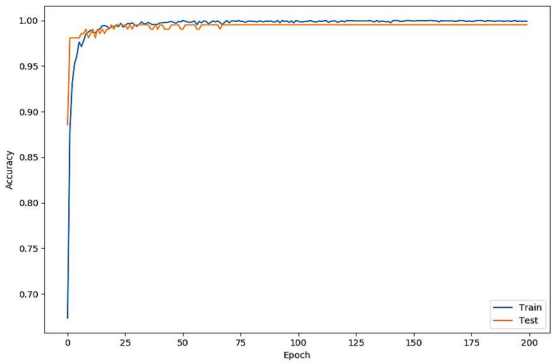

Figure 8. The Diagram Of Recognition Accuracy Versus The Number Of Training Periods

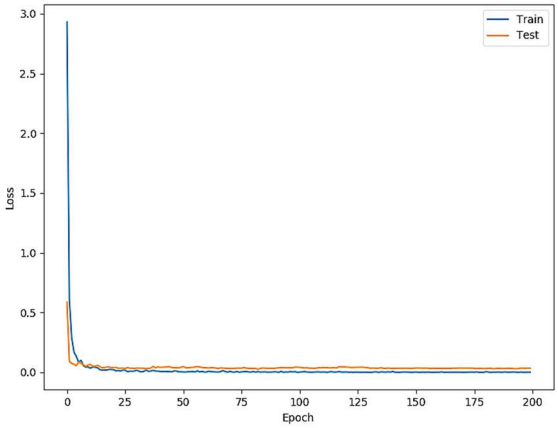

Figure 9. The Diagram Of Recognition Losses Versus The Number Of Training Periods

As the analysis of the above diagrams shows, both recognition accuracy and losses stabilize after about the 80th period of training. In this case, the recognition accuracy in test cases is about 0.98. We should note that the recognition accuracy on the validation sample not used for training is about 0.95, which is approximately 0.1-0.15 higher than in the best neural network recognition systems of similar purposes [4, 9, 10, 22].

Thus, the results of the experiments indicate the appropriate use of a SqueezeNet neural network

model for analyzing a voice signal in order to recognize the speaker's emotions. It should be noted that the voice signals presented in the TESS database were recorded in the studio conditions. Moreover, when applied, it is of interest to recognize the speaker's emotions under the influence of various noises.

To assess the influence of noise on recognition accuracy, experiments are conducted which, using the above SqueezeNet network, recognize high-noise training examples. We have determined that when exposed to white noise, the level of which is about 20dB, the recognition accuracy of emotions significantly decreases and amounts to approximately 0.6. Thus, it is advisable to correlate the trends of further research with the adjustment of neural network solutions to the recognition of the speaker's emotions under a variety of noise interference.

To assess the effect of noise on recognition accuracy, experiments were carried out in which, using the SqueezeNet network described above, recognition of noisy training examples was carried out. For an example in fig. 10 and in fig. 11 shows graphs of the dependence of accuracy and loss recognition for training and test cases depending on the number of training epochs when a noise of 10 dB is exposed to test cases.

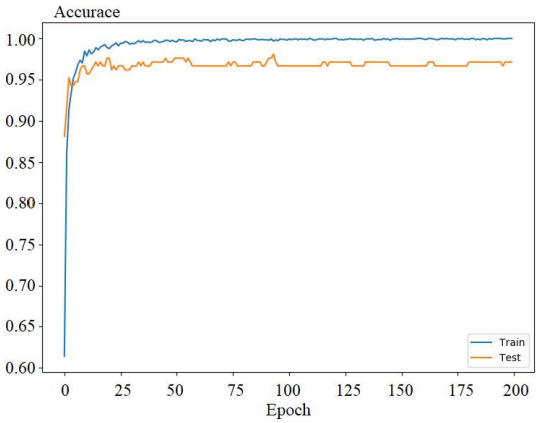

Fig. 10 Graphs Of The Dependence Of Recognition Accuracy On The Number Of Training Eras When Exposed To Noise

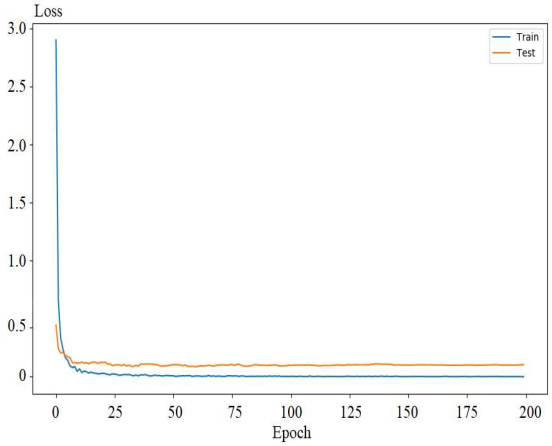

Fig. 11 Graphs Of The Dependence Of Losses On The Number Of Training Eras When Exposed To Noise

As a result of the experiments, it was determined that under the influence of white noise, the level of which was about 20 dB, the recognition accuracy of emotions significantly decreased and amounted to approximately 0.6.

Thus, it is advisable to correlate the paths of further research with the adaptation of neural network solutions to the recognition of speaker emotions in the presence of a variety of noise interference.

Also, based on the generally accepted methodology for increasing the efficiency of neural network diagnostic tools for technical systems [18, 23, 21], it is possible to suggest the prospect of research in the direction of creating a representative training sample for creating speaker-independent emotion recognition systems.

In addition, given the close relationship between the speaker's emotion recognition task and the user's voice recognition task, it is advisable to develop a method for adapting SqueezeNet architectural parameters to the conditions of the voice signal analysis task to simultaneously recognize the speaker's personality and emotions.

CONCLUSION

The results of the research show the possibility to effectively use a SqueezeNet-type neural network model for analyzing a voice signal in order to recognize the speaker's emotions. The efficiency of use is confirmed by computer experiments, which show the possibility of achieving the recognition of user's emotions at the level of 0.95 on a fairly limited training sample at 80

educational periods, which is comparable with the results of the best modern systems of similar purposes.

We have shown the necessity for further studies related to the adjustment of neural network solutions to the recognition of the speaker's emotions under a variety of noise interference.

The research has also determined the feasibility of developing a method for adjusting SqueezeNet architectural parameters to the conditions of the task of analyzing a voice signal for simultaneous recognition of the speaker's personality and emotions.

ACKNOWLEDGEMENTS

We are grateful to experts at Synk for their appropriate and constructive suggestions and sponsoring of this article.

This research was supported by grant of the program of Ministry of Digital Development, Innovations and Aerospace industry of the Republic of Kazakhstan IRN AP06851248 "Development of models, algorithms for semantic analysis to identify extremist content in web resources and creation the tool for cyber forensics".

REFERENCES

[1] Juslin P. N., Laukaa P. Communication of emotions in vocal expression and music performance: Different channels, same code? // Psychological bulletin. – 2003. – V. 129. – I. 5. – P. 770.

[2] Ranganathan H., Chakraborty S., Panchanathan S. Multimodal emotion recognition using deep learning architectures // 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). – IEEE, 2016. – C. 1-9.

[3] Ajinkya N. Jadhav, Nagaraj V. Dharwadkar, " A Speaker Recognition System Using Gaussian Mixture Model, EM Algorithm and K-Means Clustering", International Journal of Modern Education and Computer Science, Vol.10, No.11, pp. 19-28, 2018.

[4] Ingale A. B., Chaudhari D. S. Speech emotion recognition //International Journal of Soft Computing and Engineering (IJSCE). – 2012. – T. 2. – №. 1. – C. 235-238

[5] Altincay H. (2003). Speaker identification by combining multiple classifiers using Dempster–Shafer theory of evidence. Speech Communication, v.41, N4, 531–547.

[6] Vaziri G., Almasganj F., Behroozmand R. Pathological assessment of patients' speech signals using nonlinear dynamical analysis // Comput. Biol. Med. – 2010. – V. 40, № 1. – P. 54-63

[7] Ganchev T., Fakotakis N., Kokkinakis G. Comparative evaluation of various MFCC implementations on the speaker verification task // 10th International Conference on Speech and Computer. — Patras, Greece, 2005.

[8] Ing-Jr Ding, Chih-Ta Yen, Yen-Ming Hsu. Developments of Machine Learning Schemes for Dynamic Time-Wrapping-Based Speech Recognition // Mathematical Problems in Engineering. 2013, pp. 56-68

[9] Lyon R.F. Machine hearing: An emerging field // IEEE signal processing magazine. – 2010. – T. 27. – №. 5. – C. 131-139.

[10] Makarova V., Petrushin V.A. RUSLANA: a database of russian emotional utterances // ICSLP, 2002. pp. 20412044.

[11] Penagarikano, M., Bordel, G. (2004) Layered Markov models: A New architectural approach to automatic speech recognition. Machine Learning for Signal Processing XIV - Proceedings of the 2004 IEEE Signal Processing Society Workshop pp. 305-314

[12] Savchenko L.V., Savchenko A.V. Fuzzy Phonetic Encoding of Speech Signals in Voice Processing Systems. Journal of Communications Technology and Electronics. 2019. Vol. 64. No. 3. P. 238-244

[13] Zhang W.-Q., Deng Y., He L., Liu J.(2010). Variant Time-Frequency Cepstral Features for Speaker Recognition. Interspeech, pp. 2122-2125.

[14] Tereikovskyi I., Tereikovska L., Korystin O., Mussiraliyeva S., Sambetbayeva A. (2020) User Keystroke Authentication and Recognition of Emotions Based on Convolutional Neural Network. In: Hu Z., Petoukhov S., He M. (eds) Advances in Artificial Systems for Medicine and Education III. AIMEE 2019. Advances in Intelligent Systems and Computing, vol 1126, pp 283-292. Springer, Cham.

[15] Toliupa S., Tereikovskiy I., Dychka I., Tereikovska L., Trush A. (2019). The Method of Using Production Rules in Neural Network Recognition of Emotions by Facial Geometry. 3rd International Conference on Advanced Information and Communications Technologies (AICT). 2019, 2-6 July 2019, Lviv, Ukraine, Page(s): 323 – 327. DOI: 10.1109/AIACT.2019.8847847.

[16] Campbell W., Sturim D., Reynolds D. (2006). Support vector machines using GMM supervectors for speaker verification. IEEE

Signal Process. Lett., v.13, N5, pp. 308–311. doi: 10.1109/LSP.2006.870086.

[17] Karam Z., Campbell W. (2007). A new kernel for SVM MLLR based speaker recognition. In: Proc. Interspeech 2007, Antwerp, Belgium, August 2007, 290–293

[18] Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez Moreno, and Javier Gonzalez-Dominguez. Deep neural networks for small footprint text-dependent speaker verification. In Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, pages 4052–4056. IEEE, 2014

[19] Savchenko V.V., Savchenko A.V., 2016. Information Theoretic Analysis of Efficiency of the Phonetic Encoding–Decoding Method in Automatic Speech Recognition. Journal of Communications Technology and Electronics. 4(61): 430-435.

[20] Mitchell McLaren, Yun Lei, Nicolas Scheffer, and Luciana Ferrer. Application of convolutional neural networks to speaker recognition in noisy conditions. 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014, pages 686–690. ISCA, 2014

[21] Savchenko V.V., 2016. Enhancement of the Noise Immunity of a Voice-Activated Robotics Control System Based on Phonetic Word Decoding Method. Journal of Communications Technology and Electronics. 12(61): 1374-1379.

[22] Partila P., Voznak M., Tovarek J. Pattern Recognition Methods and Features Selection for Speech Emotion Recognition System // ScientificWorldJournal. – 2015. – V. 2015. – 7 p.

[23] Dychka I., Chernyshev D., Tereikovskyi I., Tereikovska L., Pogorelov V. (2020) Malware Detection Using Artificial Neural Networks. In: Hu Z., Petoukhov S., Dychka I., He M. (eds) Advances in Computer Science for Engineering and Education II. ICCSEEA 2019. Advances in Intelligent Systems and Computing, vol 938. Springer, Cham https://doi.org/10.1007/978-3-030-16621-2_1

[24] Hu Z., Tereykovskiy I., Zorin Y., Tereykovska L., Zhibek A. (2019) Optimization of Convolutional Neural Network Structure for Biometric Authentication by Face Geometry. In: Hu Z., Petoukhov S., Dychka I., He M. (eds) Advances in Computer Science for Engineering and Education. ICCSEEA 2018. Advances in Intelligent Systems and Computing, vol 754. Springer, Cham https://doi.org/10.1007/978-3-319-91008-6_57.

[25] Tereikovskyi I. A., Chernyshev D. O., Tereikovska L.A., Mussiraliyeva Sh. Zh., Akhmed G. Zh. The Procedure For The Determination Of Structural Parameters Of A Convolutional Neural Network To Fingerprint Recognition. Journal of Theoretical and Applied Information Technology. 30th April 2019. Vol.97. No 8. Pages 2381-2392.

[26] Iandola F.N., Han S., W. Moskewicz M.W. SqueezeNet: AlexNetlevel accuracy with 50x fewer parameters and <0.5MB model size. arXiv:1602.07360v4 [cs.CV], 2016. 13 p. Available at: https://arxiv.org/pdf/1602.07360.pdf.

[27] Satyanand Singh, Abhay Kumar, David Raju Kolluri,"Efficient Modelling Technique based Speaker Recognition under Limited Speech Data", International Journal of Image, Graphics and Signal Processing, Vol.8, No.11, pp.41-48

[28] Tereikovskyi, I., Tereykovska, L., Mussiraliyeva, S., Tsiutsiura, M., Achkoski, J. Markov model of unsteady profile of normal behavior of network objects of computer systems. CEUR Workshop Proceedings, 2019