Using convolutional neural networks to classify audio signal in noisy sound scenes*

M.V. Gubin
South Ural State University (national research university)
Chelyabinsk, Russia

***Abstract*—The issue of source separation of audio signals from a mixture of sounds is an important problem that arises in various industrial applications. For example, it can be mentioned that the issue of sound-based fault detection and diagnosis in industrial equipment or multi-speaker recognition task in a cocktail party problem. The last one is of great importance for the creation of new generation hearing aids that isolate and amplify a certain speech signal in noisy scenes. In this paper, I propose a new approach to solving this problem based on an ensemble of artificial neural networks. The solution of the problem is divided into two stages. At the first stage, the ensemble of convolutional neural networks determines the presence or absence of the speech signal in a noisy environment, using a set of speech signal samples prepared in advance. At the second stage, another ensemble of neural networks filters the speech signal determined at the first stage and cuts the rest of the signals as noise. The ensemble of convolutional neural networks, used at the first stage, consists of neural networks, each of which includes three convolutional layers and one fully connected layer. The analysis of the sound scene is performed on the basis of its spectrogram obtained by using the fast Fourier transform. This neural network is implemented in Python by the use of the TensorFlow and Keras software libraries. Here are the results of computational experiments on using designed and trained neural network for analyzing and filtering an audio stream that contains several superimposed male and female voices with music in the background. The performed experiments confirm the efficiency of the proposed approach.**

Keywords—neural networks; "cocktail party" effect; convolutional neural networks.

I. INTRODUCTION

Natural auditory environments, whether they are cocktail parties or rain forests, contain many things that make sounds concurrently. The cocktail party problem is the task of hearing a sound of interest, often a speech signal, in this sort of complex auditory settings. The problem is intrinsically quite difficult and there has been a longstanding interest for the way humans manage to solve it [1]. People manage to cope with the problem very effectively.

The cocktail party problem was described in the 1950s by Colin Cherry in his article "Some experiments on the recognition of speech, with one and with two ears " [2], which describes the phenomenon of the listener being able to filter out unwanted sounds in a noisy environment. The example of such a situation is the ability of the human sensory system to focus on a particular speaker in a noisy environment, hence the name of the problem.

Colin Cherry proposed a method for studying auditory attention presenting the now well-known dichotic listening paradigm to solve the following problems: Can an observer separate one speech signal from the others? Has the person kept anything about insignificant signals? How can one's attention be switched between signals? After more than three decades, Albert Bregman began to study sound segregation, calling it auditory scene analysis (ASA) [3, 4]. ASA is a complex process, the ear has access to the only one pressure wave, and it is the sum of the pressure waves emanating from individual sound sources [5]. During the analysis, it is necessary to filter out various sound waves. This has been the subject of recent researches using such methods as an independent component analysis (ICA), radiation pattern measurements, a computational auditory scene analysis (CASA), and triangulation methods between microphones. However, the implementation of these methods is still limited in a complex acoustic environment. In complex acoustic conditions, when various noisy environments mask the target signal and its extraction is partially impossible, we may use an additional analysis based on the selection of the semantic categories of the sound wave.

The article proposes a new approach to solving this problem on the basis of an ensemble of artificial neural networks. The solution of the problem is divided into two stages: determining the presence of a useful voice in a noisy environment and filtering this voice, cutting out the rest as noise.

II. REVIEW OF WORK

The problem of the source separation of audio signals from a mixture is an important task that arises in various areas of industry. For instance, I can mention the problem of isolating and classifying the noise associated with the occurrence of malfunctions in the process equipment or the problem of speech recognition for managing complex technical objects [6]. Another important task is the separation of speech and noise signals for people with a hearing loss [7]. Hard of hearing people lose the ability to analyze auditory scenes, it is sometimes difficult for them to follow a conversation between several interlocutors or it is difficult to orientate themselves in the sound space if there are several audio events and they are simultaneous. In such situations, hearing aids simply amplify these sound events, creating an incoherent noise. Using

*The study was supported by the Government of the Russian Federation according to Act 211 (contract No. 02.A03.21.0011) and by the Ministry of Science and Higher Education of the Russian Federation (government order 2.7905.2017/8.9).

intelligent hearing aids can solve this problem. The implementation of such hardware and software devices is complicated due to the presentation of overstated requirements to them: processing of audio streams must take place in real time mode; devices must have low power consumption, modest size and weight characteristics. It is expected that the global industry of hearing aids worth $ 6 billion will grow by 6% each year until 2020 [8]. There are many other tasks, that require a qualitative separation of audio sources, and most of these tasks require modest computational resources and low energy consumption when processed in real time [9].

There are several approaches to solving the problem of source separation of mixture, one of them is based on the use of artificial neural networks, and it is considered the most successful in recent researches [8]. There are many architectures of neural networks used in the audio processing, the main ones include full-connected networks, which are a basis for various auto-encoders, convolutional networks, which distinguish attributes from audio streams, and recurrent networks, which are well adapted for the processing of different time sequences. The latter have shown their effectiveness in isolating a useful signal against a noise background [6].

To separate audio sources by machine learning, audio streams are mapped to a two-dimensional image of a spectrogram with two axes: time and frequency [10], which is subsequently analyzed to extract the expected features. The analysis of spectrograms is well suited to convolutional networks, and the result of such analysis is used by other components of audio processing systems. The auditory scene analysis [11, 12] consists in the fact that the audio system splits audio mixtures into fragments (for example, the areas of the time-frequency plane). Fragments belonging to the same audio source will be assigned to the same cluster. Since there are several sources, their clusters can alternate or intersect, and their intersections can be either in time or in frequency. Clustering of such fragments and extraction of audio sources from them is based on the technology of time-frequency masks (T-F) [13]. This mask can be obtained for various signals based on the analysis of time, spectral or spatial characteristics of the sources [14].

More complex schemes for analyzing audio scenes in addition to the technology of time-frequency masks can use information about the mixing vector (MV), interaural level and phase differences (ILD and IPD) [14]. It is shown that the MV distributions are quite distinct while binaural models overlap when the sources are close to each other. There is a stereo speech recognition system on the background of noise, based on a deep neural network to create a soft T-F separation mask [15]. The neural network, which is composed of two sparse auto-encoders and a softmax regression, is used to estimate the orientations of the dominant source at each T-F unit, based on low-level features, such as mixing vector (MV), interaural level and phase difference (IPD/ILD). The method of separating stereo speech signals [15], using time-frequency masks T-F, is one of the effective ones. However, the reliable evaluation of the T-F mask from audio mixtures is a difficult task, especially when room reverberations are present in the mixtures.

The approaches to the processing of audio streams can be divided by the number of sources (microphones) into monoural [13, 16, 17, 18, 19], binaural [11, 15, 20, 21] and multi-microphone. An overview of multichannel methods is also presented. It includes beam masking, independent component analysis (ICA) and time-frequency masks (T-F) [22].

The problem of binaural segregation as a binary classification with the use of deep neural networks (DNN) for the classification problem is considered in [23]. As the main features for the classification are used binaural: intertemporal differences of waves from different pickups, and also the difference of this wave in amplitude. The work [24] proposed an algorithm for tracking the azimuths of all active sources at a certain time interval, using a hidden Markov model for the formation of continuous tracks and automatic determination of the number of active sources in time.

In work [25], psychoacoustic models have been developed that can accurately predict the effects of auditory scene analysis. Grouping, segregation and audio streaming represent the subsequent stages of processing that closely interact with attention. Audio segments can be sequentially selected and grouped with the help of primitive functions such as spatial placement and base frequency. More complex processing is required when it is necessary to use lexical, syntactic or semantic information.

In a number of recent works, quite simple architectures of neural networks have appeared, in contrast to deep architectures, which allow the reliable recognition of various audio events. In work [26], the neural network architecture consists of only three layers: convolutional, pool and softmax layer. However, the architecture is not critical to the length of the input audio fragment.

In this article, a new approach to solving the problem of sources separating of audio signals from a mixture of sounds is proposed on the basis of an ensemble of artificial neural networks. The solution of the problem is divided into two stages. At the first stage, an ensemble of convolutional neural networks determines the presence or absence of the speech signal in a noisy scene using a set of speech signal samples prepared in advance. At the second stage, another ensemble of neural networks filters the speech signal determined at the first stage and cuts the rest of signals as a noise. At the output, there are separate audio signals of voices, the ensembles of neural networks were trained to isolate.

III. Architecture of an Artificial Neural Network

My approach to solving the problem of isolating a useful voice source in a noisy environment is based on the use of ensembles of artificial neural networks. The solution of the problem is divided into two stages. At the first stage, an ensemble of convolutional neural networks determines the presence or absence of the speech signal in a noisy scene using a set of speech signals samples prepared in advance. At the second stage, another ensemble of neural networks filters the speech signal determined at the first stage and cuts the rest of signals as a noise. To separate each voice from the mixture, it is used an ensemble of neural networks consisting of one convolutional and one recurrent (Fig. 1).

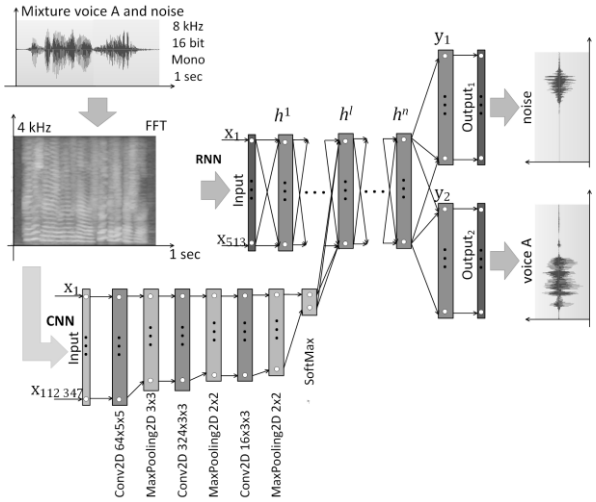

Fig. 1. Architecture of neural network ensemble: one RNN and one CNN

This approach can be generalized to extract a certain number of target votes from the audio stream by constructing a certain number of such ensembles consisting of one convolutional network and one recurrent network. The reasonable number of such ensembles is limited by the ratio of the useful signal to noise.

The analysis of the mixture of the audio signals occurs according to the spectrogram obtained by the fast Fourier transform method (Fig. 1). A convolutional neural network includes three convolutional layers and one fully connected layer. This network must reliably determine the target voice in the audio stream. The received data from the convolutional network are transmitted to the recurrent neural network, which filters the sound stream from noise in the presence of the target voice, otherwise cuts out the entire stream as noise. Data on the presence of the target voice are transmitted to the inner layer of the convolutional network.

The architecture of the recurrent neural network will be described and studied in the next paper. The method of the construction of learning and test samples is proposed below, the results of experiments are also described.

IV. Construction of Training and Test Samples

The method of the construction of training and test samples is based on the choice of voices the neural network is being tuned to define, and the separation of these samples into learning fragments and testing fragments, these fragments are disjoint. Samples of voices are collected from audiobooks, access to which is free through Internet resources. The fragments of audiobooks were selected for five speakers: two male and three female voices lasting more than 3 hours each. One voice is the basic (the female voice), on the diagrams it is named voice A, the remaining voices are considered as noise, these are the voices that hinder the network from singling out the presence of voice A, they will be called voice M1, voice M2, voice W1 and voice W2. The use of audiobooks is very convenient, the voice streams in them are cleaned from various noises, the duration of audiobooks reaches tens of hours, the speaker reads expressively and the speech is clearly discernible.

The construction of training and test samples is carried out according to the following principle: fragments with the duration of one second are selected sequentially from each audiobook, which is an audio file lasting more than 3 hours with a sampling rate of 8 KHz for one channel and a depth of 16 bits. Then a spectrogram based on the Fast Fourier Transform algorithm is built for each fragment. The spectrogram is stored in the database for training and testing a convolutional neural network. In addition to training the neural network on clean voices, clean fragments from audiobooks are used, new fragments of the mixture of these voices are built. They may contain voice A in the mixture or may not. A complete list of the mixtures in use and the separation of these mixtures in the training stage of the neural network is given in Table 1. Mixtures of voices are formed in the sound editor that levels all voices in amplitude, so all voices in the mixtures have a normalized representation.

To determine the architecture of a convolutional neural network, a preliminary stage of training is introduced. At this stage, the neural network is trained to distinguish voice A from other clean voices. This preliminary stage is discussed in detail in Section 5 below. At the same stage for voice A, the segments of the sound stream are marked. Words and phrases are selected if they last three quarters of one second or more. These are the words and phrases to carry out the experiments. This is necessary for the quality filling of the spectrogram with voice A on training samples.

To construct training samples, the first 2.5 hours of the audio track are used. They are divided into 40 segments. Each segment is divided into words or phrases lasting 1 second. The number of training samples within the segment for all mixtures is 4000. The number of training samples for the entire fragment is 160000. For the final testing of the accuracy of the neural network the remaining 30 minutes – 15000 records are used. Thus, the process of final estimation of the accuracy of the neural network takes place on the unfamiliar fragments of the neural network itself.

TABLE I. Training Stages and Mixtures in Use

Mixtures in use	Number of training fragments	Number of testing fragments
Stage 1		
Voices: A, M1, W2, M1, W2 **Mixtures:** A and M1; A and M2; A and W1; A and W2; M1 and W1; W1 and W2	160 000	3000
Stage 2		
Mixtures: M1, M2 and W1; M1,M2 and W2; M1, W1 and W2; M2,W1 and W2; A, M1 and M2; A, M1 and W1; A, M1 and W2; A, W1 and W2; A, M2 and W1; A, M2 and W2	160 000	3000
Stage 3		
Mixtures: Music, A and M1; Music, A and M2; Music, A and W1; Music, A and W2; Music, M1 and M2; Music, M1 and W1; Music, M1 and W2;Music, M2 and W1; Music, M2 and W2; Music, W1 and W2	160 000	3000
Stage 4 (final testing)		
	15000	

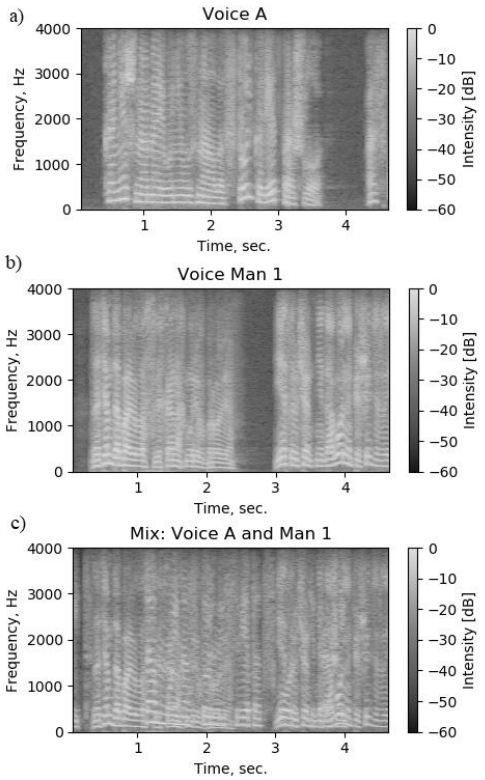

Fig. 2. Spectrograms: a) voice A; b) voice M1; c) voices A and M1

The analysis of sound fragments by a neural network takes place according to spectrograms, the examples of spectrograms are given in Fig. 2. Fig. 2a shows an example of a target voice (voice A) spectrogram with a duration of 4.5 seconds, Fig. 2b shows a spectrogram of the voice M1, and in Fig. 2c shows a spectrogram of a mixture of voices: voice A and voice M1.

From the analysis of the spectrograms (Fig. 2), it can be seen that the events or fragments belonging to the same acoustic source begin and end simultaneously and they can be attributed to the same source. Our task is to teach the network to find such fragments in the mixture that belong to voice A.

V. Description and Analysis of Experiments

In the preliminary stage of the experiments, it was necessary to determine the architecture of the convolutional neural network. That is why several architectures were selected with the number of convolution layers and the training accuracy as determining factors. We selected architectures with the number of convolutional layers from one to six. The training was conducted on small audio fragments lasting 4 minutes. The graphs of accuracy growth of these neural networks during the process of training are presented in Fig. 3, the number of learning parameters of these networks and the training accuracy are presented in Table 2. All these architectures were trained to determine voice A from the variety of other samples. The analysis of accuracy showed (Fig. 3 and Table 2) that next architectures are suitable for solving this problem: network 3, network 4, network 5 and network 6. For architecture of the network 1, the training sample may not have sufficient size. Further studies were carried out only for the architecture of neural network 3.

The first stage of neural network training when the network differs voice A from other voices and their mixtures including voice A or without it was conducted on fragments lasting 2.5 hours. Mixtures include no more than two voices, including voice A (Fig. 4). The ability of the network to determine voice A from all other voices after the full cycle of training is more than 98.7%. In the second stage, the neural network was trained on mixtures containing three voices, including voice A (Table 1). The accuracy of the neural network training in the second stage is 96.9%. In the third stage, the neural network was trained to determine the presence of voice A in mixtures with music in the background. The accuracy of training is 96.8%. The results of the accuracy of neural network training are shown in Fig. 4 and Fig. 5. To test the reliability of the neural network, the trained neural network was tested at the low level of the target signal (voice A). The level of the target audio signal was lowered by 3 dB, while the accuracy of the network did not change significantly.

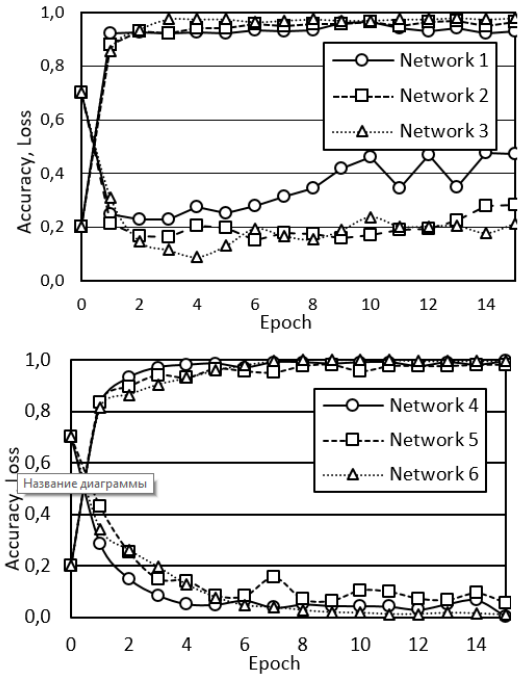

Fig. 3. Speed of neural network training, preliminary stage

TABLE II. Convolutional Neural Network Data

Network name	Number of convolutional layers	Number of training parameters	Network accuracy	Loss
Network 1	1	3,772,226	0.92	0.61874
Network 2	2	1,874,050	0.945	0.29653
Network 3	3	500,674	0.975	0.19481
Network 4	4	196,866	0.995	0.07098
Network 5	5	122,402	0.98	0.08976
Network 6	6	150,498	0.995	0.01039

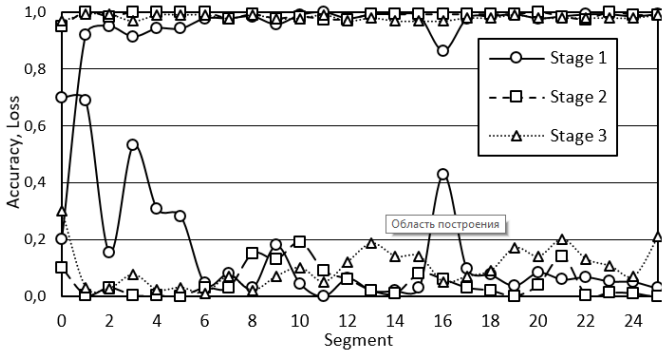

Fig. 4. Speed of neural network training, different stages

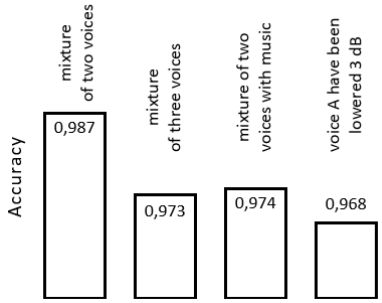

Fig. 5. Accuracy of neural network in the final testing stage

The calculations were made on a graphics accelerator NVidia Tesla K40m with 12GB of RAM and a peak performance of 4Tflops. The software is presented in the form of Python 3.6 modules using the TensorFlow and Keras libraries. This implementation is available through Internet resources https://github.com/gubinmv/cnn_in_noisy_scenes.git.

VI. Conclusion

The paper describes a new model based on an ensemble of neural network architectures that combine a convolutional neural network and a recurrent neural network. A network architecture has been designed for a convolutional neural network and experiments have been performed to confirm the reliable determination of the target voice in an audio stream containing several superimposed male and female voices with music in the background. The accuracy of neural networks shows the reliability of determining the voice in a mixture of audio signals, which is confirmed by the experiment with a lowered power of the target audio signal (voice A).

In the next work, the following tasks are planned to perform:

- to determine the degree of the influence of the target audio signal power on the accuracy of detection of its presence in the mixture;
- to determine the architecture of the RNN network and conduct experiments on the separation of audio signal sources from the mixture;
- to evaluate the reliability of the proposed method on the set of neural network ensembles to identify the set of target voices;
- to conduct experiments for audio signals with a higher sampling rate.

References

[1] J.H. McDermott, “The cocktail party problem”, Current Biology, vol. 19, № 22, pp. 1024–1027, 2009.

[2] E. C. Cherry “Some experiments on the recognition of speech, with one and with two ears”, The Journal of the Acoustical Society of America, vol. 25, № 5, pp. 975–979, 1953.

[3] A.S. Bregman, S. McAdams, “Auditory Scene Analysis: The Perceptual Organization of Sound”, The Journal of the Acoustical Society of America, vol. 95, № 2. pp. 1177–1178, 1994.

[4] W.L. Rogers, A.S. Bregman, “Cumulation of the tendency to segregate auditory streams: Resetting by changes in location and loudness”, Perception and Psychophysics, vol. 60, № 7, pp. 1216–1227, 1998.

[5] A. Bregman, P. Ahad, “Demonstrations of Auditory Scene Analysis: The Perceptual Organization of Sound. Audio compact disk”, The MIT Press, Cambridge, MA, 1996.

[6] Andrew Maas, Quoc Le, Tyler M. O’Neil, Oriol Vinyals, Patrick Nguyen, Andrew. Y.N. “Recurrent Neural Networks for Noise Reduction in Robust ASR”, 13th Annual Conference of the International Speech Communication Association, vol. 1, pp. 22–25, 2012.

[7] D. Monroe “Digital hearing”, Communications of the ACM, vol. 60, № 10, pp. 18–20, 2017.

[8] D.L. Wang, “Deep learning reinvents the hearing aid”, IEEE Spectrum, March 2017, pp. 32–37.

[9] K. Wiklund, S. Haykin “The cocktail party problem: Solutions and applications”, Canadian Acoustics, vol. 37, № 3, pp. 80–81, 2009.

[10] J. Dennis, H.D. Tran, H. Li, “Spectrogram Image Feature for Sound Event Classification in Mismatched Conditions”, IEEE Signal Processing Letters, vol. 18, № 2, pp. 130–133, 2011.

[11] D. Wang, “Time-Frequency Masking for Speech Separation and Its Potential for Hearing Aid Design”, Trends in Amplification, vol. 12, № 4, pp. 332–353, December 2008.

[12] D.S. Williamson, D. Wang, “Time-Frequency Masking in the Complex Domain for Speech Dereverberation and Denoising”, IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, № 7, pp. 1492–1501, July 2017.

[13] Po-Sen Huang, Minji Kim, Mark Hasegawa-Jonson, Paris Smaragdis, “Deep learning for monaural speech separation”, 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1562–1566, 2014.

[14] Atiyeh Alinaghi, Philip JB Jackson, Qingju Liu, and Wenwu Wang, “Joint Mixing Vector and Binaural Model Based Stereo Source Separation”, EEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, № 9, pp. 1434–1448, 2014.

[15] Y. Yu, W. Wang, P. Han, “Localization based stereo speech source separation using probabilistic time-frequency masking and deep neural networks”, EURASIP Journal on Audio, Speech, and Music Processing, № 7. p. 1-18, 2016.

[16] J. K. Perin, M. Frank, N. Gallagher, “Speaker Recognition for Multi-Source Single-Channel Recordings”, CS 229: Machine Learning Final Projects, pp. 1-5, Autumn 2014.

[17] Y. Luo, Zhud Chen, John R. Hershey, Jonathan Le Roux, Nima Mesgarani, “Deep clustering and conventional networks for music separation: Stronger together” // 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 61–65, November 2017

[18] Po-Sen Huang, Minje Kim, Mark Hasegawa-Jonson, Paris Smaragdis, “Singing-Voice Separation From Monaural Recordings Using Deep Recurrent Neural Networks”, 15th International Society for Music Information Retrieval Conference, pp. 1562–1566, 2014.

[19] Po-Sen Huang, Minji Kim, Mark Hasegawa-Jonson, Paris Smaragdis, “Joint Optimization of Masks and Deep Recurrent Neural Networks for Monaural Source Separation”, IEEE/ACM Transactions on Audio Speech and Language Processing, vol. 23, № 12, pp. 2136–2147, 2015.

[20] M.S. Pedersen, DeLiang Wang, Jan Larsen, Ulrik Kjems, “Two-Microphone Separation of Speech Mixtures”, // IEEE Transactions on Neural Networks, vol. 19, № 3, pp. 475–492, 2008.

[21] X. Zhang, D. Wang, “Deep Learning Based Binaural Speech Separation in Reverberant Environments”, IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, № 5, pp. 1075–1084, 2017.

[22] A. Hidri, S. Meddeb, H. Amiri, “About Multichannel Speech Signal Extraction and Separation Techniques”, Journal of Signal and Information Processing, vol. 3, pp. 238-247, 2012.

[23] Yi Jiang, D. Wang, R. Liu, Z. Feng, “Binaural Classification for Reverberant Speech Segregation Using Deep Neural Networks”, IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, № 12, pp. 2112–2121, 2014.

[24] N. Roman, DeLiang Wang, “Binaural Tracking of Multiple Moving Sources”, IEEE Transactions on Audio, Speech, and Language Processing, vol. 16, № 4, pp. 728–739, 2008.

[25] A.W. Bronkhorst, “The cocktail-party problem revisited: early processing and selection of multi-talker speech”, Attention, Perception, & Psychophysics, vol. 77, № 5, pp. 1465–1487, 2015.

[26] H. Phan, L. Hertel, M. Maass, A. Mertins, “Robust Audio Event Recognition with 1-Max Pooling Convolutional Neural Networks”, Proceedings of the Annual Conference of the International Speech Communication Association, pp. 3653–3657, September 2016.