Research of neural network classifier in speaker recognition module for automated system of critical use

Mykola M. Bykov[a], Viacheslav V. Kovtun[a], Andrzej Smolarz[b], Mukhtar Junisbekov[c], Aliya Targeusizova[d], Maksabek Satymbekov[e]
[a]Vinnytsia National Technical University, Khmelnytsky Hwy, 95, 21021 Vinnitsa, Ukraine; [b]Lublin University of Technology, Nadbystrzycka 38A, 20-618 Lublin, Poland; [c]M.Kh. Dulati Taraz State University, Tole Bi St 60, Taraz, Kazachstan; [d]al-Farabi Kazakh National University, Almaty, Kazakhstan; [e]Institute of Information and Computational Technologies, Almaty, Kazakhstan;

ABSTRACT

The article studies the dependence of the quality of speakers recognition by convolutional neural network from the type of chosen informative features for use it in automated systems for critical use especially when they are used in the environmental influences. The environmental influences are the noise of high level with a spectrum that correlates with the spectrum of the speech signal or the signal of speaker simulator. Convolutional network operation principles for the case of speaker signal recognition, as well as experiments on neural network training and the recognition of speakers on a test samples have been considered. According to the research, it was concluded that the bark-cepstral coefficients make it possible to perform recognition with greater reliability than the spectral parameters of the signal.

Keywords: Speech recognition, Convolutional network, cepstral analysis

1. INTRODUCTION

The task to provide an information security by increasing of demand on critical applications system, that can work under adverse conditions with guaranteed reliability, is very actual in secure information systems. In context of automated speaker recognition systems, this means the creating of a module that would perform speakers recognition procedure with a high probability of correct recognition in the critical systems in condition of high level noise, signal of the person who imitate speaker's speech, etc. As it was shown by previous authors studies the interference effect on the speaker recognition accuracy, using of the speech signal filtering are effective only in cases when the clusters of speakers in the feature space are not overlapped[1].

2. ACTUALITY OF THE RESEARCH

Existing automated speaker recognition systems are mainly used for identifying a large number of speakers with sufficiently high reliability in the lab conditions. At the same time, the speaker recognition module for automated systems for critical use have to provide recognition reliability close to 100% for a small number of speakers in any environment. Therefore it is necessary to use in the speaker recognition system a specific set of informative features and to take into account the additional requirements to the speaker classifier. Taking into account the random nature of the interference, it's impossible to predict which of the "traditional" informative features (energy of signal, linear prediction coefficients, frequency and period of the pitch and others [2-5]) will be more effective. Therefore, the actual problem is the study of the features effectiveness by using a classifier which itself would extract them from audio signal. Such task can perform deep neural network (DNN) and convolutional neural network (CNN). The authors in this work have chosen the convolutional neural network.

3. EXPERIMENT

The architecture of the convolutional neural network is shown in Figure 1. Convolutional neural network (CNN[6-8]) uses the convolution for an information processing operation[6-8], which in the context of image processing can be described by the following formula:

*mbykov123@ukr.net

Photonics Applications in Astronomy, Communications, Industry, and High Energy Physics Experiments 2017, edited by Ryszard S. Romaniuk, Maciej Linczuk, Proc. of SPIE Vol. 10445, 1044521

$$(inp * cor)[m,n] = \sum_{k,l} f_{act}[m-k, n-l] \cdot cor[k,l] \tag{1}$$

where inp denotes input matrix (for the first layer - digitized images), and cor – the convolution core.

At each step, the element-wise multiplication of the contents of all the windows on the core g, summing of the results of products and entering the amount into the results matrix are fulfilled. Depending on the method of cutting edges of the input matrix the result may be less than the original image (valid), the same size (same) or larger (full).

In the convolutional neural network the set of weights are formed, which encode all the characteristic features of the image (eg slope, width, color, lines and points). The convolution core is formed by neural network itself in the training process with classical Backpropagation method[5]. Passage of the image by each set of weights creates a unique instance of the map features, turning to a multi-dimensional neural network (set of unique feature maps in a single layer). For example, the matrix of the dimension 5 × 5 is shifted to one or two neurons (pixels) instead of five, what allows to maintain the desired feature. In a simplified form the convolution layer is described with the following formula:

$$x^l = f_{act}\left(x^{l-1} * k^l + b^l\right), \tag{2}$$

where x^l is the output of the layer l, $f_{act}(...)$ - activation function, b - shift coefficient, symbol $*$ denotes the operation of convolution of the input x with core k.

Under the given angle effects of the amount of input matrix change in such a manner:

$$x_j^l = f\left(\sum_i x_i^{l-1} * k_j^l + b_j^l\right), \tag{3}$$

where x_j^l is the map of features j (output layer l), $f_{act}(...)$- activation function, b_j^l - shift coefficient of the map of features j layer l, k_j^l - the convolution core number j, x_i^{l-1}- the features map of the previous layer.

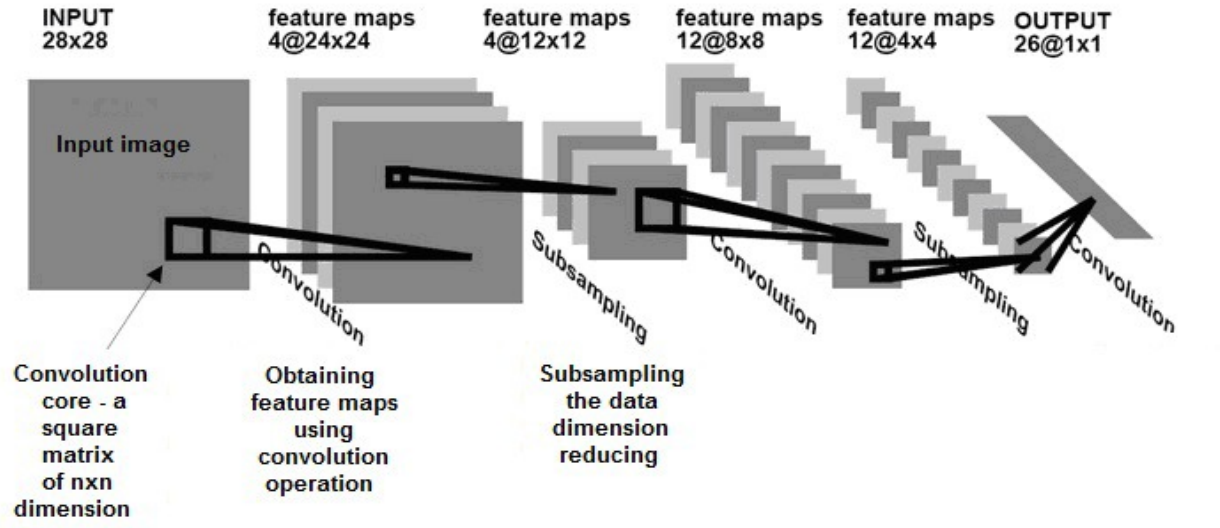

Figure. 1. Architecture of convolutional deep learning neural network

Sub-sampling operation is performed to reduce the dimension of the existing maps of features. It is believed that in this network architecture the information about the presence of informative feature is more important than data about its exact coordinates. Therefore, the neuron with the maximum excitation is selected among several neighboring neurons of features map and accepted as one neuron of a reduced dimension features map. Sometimes the operation of the average between neighboring neurons is used. These operations increase further the speed of calculations and make the neural network more

invariant to changes in the input image size. For example, if the network performs the sub-sampling operation by method of maximum element choice (max-pooling), the whole map is divided into cells of size $n \times n$ (usually $n = 2$), in each of which the element with maximum value is selected. Analytically this operation we describe as:

$$x^l = f\left(a^l \cdot sbs\left(x^{l-1}\right) + b^l\right) \tag{4}$$

where x^l is the output of the layer l, $f_{act}(...)$ - activation function, a, b - coefficients, $sbs(\ )$ – operation of the local maximum values finding.

Thus, by repeating, operations of convolution and sub-sampling turn into several layers, a convolutional neural network is building. Alternating of the layers allows to form the generalized feature maps from other feature maps that makes possible the recognition of the image using complex hierarchies features. Usually after passing of several layers the features map degenerates into a vector or even scalar, but the number of these cards can be equal to a few hundred. In the output of the network additional several layers of fully-connected neural network (perceptron) are often set, whose input is fed with final features maps outputs:

$$x_j^l = f\left(\sum_i x_i^{l-1} \cdot \omega_{ij}^{l-1} + b_j^{l-1}\right), \tag{5}$$

where x_j^l is the map of features j (output layer l), $f_{act}(...)$ – activation function, b – shift coefficient, ω – matrix of weights.

If on the first layer the convolution core passes only one input image, then on the inner layers the same core runs parallel on all feature maps of this layer, and the result of convolution is summed to form (after passing the activation function) one features map of the next layer, which corresponds to this convolution core.

Convolutional neural networks have a number of significant advantages. They are one of the best classifications and image recognition algorithms, significantly fewer number of adjustable weights (as compared to a fully-connected neural network with the same number of layers), the ability to parallelize calculation and the ability of learning with backpropagation method. Application of convolutional neural networks to the speaker recognition task is largely restricted by fact that such network is suitable only for image recognition. In addition, the convolutional neural network summarizes information on the basis of the finding of the informative feature presence in the input image without accounting of its exact coordinates. At the stage of the study authors have eliminated these limitations by selecting informative features in the form of the signal spectrum and bark-cepstral coefficients with their subsequent visualization.

4. RESEARCH OF THE INFLUENCE OF SELECTED FEATURES ON THE SPEAKER RECOGNITION RELIABILITY

The articulation apparatus of each person is unique, which causes the individual frequencies combination of speech signals which a person speaks. To study the impact of features types on the reliability of speaker recognition authors suggested to use the visualized spectrograms and frequency cepstral coefficients (FCC) of the speakers passphrase records. These features are widely used in automatic speech recognition systems[9,10].

Values of cepstral coefficients are obtained from Spectrum module (module of values of Fast Fourier transformation of input signal) using a bank of filters which are evenly distributed on the "distorted" by applicable law frequency scale. Then the resulting spectrum is weighted by filter banks. The resulting set of values is subjected to logarithm, and then decorrelated using a discrete cosine transformation to obtain cepstral coefficients vector. Depending on the used frequency scale for the distortion the Mel-cepstral (MFCC), Bark-cepstral (BFCC) and Uniform-frequency Cepstral Coefficients (UFCC) are distinguished. Windows filters, which perform the convolution of input signal spectrum in MFCC, BFCC, UFCC scales differ among themselves. So Bark filters have a narrow bandwidth at low frequencies and significantly broader at high, while UFCC has regular interval and width for triangular filters across the frequency range.

To perform the cepstral analysis of each frame of input audio signal we will divide the frequency domain of signal to ranges by the bank of triangular filter whose boundaries are calculated in the Bark scale. This scale is the result of a

research of the human ear's ability to perceive sounds at different frequencies. Converting to bark-scale is performed according to the formula:

$$B(f) \approx 6 \ln\left(\frac{f}{600} + \sqrt{1 + \left(\frac{f}{600} \right)^2} \right). \quad (6)$$

It is believed that the information that low frequency components of the speech signal are carrying is more important than the one carrying by the high frequency components, so Bark-scale is linear to 1 kHz and logarithmic above 1 kHz. That is, at low frequency filters are applied linearly, whereas at high frequencies - log. These filters are unevenly located on the frequency axis, because such filters are more at regions of the spectrum in low frequency (1 kHz) and less at high frequencies (above 1 kHz). Filters are squared modulus of the Fourier transform, and the values are logarithmed:

$$e_m = \ln\left(\sum_{k=0}^{N} X_k^2 H_{m,k} \right), m = 0, \ldots, N_{FB-1} \quad (7)$$

where N_{FB} is the number of filters, $H_{m,k}$ - weights coefficients of obtained filters.

This approach can partially eliminate the noise components in the frequency domain, as the most important frequencies of human voice are located at range from 70Hz to 3400Hz.

The last stage of the voice cepstral features selecting process is the using of discrete cosine transform (DCT, Discrete Cosine Transform), the result of which will be set of bark-frequency cepstral coefficients (BFCC), which are elements of the vector of individual characteristics of speakers:

$$c_i = \sum_{m=0}^{N_{FB-1}} e_m \cos\left(\frac{\pi i (m + 0.5)}{N_{FB}} \right), i = 0, \ldots, N_{BFCC}, \quad (8)$$

where e_m denotes the logarithm Fourier coefficients, N_{BFCC} - the number of factors (size feature vectors).

As a result, for each fragment of the original speech signal we get a finite set of bark-frequency cepstrum coefficients $c = c_1, c_2, \ldots, c_N$, containing N elements, which form the vector of the characteristic features of specific user voice. Figure 2 shows a set of speaker's voice feature vectors. Each vector contains 26 bark-frequency cepstral coefficients for 60 variants of uttered passphrase.

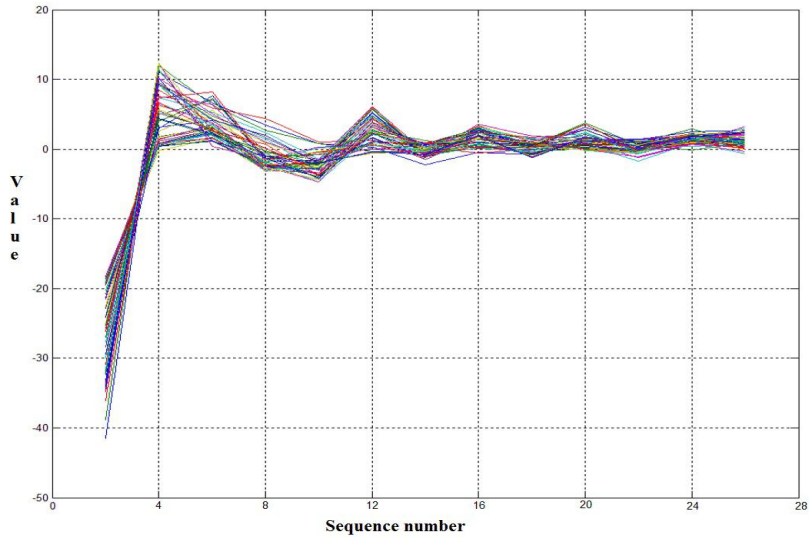

Figure.2. The set of bark-cepstral coefficients for single speaker

We have created two convolutional deep learning neural networks with using Caffe cross platform libraries of open source code [13,14] for the procedure of recognition – one for the recognition of speakers on spectrogram basis (Fig. 3) and another on the cepstral coefficients basis (Fig. 4).

Architecture is identical for each created neural network. Images with resolution of 500 by 100 pixels (in the figures it's indicated as 1@500x100) selected for speaker recognition features are fed to the inputs of these neural networks.

Image is fed to the input of convolutional layer C1, in which the convolution operation with core 4x4 collapse 4x4 is used to this image. This layer has optimized linear functions of neurons activation, as well as another layers of created neural network. This layer forms the features maps in frequency/time coordinates of spectrogram. These maps are submitted to sub-sampling layer S1, which has average-pooling activation function (all elements of the core are summed and then the sum is divided by the number of elements). This allows to increase the sensitivity of the neural network and do not use a layer of local normalization (Local Response Normalization layer), which would be necessary, if we use the max-pooling function.

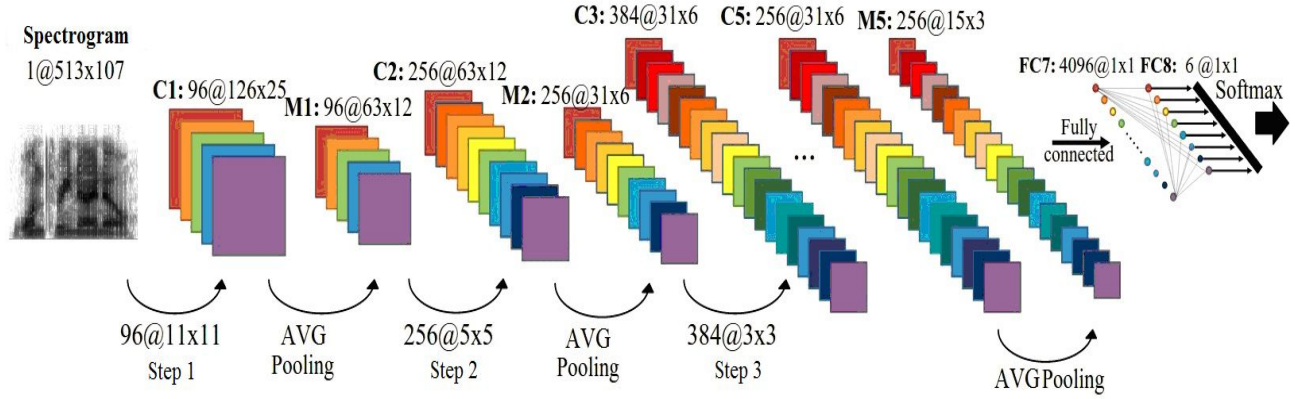

Figure. 3. Architecture of convolutional deep learning neural network for speaker recognition on spectrogram basis

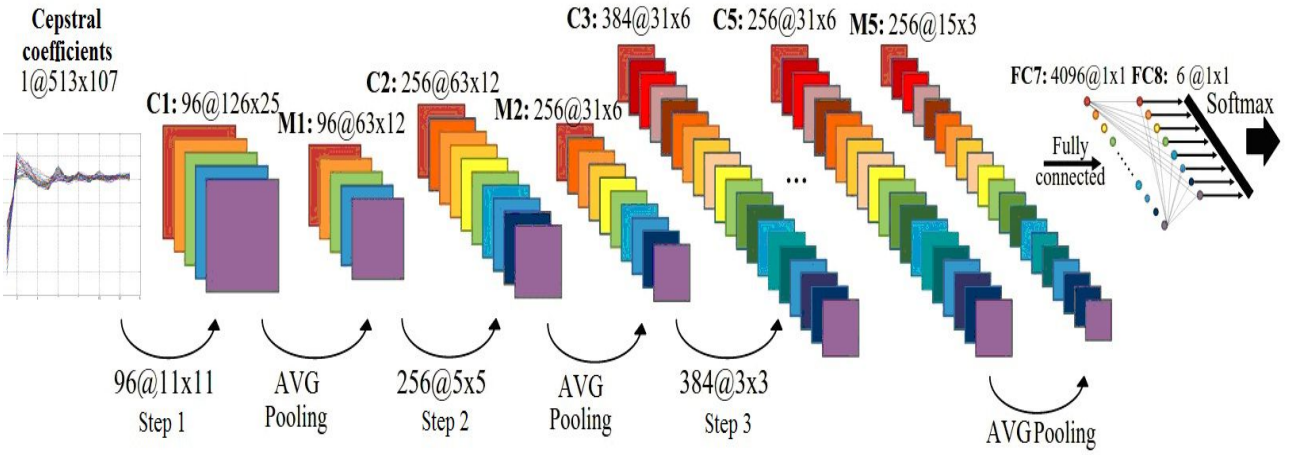

Figure.4. Architecture of convolutional deep learning neural network for speaker recognition on cepstral coefficients basis

Layer S1 is followed by convolution layer C2. Neurons in this layer are randomly united into two groups where each output is fed to the inputs of neurons in their group. Other options of this layer are similar to the previous layer (RLU, average-pooling). Convolution layers C3 and C5 are structurally similar to layer C2, the other parameters of layers are shown in Figures 3 and 4. To layers C3 and C4 grouping is not used.

Layer P6 is fully connected to the previous layer C5, and is an input layer of the perceptron that analyzes vectors of informative features, that are formed by convolutional neural network. This layer uses function Softmax of layer P7 that

allows it to identify one of the six speakers, to recognition of which neural networks were trained using the algorithm with adaptive learning rate (Adaptive Gradient learning Rate algorithm, AdaGrad)[15].

5. RESULTS AND DISCUSSION

As a reference, the database records we used is the free database NOIZEUS [16] – specialized database of the Eric Johnson School of Engineering and Computer Science at the University of Texas at Dallas, USA. It is used to analyze the algorithms of audio records improving and consists of records of typical household and industrial noises and of 30 sentences of English conversational speaking. These sentences were uttered by three men and three women (5 for each speaker, the signal sampling frequency is 25 kHz, but to adding of noise it was reduced to 8 kHz). During the experiment the training of the recognition module was carried out as to records of denoised passphrase and passphrase with added noise. The first case training sample contained 18 passphrase, the second – 576, where the artificial noise with levels of 0 dB, 5 dB, 10 dB and 15 dB was added to the pure signal.

To obtain cepstral coefficients the input signal was divided into frames of 20 ms duration, from each of them 12 cepstral coefficients, 12 delta coefficients (first derivative) and 12 double delta cepstral coefficients (second derivative) being allocated, Bank of 23 triangular bandpass filters covered frequency range of 40-4000 Hz, borders of filters were selected so that each pair of filters was overlapped by half, and at the frequency scale each filter began and ended in the middle of neighboring filter.

The training sample that was 60% of the base audio recordings was used to train the neural network. To test sample the remaining 40% audios was included. Summarized results of the experiment are presented in Figure 5, where the probability of correct recognition is calculated by the formula

$$P = \frac{\sum_i (Np_i)}{N}, \tag{9}$$

where Np_i is the number of correct recognition results of the *i*-th speaker, N – the total number of experiments.

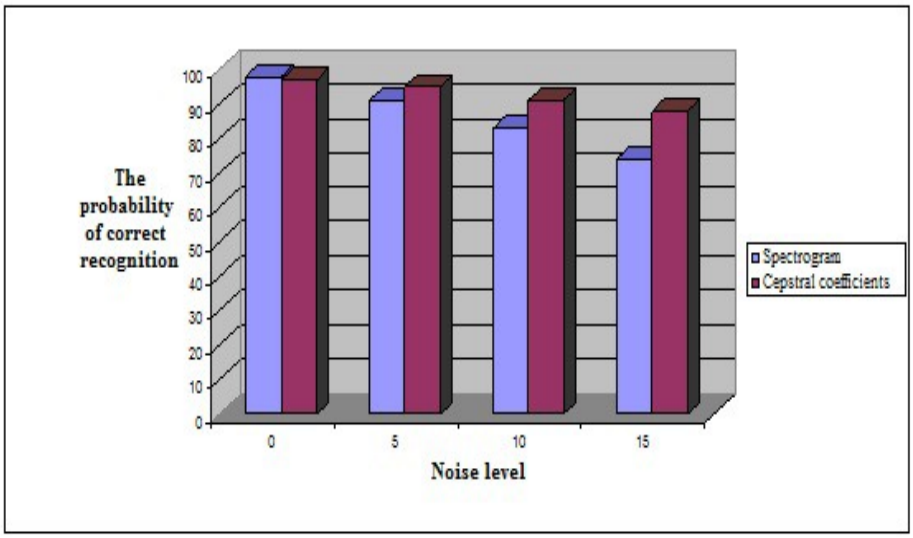

Figure. 5. Experimental results from recognition of speakers

Data in Figure 5 shows that for the worst case (noise level of 15 dB), the probability of correct recognition rate is 89% for 6 speakers.

6. CONCLUSIONS

Thus, the authors obtained the results that allow to suggest that the convolution deep learning neural network can be effectively used to make decisions in speaker recognition modules of automated system for critical application. Furthermore, the quality of the decision by the neural network depends on the type of selected features. Speakers recognition reliability with using of the signal spectrum is worse than the reliability of recognition using bark-cepstral coefficients. Additional efforts are also required for the study of neural convolutional network work directly with the acoustic parameters of the speech signal.

REFERENCES

[1] Bykov, M.M., Kovtun V.V. and Savinova N.G., "Evaluation of the noise influence on the reliability of the information-measuring system for voice recognition," Scientific Works of Vinnytsia National Technical University 3, 1-5 (2009).

[2] Alegre, F., Soldi G. and Evans N., "Evasion and obfuscation in automatic speaker verification," 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 749–753 (2014).

[3] Bennani, Y. and Gallinari P., "Connectionist approaches for automatic speaker recognition," ESCA workshop on speaker recognition, (1998).

[4] Dehak, N., Kenny P., Dehak R., Dumouchel P. and Ouellet. P.. "Front-End Factor Analysis For Speaker Verification," IEEE Transactions on Audio, Speech, and Language Processing 13(4), 788–798 (2011).

[5] Huang, P.-S., Avron H., Sainath T., Sindhwani V. and Ramabhadran B., "Kernel methods match deep neural networks on TIMIT," 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 205–209 (2014).

[6] Abdel-Hamid, O., Mohamed A., Jiang H. and Penn G., "Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition," 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 4277–4280 (2012).

[7] Dahl, G. E., Sainath T. N. and Hinton G. E., "Improving deep neural networks for LVCSR using rectified linear units and dropout," IEEE Proc. ICASSP, 8609–8613 (2013).

[8] Hinton, G. E., Deng L., Yu D., Dahl G. E., Mohamed A., Jaitly N., Senior A., Vanhoucke V., Nguyen P., Sainath T. N. and Kingsbury B., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," IEEE Signal Process. Mag. 29(6), 82–97 (2012).

[9] Zheng, F., Zhang G. and Song Z., "Comparison of different implementations of MFCC," J. Computer Science&Technology 16(6), 582-589 (2001).

[10] Shannon, B.J. and Paliwal K.K., "A comparative study of filter bank spacing for speech recognition," Proc. of Microelectronic engineering research conference, Brisbane, Australia, Nov. 2003, (2003).

[11] Skowronski, M.D. and Harris J.G., "Exploiting independent filter bandwidth of human factor cepstral coefficients in automatic speech recognition," Journal of the Acoustical Society of America 116(3), 1774–1780 (2004).

[12] Jia, Y., Shelhamer E., Donahue J., Karayev S., Long J., Girshick R., Guadarrama S. and Darrell T., "Caffe: Convolutional architecture for fast feature embedding," Proceedings of the 22nd ACM international conference on Multimedia, 675-678 (2014).

[13] Mashkov, V., Smolarz, A., Lytvynenko, V., Gromaszek, K., "The problem of system fault-tolerance," Informatyka Automatyka Pomiary w Gospodarce i Ochronie Środowiska 4(4), 41-44 (2014).

[14] Lal-Jadziak, J., "(Noise in noise measurement by means of correlation method," Przegląd Elektrotechniczny 92(11), 179-182 (2016).

[15] Anjos, A., Shafey L. E., Wallace R., Gunther M., McCool C. and Marcel S., "Bob: a free signal processing and machine learning toolbox for researchers," 20th ACM Conference on Multimedia Systems (ACMMM) Nara Japan, 1449-1452 (2012).

[16] Loizou, P., "NOIZEUS: A noisy speech corpus for evaluation of speech enhancement algorithms," <http://ecs.utdallas.edu/loizou/speech/noizeus/> (05.01.2017).