Journal of Engineering and Applied Sciences 13 (13): 5096-5104, 2018
ISSN: 1816-949X

Deep Residual Network for Sound Source Localization in the Time Domain

[1]Dmitry Suvorov, [2]Ge Dong and [1]Roman Zhukov
[1]Center for Space Research, Skolkovo Institute of Science and Technology,
143026 Moscow, Russia
[2]School of Aerospace Engineering, Tsinghua University, 100084 Beijing, China

Abstract: This study presents a system for sound source localization in time domain using a deep residual neural network. Data from the linear 8 channel microphone array with 3 cm spacing is used by the network for direction estimation. We propose to use the deep residual network for sound source localization considering the localization task as a classification task. This study describes the gathered dataset and developed architecture of the neural network. We will show the training process and its result in this study. The developed system was tested on validation part of the dataset and on new data capture in real time. The accuracy classification of 30 m sec sound frames is 99.2%. The standard deviation of sound source localization is 4°. The proposed method of sound source localization was tested inside of speech recognition pipeline. Its usage decreased word error rate by 1.14% in comparison with similar speech recognition pipeline using GCC-PHAT sound source localization.

Key words: Sound source localization, microphone array, deep neural network, residual network, audio processing, proposed method

INTRODUCTION

The purpose of the research is to develop a sound source localization system based on data obtained from a linear compact microphone array. The system should be resistant to noise and reverberation and also should be able to work in real time on conventional personal computers.

A large amount of noise and reverberation in captured sound is the key problem for distant speech recognition systems (Woelfel and McDonough, 2009). To solve this problem, a sound signal can be captured by microphone array to perform sound source localization and beamforming. In this case, the full process of sound capture and processing will consist of the following steps (Kumatani *et al*., 2012):

- Sound capture with microphone array
- Sound source localization and tracking
- Beamforming
- Post-filtering

Sound source localization is the key element in this architecture because its accuracy defines quality of algorithms for implementation at further stages. Beamforming and post-filtering use previously defined sound source direction as input parameter.

At the moment there are a large number of methods for sound source localization: weighted GCC-PHAT (Grondin and Michaud, 2015) and its analogs which use sound channels correlation. The baseline Version of GCC-PHAT is presented in Eq. 1 and 2:

$$GCC_{Kl}(\tau)=\int \frac{Y_k(\omega)Y_l(\omega)e^{i\omega}}{\left|Y_k(\omega)\right|\left|Y_l(\omega)\right|}d\omega \quad (1)$$

where, $Y_k(\omega)$ and $Y_l(\omega)$ are discrete fourier transforms of k and l channels of the sound frame from the microphone array. Likelihood of presence of active sound source at direction Θ_i:

$$\log lik(\theta_i)=\frac{1}{M}\sum_{Kl}^{M} GCC_{kl}\left(\tau_{kl}^{*}(\theta_i)\right) \quad (2)$$

Where:

M = A number of channels
Θ = A direction (azimuth for a linear microphone array, azimuth and elevation for planar and 3D configurations)
$\tau_{kl}^{*}(\Theta_i)$ = A theoretical delay between k and l channels for Θ_i direction of arrival

IDOA algorithms (Tashev and Acero, 2006) estimating phase delays on different frequencies between channels of captured multichannel sound (Eq. 3-7).

Corresponding Author: Dmitry Suvorov, Center for space research, Skolkovo Institute of Science and Technology, 143026 Moscow, Russia

Likelihood of presence of active sound source with frequency ω at direction Θ_i:

$$\log lik(\theta_i \mid \omega) = \frac{-\left\| mod(\delta(\omega) - \Delta(\omega, \theta_i), 2\pi) \right\|_2^2}{\left\| \frac{\partial \Delta}{\partial \theta}(\omega, \theta_i) \right\|} \quad (3)$$

Where:

$\Delta(\omega, \Theta_I)$ = A vector of theoretical phase differences between k and zero microphones at frequency

ω = The active sound source located at direction

Θ_i and = A vector of measured phase differences between k and zero microphones at frequency ω

$$\delta_k(\omega) = \angle Y_k(\omega) - \angle Y_0(\omega) \quad (4)$$

$$\delta(\omega) = \left[\delta_1(\omega), \delta_2(\omega), ..., \delta_{M-1}(\omega)\right] \quad (5)$$

Probability of presence of active sound source with frequency ω at direction Θ_i:

$$P(\theta_i \mid \omega) = \frac{\exp\left(\frac{\log lik(\theta_i \mid \omega)}{\sigma}\right)}{\left(\frac{\log lik(\theta_i \mid \omega)}{\sigma}\right)} \quad (6)$$

The most probable direction to the wideband sound source:

$$\hat{\theta} = \arg\max_{\theta}\left(\Sigma_{\omega} P(\theta_i \mid \omega)\right) \quad (7)$$

Scanning of the surrounding area with delay-and-sum beamformer (Valin *et al*., 2007) or other types of beamformers.

Likelihood of presence of active sound source with frequency ω at direction Θ_i when scanning is performed using delay-and-sum beamformer:

$$\log lik(\theta_i \mid \omega) = \frac{1}{M}\sum_{m=0}^{M-1} e^{j\omega\tau_m^*(\theta_i)} Ym(\omega) \quad (8)$$

The most probable direction to sound source can also be calculated using Eq. 6 and 7. MUSIC algorithms (Ishi *et al*., 2009) and their modifications. Probability of presence of active sound source with frequency ω at direction Θ_i:

$$\log lik(\theta_i\ \omega) = \frac{1}{\alpha(\omega, \theta_i)^H \left(I - U_s U_S^H\right)\alpha(\omega, \theta_i)} \quad (9)$$

Where:

$\alpha(\omega, \Theta_i)$ = The capturing matrix with size M by J` (J is a number of sound sources)

U_s = Signal subspace eigenvectors matrix (Tashev, 2009)

The most probable direction to sound source can also be calculated using Eq. 6 and 7. Sound source localization algorithms based on deep neural networks (Yalta *et al*., 2017) using convolutional and residual layers. Vector of probabilities of presence of sound source at possible directions:

$$P(\theta_0, ..., \theta_{N-1}) = F\left(Y_0(\omega), ..., Y_{M-1}(\omega)\right) \quad (10)$$

where, N is a number of checking directions. The architecture proposed by Yalta *et al*. (2017) is shown in Fig 1. Human speech localization algorithms based on processing data from microphone array and video camera (Suvorov and Zhukov, 2017).

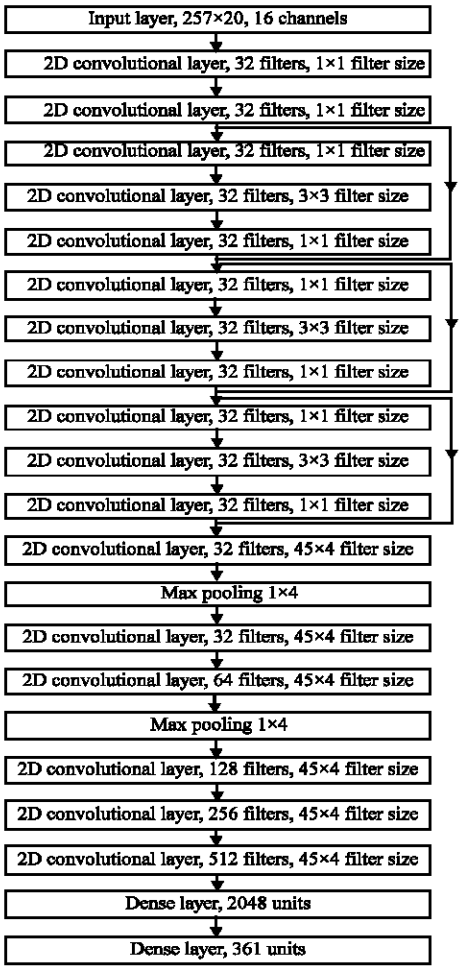

Fig. 1: Residual CNN proposed by Yalta *et al*. (2017). Each convolutional layer is followed by the ReLU non-linearity

All the methods except the one based on neural networks consider the localization problem as a problem of testing the hypothesis about sound source presence in a specific space sector which leads to an increase in required computing power as it is needed to check sound source presence in the surrounding space with a specified step (Tashev, 2009). Also, their implementations use assumptions about the plane front of an acoustic wave (Tashev, 2009) which leads to errors in localization of sound sources located closely to a microphone array (Ronzhin and Karpov, 2008).

The method based on neural networks described by Yalta *et al.* (2017) considers the localization problem as a problem of sound frame classification into sound source direction classes and the classification of active sound source absence. As input data, the algorithm uses the Discrete Fourier Transform (DFT) for every channel with sound duration of about several tens of milliseconds with some previous frames. Necessity in DFT computation for each channel on each iteration and use of two-dimensional convolutional layers, lead to increased computing complexity of the algorithm. Moreover, the algorithm uses only amplitude information from DFT and doesn't use phase information. It can also negatively affect the accuracy of localization.

Further in the study, a sound source localization method based on deep convolutional neural networks using as input, multichannel sound frames with fixed duration from microphone array will be proposed. Unlike in the method introduced in (Yalta *et al.*, 2017) the network uses only one-dimensional convolutions which significantly reduces its computing complexity. Also the process of training dataset collection, neural network training and system testing will be described.

MATERIALS AND METHODS

Dataset: To perform experiments with deep neural network training, a big dataset of labeled data is required. To solve this problem a python application was developed which plays sound via. a speaker whilst simultaneously recording it with an 8-channel microphone array with 3 cm spacing, implemented on the basis of MEMS microphones with PDM interface (Suvorov and Zhukov, 2017) which is shown in Fig. 2. The application randomly chooses and plays a music file for a duration of 30 sec from an array of one-channel sound files from "GTZAN genre collection" collected in the framework by Tzanetakis and Cook (2002). In this way, one-hour multichannel sounds for each direction with a 10° step from 0-180° were recorded. One hour of silence was also recorded. Everything was recorded in a 2×3 m room. The sound was recorded with 16 kHz frequency with 16 bit resolution.

As dataset was collected with a linear microphone array, further the task of sound source localization was considered as a task for estimation of azimuth to sound source because the use of linear microphone array makes it impossible to determine an elevation angle for obvious geometric reasons.

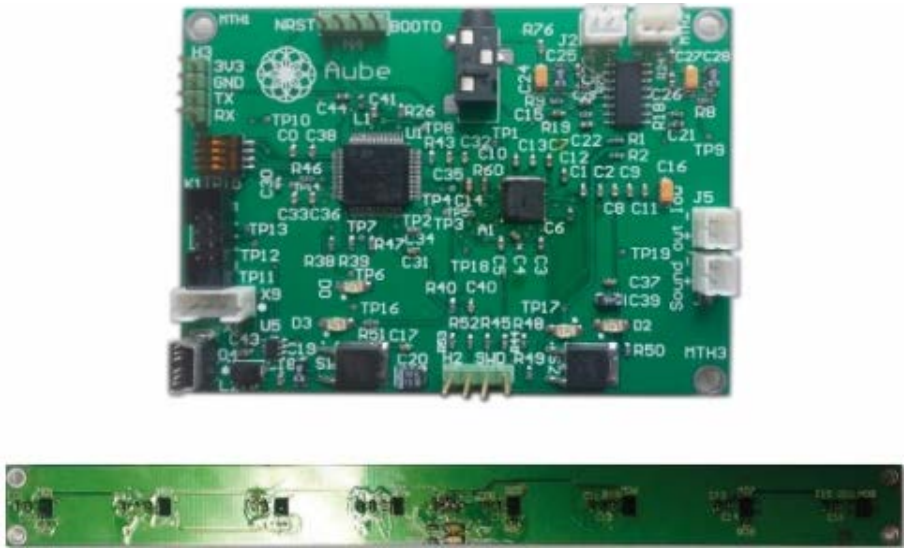

Fig. 2: Linear microphone array used for capturing the dataset and real time experiments with proposed sound source localization

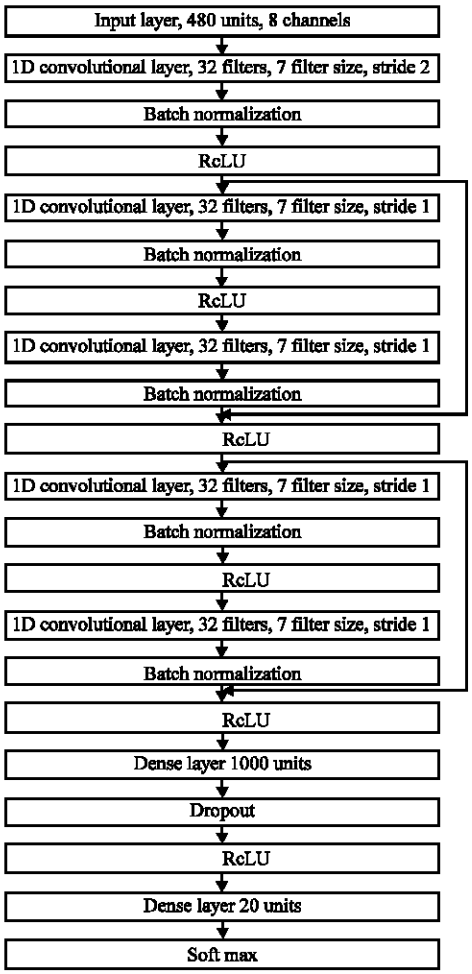

Fig. 3: A deep neural network, used for sound source localization

Neural network architecture: The developed neural network architecturally consists of four big blocks (Fig. 3):

- Input layer, accepting 8-channel sound frames from microphone array with duration of 480 samples (30 m sec) in float format
- First 1D convolutional layer (Eren, 2017) performing primary feature extraction (Eq. 11)
- Block consisting of two residual layers (He *et al*., 2016). Residual layers allow a delay in overfitting of neural networks and therefore train deeper networks
- Decision-making blocks, consisting of two fully connected layers, create outputting probabilities that a sound frame has a sound from one of the possible azimuths and probability of absence of any active sound sources in the frame:

$$V(x,t)=\sum_{i=x-\frac{L-1}{2}}^{x+\frac{L-1}{2}}\sum_{S=1}^{S}K\left(i-x+\frac{L-1}{2},S,t\right)U \qquad (11)$$

Where:

U(x, s) = A 1D input signal containing S channels
t = Number of output channel
K(x, s, t) = A matrix of size L by S of the filter for t output channel

After each convolutional layer, a batch norm layer is used to allow to train neural networks with a lesser number of iterations to postpone overfitting (Ioffe and Szegedy, 2015). Batch Normalization Transform is shown in Eq. 12-15.

Mini-batch mean:

$$\mu\beta=\frac{1}{m}\sum_{i=1}^{m}x_i \qquad (12)$$

Mini-batch variance:

$$\sigma_\beta^2=\frac{1}{m}\sum_{i=1}^{m}(x_i-\mu\beta)^2 \qquad (13)$$

Normalize:

$$\hat{x}_i=\frac{x_i-\mu\beta}{\sqrt{\sigma_\beta^2-\varepsilon}} \qquad (14)$$

Final scale and shift:

$$y_i=\gamma x_i+\beta \qquad (15)$$

Where:

m = A batch size
x = A batch of input data
γ and β = Parameters to be learned

After the first fully connected layer, a dropout layer is also used to delay the moment of overfitting to later iterations (Srivastava *et al*., 2014). Feed-forward operation of the dropout layer:

$$\gamma_j^l=\text{Bernoulli}(p) \qquad (16)$$

$$\hat{y}^l=r^l*y^l \qquad (17)$$

Where:

r^l = A vector of independent bernoulli random variables each of which has probability p of being 1
* = An element-wise product

ReLU non-linearity (Maas *et al*., 2013) is used after all the convolutional and fully connecting layers with the exception of the last dense layer:

$$f(x)=\max(0,x) \qquad (18)$$

On the last layer SoftMax non-linearity (Maas *et al*., 2013) is used as it is needed to normalize output of the neural network in such a way that the sum of all probabilities were equal to 1:

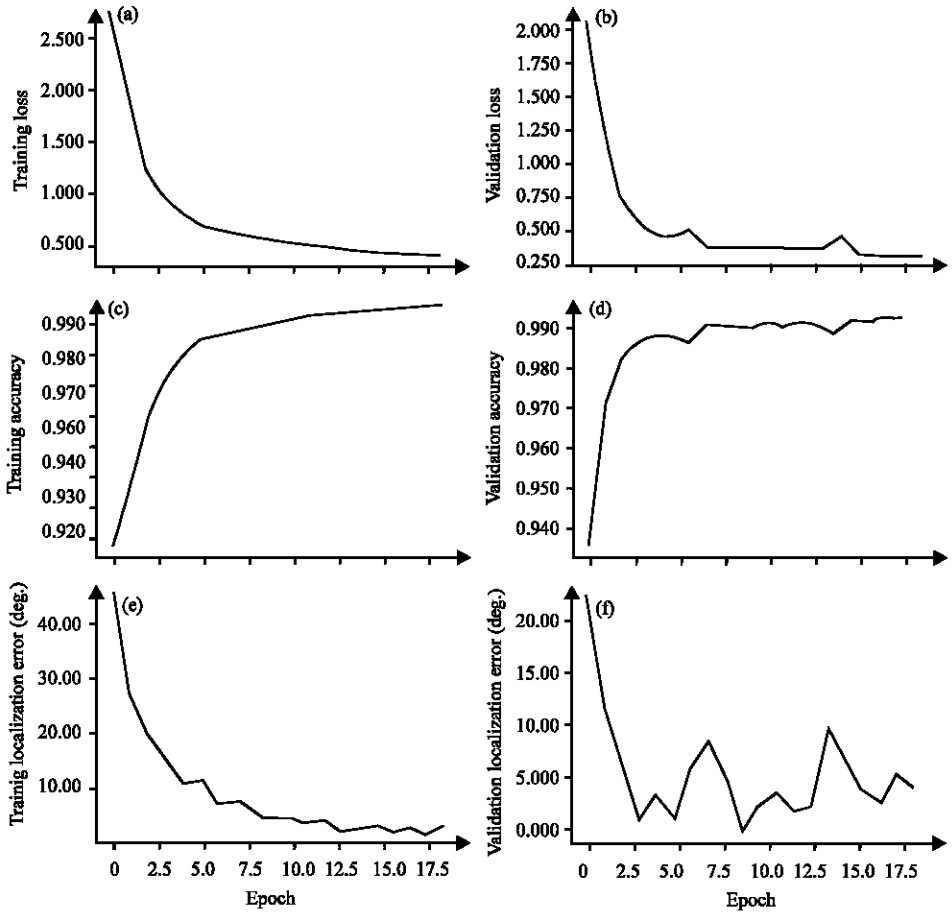

Fig. 4: The learning process of the developed system

$$f(x_i)=\frac{e^{x_i}}{\sum_j e^{x_j}} \quad (19)$$

Unlike the solution proposed in (Yalta *et al*., 2017) the network accepts original signals but not its fourier image. It is possible as the fourier transformation is essentially decomposition into narrowband components and therefore one-dimensional convolutional layers are able to learn this decomposition themselves.

Training and testing: A prototype of the proposed system was realized with python based on the Theano and Lasagne libraries. The learning was done with the Adam optimization algorithm (Kingma and Ba, 2015) with a low parameter of training speed (Eq. 21-23). The learning was done in 20 epochs on a NVIDIA GeForce GTX 1070 graphics card using cuda technology. Cross-entropy was used as a loss function (Eq. 24). In the training process, the values of the loss function, the accuracy of the classification and the standard deviation of the azimuth determination error were monitored. In Fig. 4, it can be seen that overfitting happened only after 18 epochs:

$$w[t+1] = w[t]-\alpha\frac{1}{\sqrt{g[t+1]+\varepsilon}}v[t+1] \quad (20)$$

$$g[t+1] = \mu g[t]+(1-\mu)\nabla(L,w[t])\nabla(L,w[t]) \quad (21)$$

$$v[t+1] = \beta v[t]+(1-\beta)\nabla(L,w[t]) \quad (22)$$

Where:

t = Iteration number
L = Loss function
w = Set of trainable parameters of the network
ε, μ and β = Scalar parameters of the algorithm

$$L(p,y)=\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}y_{ij}\log(p_{ij}) \quad (23)$$

where, matrix p is N×M output of the neural network, matrix y is a one-hot encoded real class identifier, M is set to a number of classes, N is a batch size.

To analyze the training results t-SNE visualization (Maaten and Hinton, 2008) for features generated by the penultimate fully connected layer was implemented (Fig. 5). It clearly shows the cluster structure of features and the mutual arrangement of clusters corresponding to the spatial arrangement of real azimuths which indicates the good quality of the neural network training.

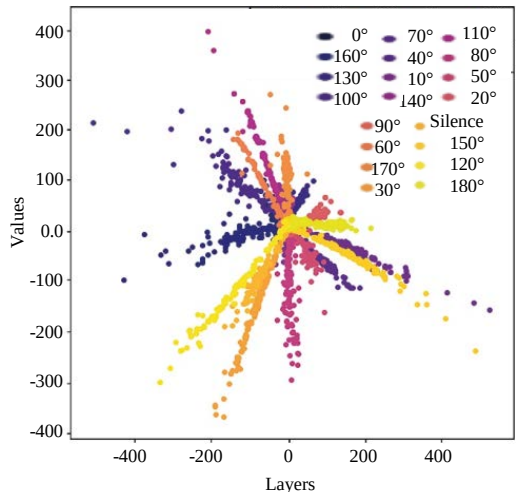

Fig. 5: T-SNE visualization of penultimate layer features

RESULTS AND DISCUSSION

Evaluation in real time: An application was developed which allows sound capture from the microphone array in real time and determines the direction of the sound source azimuth. Using this application, sound source directions were calculated in real time for sound sources in a previously known position. Measurements were done in a room where the training dataset was recorded and in another room that had a significantly different area and filling meaning it had different reverberation parameters. It can be seen in Fig. 6 that average absolute values of sound source direction azimuth determination error did not exceed 12° in both cases which is a good result, considering that the neural network was trained with an azimuth step of 10°. Mostly the same accuracy of localization in the new room and room where the train dataset was recorded, indicates a good generalization property for the trained neural network.

Figure 7 gives an example of results of continuous localization of a stationary source that plays music. It can be seen that localization error is complexity of choice between neighbouring classes of neural networks.

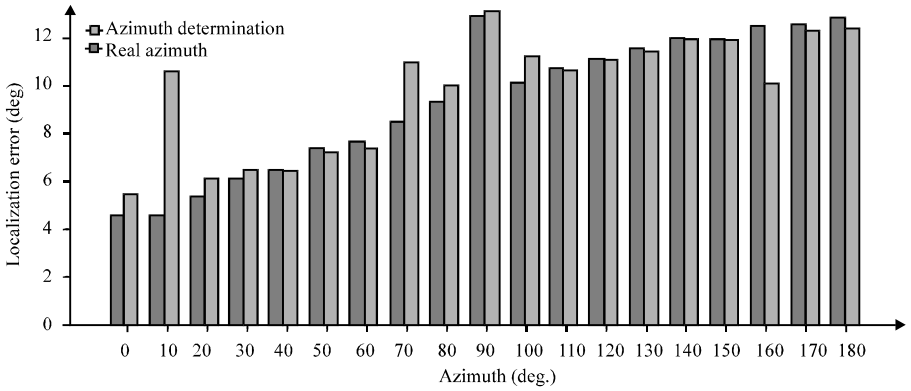

Fig. 6: Dependence of average absolute values of azimuth determination error from real azimuth with active sound source. Light grey bars show measurements conducted in the new room. Dark grey bars show measurements conducted in the room where the training dataset was recorded

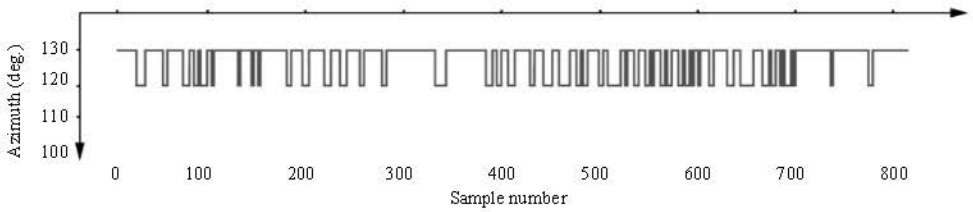

Fig. 7: Values of calculated azimuths for an active sound source as a function of sound frame number with real direction of 110°

Impact on the whole speech recognition pipeline: A comparison of the accuracy of far-field speech recognition was performed with three configurations of speech recognition pipeline to understand the impact of developed sound source localization on the final result of speech recognition. The first speech recognition pipeline didn't use microphone array processing:

- Audio capturing from the first channel of the microphone array
- Voice activity detection using code from the WebRTC project
- Speech recognition using Google speech API

The second speech recognition used sound source localization developed by Grondin and Michaud (2015):

- Audio capturing from the microphone array
- Sound source localization using weighted GCC-PHAT with Kalman filtering
- MVDR beamformer (Tashev, 2009)
- Zelinski post-filter (Aleinik, 2017)
- Speech recognition using Google speech API

Implementations of MVDR beamformer and Zelinski post-filter were used from BTK toolkit. And the last speech recognition pipeline used developed sound source localization based on the residual network:

- Audio capturing from the microphone array
- Proposed sound source localization using residual network with Kalman filtering
- MVDR beamformer
- Zelinski post-filter
- Speech recognition using Google speech API

The 100 phrases were recognized simultaneously through 3 described speech recognition pipelines. Speech recognition pipelines shared the same microphone array during the experiment. Voice sound sources were located on distance 1.5 m from the microphone array at different directions. Word Error Rates (WER) were calculated and compared for results from pipelines (Table 1).

The best result was shown by the solution with proposed sound source localization. High WER is shown by the solution with GCC-PHAT because the width of the beam pattern formed by the MVDR beamformer is lower than the accuracy of the sound source localization achieved GCC-PHAT on used microphone array, so, sometimes beam pattern became orientated not to the sound source. The width of the beam pattern of the MVDR beamformer is about 20° (Fig. 8). Average localization error of the developed sound source

Table 1: WER value for different configurations of speech recognition pipeline

Speech recognition pipeline	WER (%)
Mono audio capturing without any speech enhancement	2.21
Speech enhancement using beamforming and GCC-PHA sound source localization	2.99
Speech enhancement using beamforming and proposed sound source localization	1.85

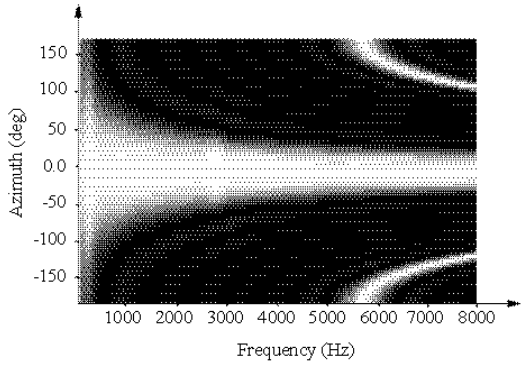

Fig. 8: The beam pattern of the MVDR beamformer in the endfire orientation for used microphone array

localization system is not higher than 12°. So, their combination achieves a good quality of speech enhancement resulting in low WER. Directivity pattern was modeled using following equations (Vary and Martin, 2006):

$$\psi(\omega,\Theta) = \left|H(\omega)^{H} u(\omega,\Theta)\right|^{2} \quad (24)$$

MVDR filter coefficient vector (Vary and Martin, 2006):

$$H(\omega) = \frac{\Phi(\omega)_{NN}^{-1} a(\omega)}{a(\omega)^{H} \Phi(\omega)_{NN}^{-1} a(\omega)} \quad (25)$$

The noise cross-power spectral matrix (Tashev, 2009):

$$\Phi NN(\omega) = \Phi_{N'N'}(\omega) + \Phi_{II}(\omega) \quad (26)$$

The instrumental noise cross-power spectral matrix (Ishi *et al*., 2009):

$$\Phi_{II}(\omega) = N_{I}^{2}(\omega) I \quad (27)$$

where, $N_I(\omega)$ is the magnitude of the instrumental noise in single microphone. The cross-spectral density for an isotropic noise field (Tashev, 2009):

$$\Phi_{ij}(\omega) = N_{o}(\omega) \sin c\left(\frac{wd_{ij}}{v}\right) \quad (28)$$

$$\Phi_{N'N'}(\omega)=\begin{bmatrix}\Phi_{11}(\omega) & \cdots & \Phi_{1M}(\omega)\\ \vdots & \ddots & \vdots\\ \Phi_{M1}(\omega) & \cdots & \Phi_{MM}(\omega)\end{bmatrix} \quad (29)$$

Where:

$N_O(\omega)$ = The noise spectrum captured by an omnidirectional microphone
v = The speed of sound
d_{ij} = A distance between i and j microphones

The propagation vector for linear microphone array (Vary and Martin, 2006):

$$a(\omega) = \left[a_i: e^{\frac{j\omega p_i}{v}}, i=1, ..., M\right] \quad (30)$$

where, p_i is a position of i microphone. The unit vector in the required direction of beam pattern (Vary and Martin, 2006):

$$u(\omega,\Theta) = \left[u_i: e^{\frac{j\omega\cos(\Theta)p_i}{v}}, i=0, ..., M-1\right] \quad (31)$$

CONCLUSION

A sound source localization method based on deep residual neural networks was developed. It doesn't require a captured signal to be transformed from time domain to frequency domain with fourier transformation which positively affects system performance. The developed method demonstrated good accuracy of the sound source direction azimuth determination with a linear compact microphone array even without the consideration of object dynamics with a Kalman filter or particle filter. As a further improvement to the method, the system can be trained in such a way that will allow us to determine several sound source locations simultaneously.

SUGGESTIONS

Also in future research, the architecture should be complemented by LSTM, BLSTM or GRU layers (Chung *et al*., 2014) to make the network able to consider object dynamics.

ACKNOWLEDGEMENT

This research was supported by the Russian Innovation Support Fund (project 102GRNTIS5/26071).

REFERENCES

Aleinik, S., 2017. Acceleration of Zelinski post-filtering calculation. J. Signal Process. Syst., 88: 463-468.

Chung, J., C. Gulcehre, K. Cho and Y. Bengio, 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. Proceedings of the NIPS Workshop on Deep Learning, December 8-13, 2014, NIPS, Montreal, Quebec, Canada, pp: 1-9.

Eren, L., 2017. Bearing fault detection by one-dimensional convolutional neural networks. Math. Prob. Eng., 2017: 1-9.

Grondin, F. and F. Michaud, 2015. Time difference of arrival estimation based on binary frequency mask for sound source localization on mobile robots. Proceedings of the IEEE-RSJ International Conference on Intelligent Robots and Systems (IROS), September 28-October 2, 2015, IEEE, Hamburg, Germany, ISBN:978-1-4799-9994-1, pp: 6149-6154.

He, K., X. Zhang, S. Ren and J. Sun, 2016. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 26-July 1, 2016, IEEE, Las Vegas, Nevada, USA., ISBN:9781509014385, pp: 770-778.

Ioffe, S. and C. Szegedy, 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. Proceedings of the 32nd International Conference on Machine Learning, July 07-09, 2015, Microtome Publishing, Lille, France, pp: 448-456.

Ishi, C.T., O. Chatot, H. Ishiguro and N. Hagita, 2009. Evaluation of a MUSIC-based real-time sound localization of multiple sound sources in real noisy environments. Proceedings of the IEEE-RSJ International Conference on Intelligent Robots and Systems IROS, October 10-15, 2009, IEEE, St. Louis, Missouri, USA., ISBN:978-1-4244-3803-7, pp: 2027-2032.

Kingma, D. and J. Ba, 2015. Adam: A method for stochastic optimization. Proceedings of the International Conference on Learning Representations ICLR, May 7-9, 2015, San Diego, California, USA., pp: 1-15.

Kumatani, K., J. McDonough and B. Raj, 2012. Microphone array processing for distant speech recognition: From close-talking microphones to far-field sensors. IEEE. Signal Process. Mag., 29: 127-140.

Maas, A., A. Hannun and A. Ng, 2013. Rectifier nonlinearities improve neural network acoustic models. Proc. ICML., 30: 1-6.

Maaten, L.V.D. and G. Hinton, 2008. Visualizing data using T-SNE. J. Machine Learn. Res., 9: 2579-2605.

Ronzhin, A. and A. Karpov, 2008. [Comparison of methods for localization of multimodal system user by his speech (In Russian)]. J. Instrum. Eng., 51: 41-47.

Srivastava, N., G.E. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, 2014. Dropout: A simple way to prevent neural networks from overfitting. J. Mach. Learn. Res., 15: 1929-1958.

Suvorov, D. and R. Zhukov, 2017. Device for synchronous data capturing from the array of MEMS microphones with PDM interface (In Russian). IFI CLAIMS Patent Services Company, Madison, Connecticut.

Tashev, I. and A. Acero, 2006. Microphone array post-processor using instantaneous direction of arrival. Proceedings of the International Workshop on Acoustic, Echo and Noise Control, September 12-14, 2006, IWAENC, Paris, France, pp: 1-4.

Tashev, I., 2009. Sound Capture and Processing: Practical Approaches. John Wiley & Sons, New York, USA., ISBN:9780470319833, Pages: 388.

Tzanetakis, G. and P. Cook, 2002. Musical genre classification of audio signals. IEEE Trans. Speech Audio Process., 10: 293-302.

Valin, J.M., F. Michaud and J. Rouat, 2007. Robust localization and tracking of simultaneous moving sound sources using beam forming and particle filtering. Rob. Auton. Syst., 55: 216-228.

Vary, P. and R. Martin, 2006. Digital Speech Transmission: Enhancement, Coding and Error Concealment. John Wiley & Sons, New York, USA., ISBN-13:978-0-471-56018-9, Pages: 607.

Woelfel, M. and J. McDonough, 2009. Distant Speech Recognition. John Wiley & Sons, New York, USA., ISBN:9780470517048, Pages: 594.

Yalta, N., K. Nakadai and T. Ogata, 2017. Sound source localization using deep learning models. J. Rob. Mech., 29: 37-48.