ROBUST ADAPTIVE METHOD FOR SPEECH SIGNAL WAVEFORM ESTIMATION USING MICROPHONE ARRAY

A.S. Ivanenkov, A.A. Rodionov

Institute of Applied Physics of the Russian Academy of Sciences

ABSTRACT

This paper considers the scenario when a mix of signals from multiple acoustic sources is received by a microphone array. The problem is to estimate the waveform of the source of interest located in the near field of the array. The considered problem can arise in many applications such as video conferencing, acoustic room surveillance and others, when it is necessary to capture human speech against acoustic interferences. To solve this problem, an adaptive algorithm is independently applied to each narrow band of the received signal. Based on the model of interference including rank-deficient correlation matrix, a new method of robust adaptive processing is proposed. The results of numerical simulation and experiment demonstrating the robustness of the proposed method to imperfections in desired signal spatial model and to the finite sample size effect are presented.

Index Terms**—** microphone arrays, robust adaptive beamforming, maximum likelihood estimation

1. INTRODUCTION

Microphone array (MA) signal processing originated over 20 years ago is now one of the actual research trends in modern acoustics [1]. With the use of MAs, many problems related to localization, identification of acoustic sources, detection and estimation of acoustic signals can be solved [1, 2]. In the present work, the problem of adaptive estimation of speech signal of a single source via MA is considered. It is supposed that a speech source is located in the near field of the MA. This problem can be actual in various applications, when it is necessary to estimate speech of one or several people against multiple acoustic interferences (such as video conferencing, "virtual" microphone applications, etc. [1,2]).

Basic advantages of adaptive beamforming techniques, as compared to data-independent beamforming methods, include improved resolution and much better interference rejection capability. However, it is known that if the desired signal is present in beamforming training data and the steering vector corresponding to the signal of interest (SOI) is imprecisely known, the performance of the adaptive methods can degrade severely [3]. This situation is often the case in practice due to modeling errors such as array calibration errors, differences between the assumed and real positions of the signal source, errors of signal modeling, etc.

In the case of wideband signals, several approaches to design of adaptive beamformers have been proposed. In this work, the approach based on decomposition of received wideband signals into narrowband signals will be considered [4]. The idea of the technique is to apply narrowband beamformer to each narrowband signal. Because of independent optimization of each subband weight vector and, hence, ignoring of inter-subband relations of the received signal, this type of beamformer may be less effective as compared to other types of beamformers. However, the narrowband beamformer technique is more explicit and has been more intensively studied than the broadband technique.

Many techniques have been proposed to improve the robustness of narrowband beamformers. Popular approaches include the algorithms with imposed linear constraints, eigenspace-based beamformers, the diagonal loading method [5]. Recently, several similar beamformers have been proposed (see, e.g. [5-7]), which, unlike early techniques, explicitly model the error in the SOI steering vector and use the idea of worst-case-performance optimization. The main disadvantage of existing narrowband beamformer methods is that in most techniques, the choice of their parameters is not directly related to uncertainty of the steering vector, or the performance of the beamformer is dependent on the quantity of the uncertainty, which is often unknown in practice. So the additional problem – estimation of optimal algorithm parameters or the uncertainty of the steering vector – should be considered as well.

In the present work, an algorithm based on the original method presented in [8] is proposed to be used for estimating the SOI. In contrast to existing algorithms, in the proposed method, the interference correlation matrix is considered as some set of rank-one matrixes that are supposed to be unknown. Each rank-one matrix can be interpreted as correlation matrix of a single interference. The case when the number of interference sources is equal to the number of array elements N is equivalent to the case of completely unknown interference correlation matrix. Thus, the problem can be solved for a wide class of possible interferences. The algorithm is derived from maximum-likelihood estimation (MLE) method which ensures asymptotic efficiency of parameters estimation. In this case, the greatest signal-to-noise ratio is guaranteed.

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

2. SIGNAL WAVEFORM ESTIMATION ALGORITHM

Assume that a broadband SOI impinges on an array with *N* sensors. We divide the time dependence at each element output into J non-overlapping blocks with each block consisting of I samples. We then apply an *I*-point fast Fourier transform (FFT) to each block to obtain *I* narrowband frequency bins (frequency subbands). The data vector $\mathbf{x}_i^j$ for the *i*-th frequency bin and the *j*-th snapshot can be written as

$$\mathbf{x}_j^i = \mathbf{a}^i(\boldsymbol{\theta}^i)s_j^{i*} + \xi_j^i,\ i = 1,...,I,\ j = 1,...,J \qquad (1)$$

where $(\cdot)^*$ stands for complex conjugation, s_j^{i*} represents the unknown waveform of the SOI in the *i*-th frequency bin, $\mathbf{a}^i(\boldsymbol{\theta}^i)$ stands for SOI steering vector depending on the vector of unknown parameters $\boldsymbol{\theta}^i$ in the *i*-th frequency bin, ξ_m^i is the $N \times 1$ sensor noise plus interference vector that is considered to be white Gaussian zero-mean noise with unknown correlation matrix $\mathbf{K}_n^i$. Further, the signal in *i*-th narrow band will be considered and, for the sake of simplicity, the index "*i*" will be omitted. One of basic features of the method is to represent the unknown interference correlation matrix $\mathbf{K}_n$ as

$$\mathbf{K}_n = \sigma_0^2\mathbf{I} + \mathbf{K}_a \qquad (2)$$

where σ_0^2 is the power of microphone self-noise that is assumed to be a zero-mean spatially and temporally white Gaussian process, $\mathbf{K}_a = \sum_{m=1}^{M}\mathbf{a}_m\mathbf{a}_m^H$ stands for the part of interference correlation matrix that consists of M rank-one matrixes formed with the use of unknown vectors $\mathbf{a}_m$. For the sake of convenience we represent latter matrix as $\mathbf{K}_a = \mathbf{A}\mathbf{A}^H$, where $\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_M)$. The model (1) can be written in the matrix form as follows

$$\mathbf{X} = \mathbf{a}\mathbf{s}^H + \boldsymbol{\Xi} \qquad (3)$$

where $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_J)$, $\boldsymbol{\Xi} = (\xi_1, \ldots, \xi_J)$. The estimates of unknown parameters can be obtained with the use of MLE method [9]. For this case Log-likelihood function (LLF) can be easily obtained and written as

$$\Lambda = -J\left[\ln\det(\sigma_0^2\mathbf{I} + \mathbf{A}\mathbf{A}^H) + tr((\sigma_0^2\mathbf{I} + \mathbf{A}\mathbf{A}^H)^{-1}\mathbf{K}_s)\right], \qquad (4)$$

where $\mathbf{K}_s = J^{-1}(\mathbf{X} - \mathbf{a}\mathbf{s}^H)(\mathbf{X} - \mathbf{a}\mathbf{s}^H)^H$. Here the unknown parameters are the matrix $\mathbf{A}$ and the vector $\mathbf{s}$. Maximization of LLF for the considered model over the unknown vector of SOI waveform $\mathbf{s} = [s_1, ..., s_J]^T$ gives its estimate as [8]:

$$\hat{\mathbf{s}} = \mathbf{X}^H\mathbf{P}_a\mathbf{a}\big/\left(\mathbf{a}^H\mathbf{P}_a\mathbf{a}\right) \qquad (5)$$

where $\mathbf{P}_a = \mathbf{I} - \mathbf{A}(\sigma_0^2\mathbf{I} + \mathbf{A}^H\mathbf{A})^{-1}\mathbf{A}^H$, $\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_M)$. LLF maximization over other unknown parameters yields the system of equations that cannot be solved in analytical form. In this work, the following way of solution is proposed. Let us consider the model in which s_j is white Gaussian noise

$$\mathbf{x}_j = \boldsymbol{\tau}_j + \xi_j,\ j = 1, \ldots, J \qquad (6)$$

where $\boldsymbol{\tau}_j$ is the SOI, ξ_j is the interference, $E\{\boldsymbol{\tau}_i\boldsymbol{\tau}_j^H\} = \mathbf{K}_{\boldsymbol{\theta}}\delta_{i,j}$, $E\{\xi_i\xi_j^H\} = \mathbf{K}_a\delta_{i,j}$, $\mathbf{K}_{\boldsymbol{\theta}} = \sigma_1^2\mathbf{a}\mathbf{a}^H + \sigma_0^2\mathbf{I}$ stands for the correlation matrix of the desired signal depending on the vector of unknown parameters $\boldsymbol{\theta}$, σ_1^2 is the power of the desired signal and σ_0^2 is the power of microphone self-noise, $\mathbf{K}_a = \sum_{m=1}^{M}\mathbf{a}_m\mathbf{a}_m^H$. For the sake of computational simplicity, the component $\sigma_0^2\mathbf{I}$ was included to the desired signal. In contrast to the model (1), the model (6) is simpler since it includes only one unknown parameter σ_1^2 instead of J unknown parameters as in the model (1). However, the model (6) requires signal samples to be independent and Gaussian. It is clear that the model (6) can be used whether this condition is satisfied or not. LLF function for the considered model can be represented as

$$\Lambda = -J\left[\ln\det(\mathbf{K}_{\boldsymbol{\theta}} + \sum_{m=1}^{M}\mathbf{a}_m\mathbf{a}_m^H) + tr((\mathbf{K}_{\boldsymbol{\theta}} + \sum_{m=1}^{M}\mathbf{a}_m\mathbf{a}_m^H)^{-1}\hat{\mathbf{K}})\right]$$

where $\hat{\mathbf{K}} = J^{-1}\sum_{j=1}^{J}\mathbf{x}_j\mathbf{x}_j^H$ stands for the sample correlation matrix. Maximization of this LLF for the model (6) provides the following estimates of unknown parameters:

$$\hat{\sigma}_0^2 = \tfrac{1}{N-M}\sum_{M+1}^{N}c_m\left(\boldsymbol{\theta}\right), \mathbf{a}_m = \left(c_m\left(\boldsymbol{\theta}\right) - 1\right)/c_m\left(\boldsymbol{\theta}\right)\hat{\mathbf{K}}^{0.5}\mathbf{U}_m,$$

where $c_m(\boldsymbol{\theta})$, $\mathbf{U}_m$ stand for the eigenvalues and eigenvectors of the matrix $\mathbf{C}_\theta^{-1} = \hat{\mathbf{K}}^{0.5}(\mathbf{I} + \beta\mathbf{a}(\boldsymbol{\theta})\mathbf{a}(\boldsymbol{\theta})^H)^{-1}\hat{\mathbf{K}}^{0.5}$, parameter $\beta = \dfrac{\sigma_1^2}{\sigma_0^2}$ represents signal-to-noise ratio. The estimates of the parameters β and $\boldsymbol{\theta}$ can be found maximizing the expression

$$L_{\beta,\boldsymbol{\theta}} = -(N-M)\ln\sum_{m=M+1}^{N}c_m(\boldsymbol{\theta}) + \sum_{m=M+1}^{N}\ln c_m(\boldsymbol{\theta}).$$

In practice, the effective number of interference sources M is usually unknown. In the present work, it is proposed to estimate M using maximization of the signal-to-noise ration β over the number M, i.e.,

$$\hat{M} = \arg\max_M \beta\left(M\right).$$

The algorithm to compute waveform $\mathbf{s}$ can be summarized as follows:

1. Calculate sample covariance matrix $\hat{\mathbf{K}} = J^{-1}\sum_{j=1}^{J}\mathbf{x}_j\mathbf{x}_j^H$.

2. For the fixed number of interference sources M, calculate estimate of β : $\hat{\beta} = \arg\max_\beta \Lambda(\beta, \boldsymbol{\theta})$, where

$$\Lambda(\beta,\boldsymbol{\theta}) = -J(N-M)\ln\sum_{m=M+1}^{N}c_m(\boldsymbol{\theta}) + J\sum_{m=M+1}^{N}\ln c_m(\boldsymbol{\theta}) \ \ldots$$

$$-J\left(M + \ln\det\hat{\mathbf{K}}\right).$$

3. Estimate $\hat{\sigma}_0^2 = \frac{1}{N-\hat{M}} \sum_{\hat{M}+1}^{N} c_m(\boldsymbol{\theta})$ and $\hat{\mathbf{a}}_m = \sqrt{\left(c_m(\boldsymbol{\theta})-1\right)/c_m(\boldsymbol{\theta})}\,\hat{\mathbf{K}}^{0.5}\mathbf{U}_m$ where $c_m(\boldsymbol{\theta})$ and $\mathbf{U}_m$ are eigenvalues and eigenvectors of the matrix $\mathbf{C}_\theta^{-1} = \hat{\mathbf{K}}^{0.5}(\mathbf{I} + \hat{\beta}\mathbf{a}(\boldsymbol{\theta})\mathbf{a}(\boldsymbol{\theta})^H)^{-1}\hat{\mathbf{K}}^{0.5}$.

4. Obtain matrix $\mathbf{P}_a = \mathbf{I} - \hat{\mathbf{A}}(\hat{\sigma}_0^2\mathbf{I} + \hat{\mathbf{A}}^H\hat{\mathbf{A}})^{-1}\hat{\mathbf{A}}^H$ where $\hat{\mathbf{A}} = (\hat{\mathbf{a}}_1, \hat{\mathbf{a}}_2, \ldots, \hat{\mathbf{a}}_{\hat{M}})$.

5. Compute an estimate of desired signal waveform $\hat{\mathbf{s}} = \mathbf{X}^H\mathbf{P}_a\mathbf{a}/\left(\mathbf{a}^H\mathbf{P}_a\mathbf{a}\right)$.

6. Repeat the steps 2-5 for all possible M, e.g. from 1 to N. The best estimate of $\hat{\mathbf{s}}$ would be achieved when the value of the signal-to-noise parameter $\beta(M)$ is maximum.

3. SIMULATION RESULTS

The proposed method was tested with the use of numeric simulations. The MA consisted of ideal omnidirectional microphones that were grouped into two linear equidistant 16 element subarrays with the array spacing of 0.2 m which were located at the height of 2 meters (see Figure 1). In the horizontal plane of the room, at the height of 1.6 meters, five point sources emitting speech signal in the range from 0.1 to 3.2 KHz were placed. In Figure 3, the correlation coefficient of estimated and emitted signals of the first source in dependence on central frequency of subband is presented. The frequency subband width was equal to 3.8 Hz. The power of the SOI was less by 30 dB than the power of other sources and by 20 db greater than the power of the microphone self-noise. It was supposed that the location of the desired source was *a priori* known and the initial phases of microphones were randomly distributed in the range from minus 10 to plus 10 degrees. The steering vector of the desired signal source was modeled as steering vector of a spherical wavefront source. Four lines show the results for the proposed method (red line), conventional MVDR beamformer [10] (blue line), delay-and-sum beamformer [10] (green line) and worst-case performance optimization method [7] (black line), that explicitly model the uncertainty of the steering vector. One can see that the conventional MVDR beamformer and the nonadaptive delay-and-sum beamformer degrade severely in their performance. This is because of significant power of the interference sources and steering vector errors. The worst case performance optimization method demonstrates the best result. However, in order to implement this method one must know maximum Euclidian norm of the steering vector error in each frequency subband. This is a problem in practice. On the contrary, the proposed method (red line) demonstrates high quality of desired signal waveform estimation without *a prior* knowledge about desired signal error model. In the considered simulation the number of interference sources M was assumed to be unknown and was estimated using the proposed algorithm.

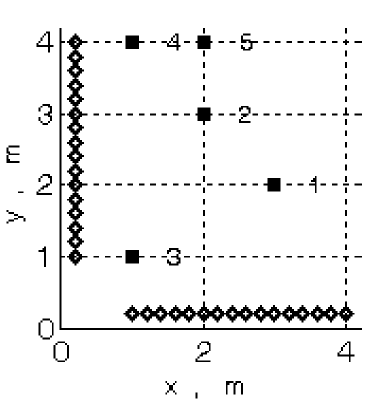

Fig. 1. Configuration of microphone array (diamonds) and sources (squares), top view.

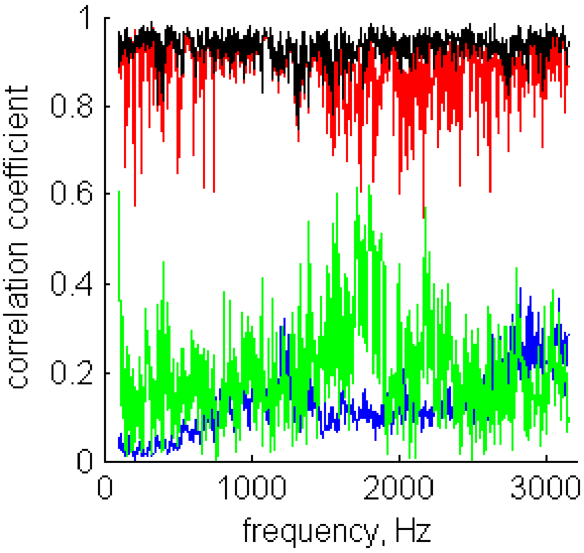

Fig. 2. Absolute value of correlation coefficient of emitted and estimated signals in dependence on frequency of subband with the width of 3.8 Hz. The results are presented for four methods: delay-and-sum beamformer (green line), conventional MVDR beamformer (blue line), worst-case performance optimization method (black line) and the proposed algorithm (red line). The signal length is 15 s.

The proposed method was also tested with the use of numerical simulations in which rectangular room with dimensions 5x5x3m was modeled (the reflections coefficients from the wall, the floor and the ceiling were 0.7, 0.4, 0.2, respectively). The reverberation was modeled with the use of the image method introduced in [11]. In Figure 3, absolute value of the correlation coefficient of estimated and emitted signals of the first source in dependence on subband central frequency is presented. The power of the SOI was less by 20 dB than the power of the other sources and by 20 db greater than the power of the microphone self-noise. Other parameters were as in previous simulation. Two graphs show the results for the proposed method (black line) and the delay-and-sum beamformer (red line). Besides, the results for the conventional MVDR beamformer and for the worst-case performance optimization method were obtained. In order to not obstruct the graph these results were not shown in

Figure 3. In the first case, the performance of the beamformer degrades severely in the presence of uncertainty in the initial phase of microphone and reverberation. In the latter case, the results are almost the same as for the case of the proposed method. But in this case, the Euclidian norm of the error of the steering vector should be known; this is the problem in practice. In contrast to the considered methods, the proposed method is shown to be quite robust to steering vector uncertainties without extra information about the error in the desired signal spatial model. Figure 4 shows the estimated number M of interference sources in dependence on central frequency of subband. One can see that the estimated number of interference sources does not equal to the real number of interference sources. The reason of this is reverberation which increases the number of effective sources.

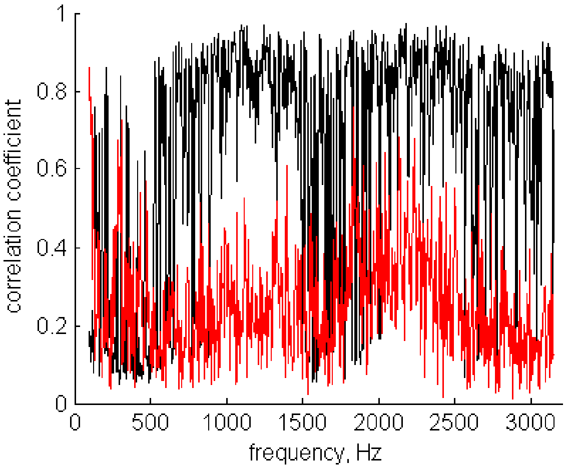

Fig. 3. Absolute value of correlation coefficient of emitted and estimated signals in dependence on frequency of subband with the width of 3.8 Hz. The results are presented for the two methods: delay-and-sum beamformer (red line) and the proposed algorithm (black line). The signal length is 15 s. The case with reverberation was considered.

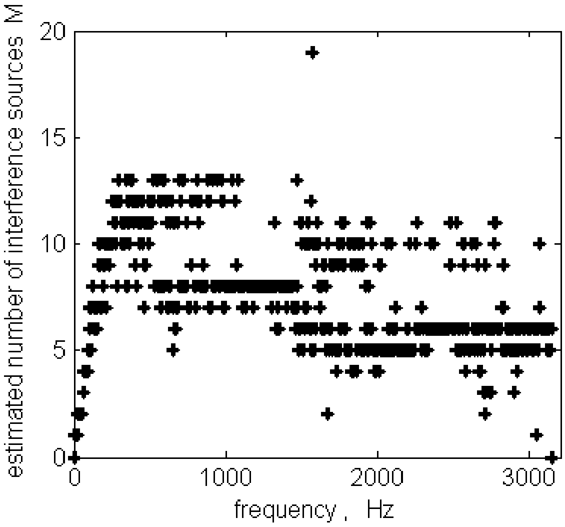

Fig. 4. Estimates of interference sources number M in dependence on central frequency of subband. The case with reverberation was considered.

4. EXPERIMENTAL RESULTS

In order to approve the proposed method, an experiment was conducted. The setup of the experiment included two linear microphone subarrays with 0.2 m spacing. Each subarray consisted of 14 elements. Subarrays were placed alongside x and y axes at the height of one meter from the horizontal plane where the sound sources were located. The experiment was carried out in anechoic chamber that enabled to eliminate reverberation and outside noise. Three loudspeakers emitting speech from 0.1 to 3.2 kHz were used as sound sources. In order to determine the coordinates of the sound sources, Capon method [10] was used. The idea of the method is to plot the function $F(x, y) = (\mathbf{a}^H \hat{\mathbf{K}}^{-1} \mathbf{a})^{-1}$ in dependence on unknown parameters. In our case, the unknown parameters are x and y source coordinates in the horizontal plane. In Figure 5, an acoustical image in logarithmic scale as a result of source localization is presented. The presented image was obtained with the use of Capon method. One can see that all three sources are localized but estimates of their coordinates have an appreciable error. Nevertheless, the weakest source was selected as a source of desired signal. Its coordinates were estimated using the acoustical image shown in Figure 5. The source of desired signal is marked with a black circle; its mean power was at least 14 dB less than the mean power of each interference source.

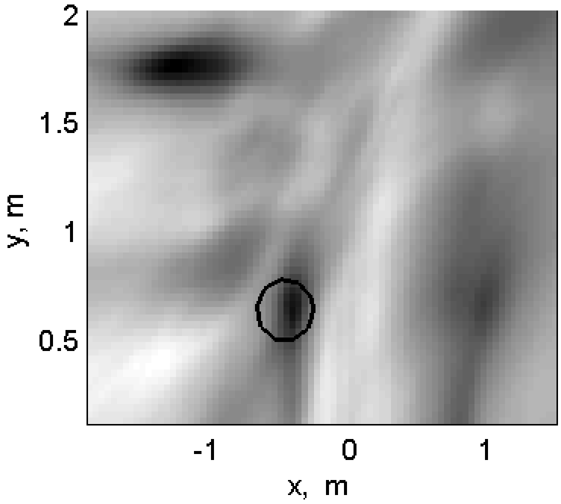

Fig. 5. Acoustical image of sources obtained with the use of Capon method. The source of desired signal is marked with the black circle. The mean power of the desired signal was 14 dB less than the mean power of each interference source.

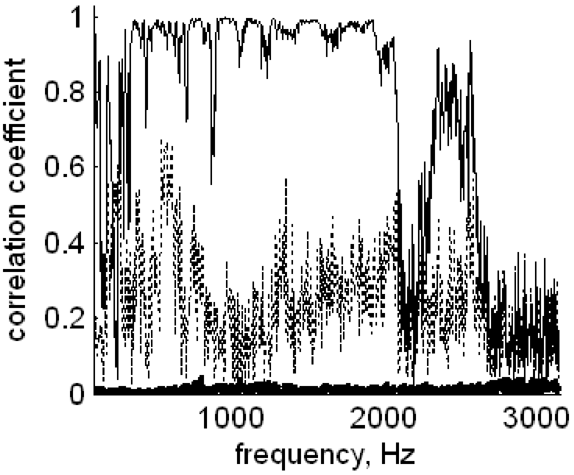

Fig. 6. Absolute value of correlation coefficient of emitted and estimated signals in dependence on frequency of subband with the width of 3.8 Hz. The results are presented for three methods: delay-and-sum beamformer (dotted line), MVDR beamformer (thick solid line) and the proposed algorithm (solid line). The signal length is 15 s.

Figure 6 shows the correlation coefficient of estimated signal with emitted desired signal in dependence on central frequency of subband. The results are presented for the three methods: delay-and-sum beamformer (dotted line), Capon method (thick solid line) and the proposed algorithm (solid line). It follows from Figure 6 that MVDR method degrades severely because of the steering vector errors that took place in the experiment. The delay-and-sum beamformer also does not give a satisfactory result because of the existence of powerful interference sources. On the contrary, the proposed method enables to estimate signal source in presence of powerful interferences. The resulting correlation coefficient of estimated signal and emitted signal equals 0.92 demonstrating the high quality of sound estimation. This result was also confirmed by high speech intelligibility of the estimated signal.

REFERENCES

[1] U. Michel., "History of acoustic beamforming", in *Proc. of the Berlin Beamforming Conf.*, Berlin, 2006, pp. 1-17.

[2] H. Wal, P. Sijtsma, "Source Localization Techniques with Acoustic Arrays", in *NAG/DAGA*, 2009, pp. 1043-1046.

[3] A.B. Gershman, “Robust adaptive beamforming in sensor arrays”, *Int. J. Electronics and Communications*, vol. 53, no 3, pp. 305–314, 1999.

[4] Z. Wang, J. Li, P. Stoica et. al. “Constant-beamwidth and constant-powerwidth wideband robust Capon beamformers for acoustic imaging”, *J. of the Acoust. Soc. of America*, vol. 116, no. 3, pp. 1621-2631, 2004.

[5] Ed. J. Li, P. Stoica, *Robust Adaptive Beamforming,* Hoboken, NJ, USA: John Wiley & Sons, Inc., 2005.

[6] J. Li, P. Stoica, Z. Wang, “On robust Capon beamforming and diagonal loading”, *IEEE Transactions on Sig. Proc.*, vol. 51, no. 7, pp. 1702–1715, 2003.

[7] S.A. Vorobyov, A.B. Gershman, Z.-Q. Luo., “Robust adaptive beamforming using worst-case performance optimization: A solution to the signal mismatch problem”, *IEEE Trans. Signal Processing*, vol. 51, no. 2, pp. 313–324, 2003.

[8] V.I. Turchin, A.A. Rodionov, "Array signal processing based on interference model with incomplete correlation matrix", *Proc. of the ICATT 2013*, Odessa, Ukraine, 2013, pp. 249-251.

[9] S.M. Kay, *Fundamentals of Statistical Signal Processing. V. I. Estimation Theory*, Prentice-Hall PTR, 1998.

[10] Van Trees H.L. *Detection, Estimation, and Modulation Theory, Part IV*, Optimum Array Processing, N.Y, Wiley, 2002.

[11] J.B. Allen, “Image method for efficiently simulating small-room acoustics”, *The J. of the Acoust. Soc. of America*, vol. 65, no. 4, pp. 943-950, 1979.