ADAPTIVE MICROPHONE ARRAY BASED ON MAXIMUM LIKELIHOOD CRITERION

Zoran Šarić[1], Slobodan Jovičić[2], Srbijanka Turajlić[2]

[1]Institute of Security, Kraljice Ane bb, 11000 Belgrade
[2]ETF, Kralja Aleksandra 73, 11000 Belgrade
e-mail: sare@yubc.net

ABSTRACT

The Minimum Variance (MV) criterion is widely used for weight vector estimation of the adaptive microphone array (AMA). The drawback of this criterion is the cancellation of the desired speech signal and its degradation when the microphone array is in a room with reverberation. Applying the Maximum Likelihood (ML) instead of MV criterion has two benefits. The first is the cancellation of interference and the second is the desired speech enhancement. Applying the ML criterion calls for the estimation of the signal and the interference covariance matrices. Both matrices can be estimated from the available microphone signals using the pause detection algorithm based on signal to noise ratio. The proposed speech enhancement algorithm was evaluated by simulating a room with reverberation. Experiments showed the superiority of this algorithm compared to MV based algorithms.

1. INTRODUCTION

The problem of high quality speech recording in a room with reverberation and the cocktail-party interference has long been under consideration. It has been established that microphone arrays, compared to a single microphone, render a better quality of speech recording. The reason being that they usually adapt their beam pattern so to maximally suppress interferences, while maintaining the unity gain for the desired speech signal. The minimum variance (MV) criterion is commonly used to estimate the weight vector. The typical adaptive structures of the microphone array are Frost's adaptive beamformer [1] and the Generalized Sidelobe Canceller (GSC) [2].

The common drawback of MV based adaptive microphone arrays is their sensitivity to room reverberation which causes the cancellation of the desired speech signal. A quantitative analysis of the desired signal cancellation can be found in [3].

Cancellation of the desired speech signal can be prevented if the weight vector is estimated within pauses of the desired speech signal [3], [4], [5]. The estimation algorithm that includes pause detection is presented in [6].

The results presented in this paper will demonstrate that the quality of the restored speech signal can be further improved by applying the maximum likelihood criterion which maximizes the desired signal power, while minimising the interference power. Applying the ML criterion calls for the estimation of the two covariance matrices. The first covariance matrix, denoted as *interference matrix*, is estimated within the pauses of the desired speech signal. The second, *signal matrix*, is estimated when the signal-to-interference ratio is high. Time intervals of both high signal power and pause, are detected by the pause detector based on signal to interference ratio.

The weight vector is estimated by solving the generalized eigenvalue problem for the signal and interference matrix pair [7]. In addition, in order to prevent a certain phase distortion, the appropriate phase correction is introduced.

The proposed estimation algorithm is experimentally verified by simulating the room with reverberation. The quality of the restored speech signal is evaluated by the cepstral distance measure. The comparison to other MV based algorithms proved the superiority of the proposed algorithm.

2. WEIGHT VECTOR ESTIMATION

The microphone signals are processed in the DFT domain. All signals are represented by complex DFT coefficients with central frequency f. For the sake of simplicity the index f will be omitted, i.e. $x = x(f)$. Every DFT bin is processed independently.

Let us assume there is an n microphone array in a room with reverberation. Column vector $\mathbf{X}$ of the n microphone signals can be expressed by

$$\mathbf{X} = \mathbf{S} + \mathbf{U},$$

where the vector $\mathbf{S}$, $\mathbf{S} = \mathbf{h}_1 s_1$, is the room response to the desired signal s_1, with $\mathbf{h}_1$ being the vector of transfer functions from the desired source s_1 to each of the microphones. The vector of the transfer functions $\mathbf{h}_1$ describes both the direct wave path and the reflections from the walls. Vector $\mathbf{U}$ is the room response to all acoustic interferences.

Microphone signals are processed by the adaptive algorithm depicted in fig. 1. The estimation of the desired signal $\hat{s}_1$ is the weighted sum of the microphone signals expressed as

$$\hat{s}_1 = \mathbf{W}^* \mathbf{X}, \qquad (1)$$

where **W** is the weight vector, while superscript * denotes a complex conjugate transpose.

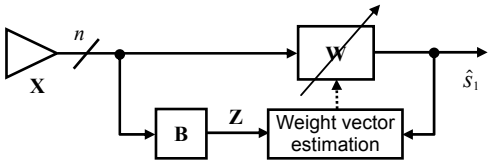

Figure 1. The adaptive beamformer.

The power of the desired signal and the interference power at the beamformer's output are expressed respectively as,

$$P_{\mathrm{s}} = \mathbf{W}^{*}\mathbf{\Phi}_{\mathrm{s}}\mathbf{W},$$

and

$$P_{\mathrm{u}} = \mathbf{W}^{*}\mathbf{\Phi}_{\mathrm{u}}\mathbf{W},$$

where $\mathbf{\Phi}_{\mathrm{s}} = E\{\mathbf{S}^{*}\mathbf{S}\}$ is the *signal matrix* and $\mathbf{\Phi}_{\mathrm{u}} = E\{\mathbf{U}^{*}\mathbf{U}\}$ is the *interference matrix*. The adaptive algorithm should adjust **W** so as to maximize the desired signal power at the output, while minimizing the output interference power. A possible criterion function, which encompasses these two goals, can be formulated as

$$\xi_0 = \frac{P_s}{P_u} = \frac{\mathbf{W}^{*}\mathbf{\Phi}_s\mathbf{W}}{\mathbf{W}^{*}\mathbf{\Phi}_u\mathbf{W}}. \qquad (2)$$

However, estimating **W** by maximizing ξ_0 might not be satisfactory for the following reasons:

a) The estimation error of the $\mathbf{\Phi}_{\mathrm{u}}$ might yield the poorly conditioned solution. This problem can be avoided by adding the term $\beta\mathbf{I}$ to the estimated matrix $\mathbf{\Phi}_{\mathrm{u}}$, $\overline{\mathbf{\Phi}}_{\mathrm{u}}=\beta\mathbf{I}+\mathbf{\Phi}_{\mathrm{u}}$, where **I** is the unit matrix and β, $\beta > 0$, is a scalar by which a compromise between stability and high interference suppression can be achieved [7].

b) The power of the desired signal P_s contains the power of the direct wave and the reflections from the wall

$$P_s = \mathbf{W}^{*}\mathbf{\Phi}_s\mathbf{W} = \mathbf{W}^{*}\mathbf{h}_d^{*}\sigma_s^2\mathbf{h}_d\mathbf{W} + \mathbf{W}^{*}\mathbf{h}_r^{*}\sigma_s^2\mathbf{h}_r\mathbf{W} \qquad (3)$$

where $\mathbf{h}_d$ and $\mathbf{h}_r$ are respectively *n*-column vectors of transfer functions of the direct wave from the desired source to each microphone, and the reflections from the wall. The value σ_s^2 in (3) is the variance of the desired signal.

While the criterion (2) will equally treat both terms, it is necessary to enhance the direct wave. This can be achieved by adding the term $\alpha\,\mathbf{h}_d^{*}\mathbf{h}_d$ to the estimated signal matrix $\overline{\mathbf{\Phi}}_s = \mathbf{\Phi}_s + \alpha\mathbf{h}_d^{*}\mathbf{h}_d$, with $\mathbf{h}_d$ being

$$\mathbf{h}_d = \left[1 \quad e^{-j2\pi f\tau} \quad \cdots \quad e^{-j2\pi f(n-1)\tau}\right]^{*}, \quad \tau = \frac{d\,sin(\theta)}{c},$$

where θ is the arriving angle of the desired signal, *f* is the central frequency of the DFT bin and *c* is the sound velocity. The scalar α, $\alpha > 0$ compromises between the enhancement of the direct wave versus the reflections.

It should be pointed out that the added term, which can be exactly calculated, also contributes towards the reduction of the effects of the estimation error in the *signal matrix* $\mathbf{\Phi}_{\mathrm{s}}$. Summing up all the arguments, the new criterion ξ_1 becomes

$$\xi_1 = \frac{\mathbf{W}^{*}\left(\mathbf{\Phi}_s + \alpha\mathbf{h}_d\mathbf{h}_d^{*}\right)\mathbf{W}}{\mathbf{W}^{*}(\mathbf{\Phi}_u + \beta\mathbf{I})\mathbf{W}},$$

The maximization of the ξ_1 is known as the generalized eigenvalue problem involving the numerator and denominator matrix pair. The solution is an eigenvector corresponding to the largest eigenvalue of the generalized eigenvalue problem

$$\left(\mathbf{\Phi}_s + \alpha\mathbf{h}_d\mathbf{h}_d^{*}\right)\mathbf{W} = \lambda(\mathbf{\Phi}_u + \beta\mathbf{I})\mathbf{W}\,.$$

Matrices $\mathbf{\Phi}_{\mathrm{s}}$ and $\mathbf{\Phi}_{\mathrm{u}}$ have to be estimated from the available data sequence **X**. The interference matrix $\mathbf{\Phi}_{\mathrm{u}}$, can be estimated within pauses of the desired speech using appropriate pause detector. Since the interference is almost always present, estimation of the matrix $\mathbf{\Phi}_{\mathrm{s}}$ is somewhat more difficult. A reasonable approach might be to select time intervals where the signal-to-interference ratio is high and use these intervals for estimating the $\mathbf{\Phi}_{\mathrm{s}}$.

3. PAUSE DETECTION

Pause detection is based on the estimation of the signal to interference ratio (SIR). The best available estimation of the desired signal s_1 is the output of the beamformer (1). Taking into account the strong correlation of the neighbouring frequency bins of the speech signal, the speech signal power at central frequency *f* is estimated by

$$\overline{p}_e(f) = \frac{1}{2\Delta+1}\sum_{i=f-\Delta}^{f+\Delta}\left|\hat{s}_1(i)\right|^2,$$

where Δ defines the length of the frequency smoothing window. The interference power can be obtained by taking the average power of the output of the blocking matrix **B**, arranged in vector **Z** (see fig. 1) and applying

$$\overline{p}_z(f) = \frac{1}{(n-1)(2\Delta+1)}\sum_{f=f-\Delta}^{f+\Delta}\mathbf{Z}^{*}(f)\mathbf{Z}(f),$$

$$\mathbf{Z} = \mathbf{B}^{*}\mathbf{X}, \quad \mathbf{B}^{*} = \begin{bmatrix} 1 & -1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 1 & -1 \end{bmatrix}$$

to smooth in the frequency domain. Estimation of the signal to interference ratio at frequency bin *f*, is thus obtained as

$$l(f) = 10 * \log\left(\frac{\overline{p}_e(f)}{\overline{p}_z(f)}\right),$$

Pause detection is performed in every frequency bin independently. The decision rule can be expressed by

$$l(f) \underset{D_{H_0}}{\overset{D_{H_1}}{\gtrless}} \lambda_{\mathrm{p}}(f),$$

where D_{H_0} and D_{H_1} are decisions of the selecting hypothesis H_0 (pause in s_1) and H_1 (speech in s_1) respectively, and $\lambda_p(f)$ is frequency dependent threshold. It is reasonable to adjust threshold $\lambda_p(f)$ in such a way that the probability of selecting D_{H_0}, is equal to the probability of hypothesis H_0, ($p(H_0)$) expressed by

$$\lambda_p(f) = \lambda \,|\, \left(p_\lambda(D_{H_0}) = q \right),$$

where $p_\lambda(D_{H_0})$ is the probability of selecting D_{H_0} with a threshold λ and q is the assumed value of $p(H_0)$. The probability of selecting D_{H_0}, is calculated by

$$p_\lambda(D_{H_0}) = \frac{1}{N} \sum_{l=l_{\min}}^{\lambda} hist(l) \ ,$$

where $hist(l)$ is histogram of $l(f)$ calculated on the last N blocks of data. Since the value of $p(H_0)$ is not known, it is replaced by an assumed value ($q \approx p(H_0)$) that meets two requirements: a high probability of pause detection, and a low probability of false pause detection. The recursive estimation of the interference matrix $\mathbf{\Phi}_u$ is expressed by

$$\Phi_u(f,t) = \Phi_u(f,t-1) - \mu L_p(f,t) X(f,t) X^*(f,t)$$

with $L_p(f,t)$ representing a soft decision between hypothesis H_0 and H_1 expressed by

$$L_p(f,t) = \frac{1}{e^{\beta(l(f,t)-\lambda_P(f))}+1},$$

where β is an experimentally determined positive constant. Estimation of the signal matrix $\mathbf{\Phi}_s$ have to be done when the signal to interference ratio is high. The soft decision function of the high signal to interference ratio is expressed by

$$L_s(f,t) = \frac{1}{e^{\beta(\lambda_s(f)-l(f,t))}+1},$$

where $\lambda_s(f)$ is a frequency dependent threshold determined in the same way as the threshold $\lambda_p(f)$. The signal matrix is recursively estimated by

$$\Phi_s(f,t) = \Phi_s(f,t-1) - \mu L_s(f,t) X(f,t) X^*(f,t)$$

4. PHASE SHIFT COMPENSATION

The criterion function ξ_1 is not sensitive to the phase shift in the vector $\mathbf{W}$. Namely, if the $\mathbf{W}_m$ is the solution of the maximization of the ξ_1, then the phase shifted vector $\tilde{\mathbf{W}}_m$, $\tilde{\mathbf{W}}_m = \exp(-j\varphi)\mathbf{W}_m$ is also the solution. The shift angle φ is a random variable in every frequency bin. In order to prevent additional speech degradation, which sounds like reverberation, the phase correction has to be applied so as to obtain the zero phase shift angle for the direct wave. The phase corrected weight vector $\overline{\mathbf{W}}$ can be defined as

$$\overline{\mathbf{W}} = \frac{\mathbf{W}^*\mathbf{C}}{\left|\mathbf{W}^*\mathbf{C}\right|}\mathbf{W}, \quad \mathbf{C} \equiv \mathbf{h}_d \, .$$

5. EXPERIMENTAL RESULTS

The proposed algorithm has been verified in a room with reverberation simulated by Allen's image method [8]. The reverberation time was T_{60}=270ms. The number of sources was 2. Source s_1 was the desired speaker while s_2 was the interference (Fig. 2). The microphone array consisted of 8 microphones 6cm apart. The sampling rate of the speech signals was 10KHz. The duration of the each test signal was 10s. To obtain high interference suppression in a room with long reverberation time, a DFT with 4096 points was used.

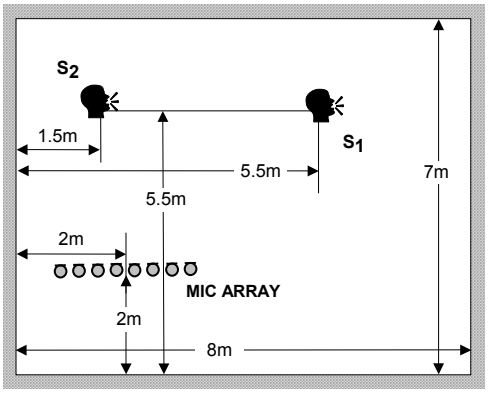

Figure 2. Experimental setup: simulated room with a reverberation time of 270ms and an 8 microphone array.

The following algorithms were compared:

1) The conventional beamformer (CBF),
2) Ordinary GSC with full adaptation,
3) GSC with weights estimated within hand labelled pause intervals [3],
4) GSC with weights estimated under an ideal scenario where only interference is present [3],
5) GEVBF algorithm with covariance matrices $\mathbf{\Phi}_s$ and $\mathbf{\Phi}_u$ estimated within time intervals detected by proposed pause detection algorithm.
6) The proposed generalized eigenvector based beamformer (GEVBF) with covariance matrices $\mathbf{\Phi}_s$ and $\mathbf{\Phi}_u$ estimated within hand labelled time intervals of speech and pause respectively.
7) The proposed GEVBF algorithm with covariance matrices $\mathbf{\Phi}_s$ and $\mathbf{\Phi}_u$ estimated under an ideal scenario where either a speech signal or interference is exclusively present.

The signal s_1 restored by different algorithms is depicted in Fig. 3. The quality of the speech signal restoration was evaluated by the cepstral distortion measure, and the results are presented in Table 1. As was expected, the worst result is obtained by the CBF algorithm. A better result is obtained by the full adaptation GSC, but the restored signal is obviously degraded due to signal cancellation. Further improvement is obtained by GSC weights estimated within the hand labelled pauses (Table 1, row 3). It should be pointed out that the best achievable quality by the MV criterion is under the ideal scenario where the desired signal

is muted and only interference is present (Table 1, row 4). Yet even this result is surpassed by applying the proposed generalized eigenvalue based algorithm (GEVBF), with covariance matrices $\mathbf{\Phi}_s$ and $\mathbf{\Phi}_u$ estimated within hand labelled time intervals of high speech power and pause in speech, respectively. The proposed pause detection algorithm reveals a slight deterioration with respect to the hand labelled time intervals, but is still somewhat superior even to the ideal GSC. Naturally, the best quality is obtained by applying the proposed GEVBF algorithm with covariance matrices $\mathbf{\Phi}_s$ and $\mathbf{\Phi}_u$ estimated under an ideal scenario where either a speech signal or interference is exclusively present.

5. CONCLUSIONS

In this paper, a new algorithm for cocktail party interference suppression is proposed. This algorithm is based on the maximum likelihood criterion that maximizes the signal-to-interference ratio. The beamformer weights are estimated by solving a generalized eigenvalue problem involving signal and interference covariance matrices.

The quality of the signal estimated by the proposed GEVBF algorithm is superior to algorithms based on the MV criterion. The signal and interference covariance matrices are estimated from microphone signals applying the proposed pause detection algorithm.

6. ACKNOWLEDGEMENTS

This work was supported in part by the Ministry of Science and Environment Protection of Republic Serbia under Grant number OI-1784.

REFERENCES

[1] Frost, O.L. “An algorithm for linearly constrained adaptive array processing”, Proceedings of IEEE, vol. 60, no.8, pp. 926-935, Aug. 1972.

[2] L. J. Griffiths, C. W. Jim, “An alternative approach to linearly constrained adaptive beamforming”, *IEEE Trans. Antennas Propagat.,*. vol. AP-30, pp.27-34, Jan.1982.

[3] Zoran Šarić, Slobodan Jovičić, “Adaptive Beamforming in Room with Reverberation”, *Eurospeech 2003, September 1-4 2003, Geneva, Switzerland,* pp 529-532.

[4] O. Hoshuyama et al. "A realtime robust adaptive microphone array controlled by an SNR estimate," Proc. ICASSP98, pp. 3605-3608.

[5] Zoran M. Saric, Slobodan T. Jovicic, “Adaptive microphone array based on pause detection”, *Acoustics Research Letters Online (ARLO) 5(2)*, pp 68-74 April 2004.

[6] Zoran Šarić, Slobodan Jovičić, “Subband pause in speech signal detection using microphone array in room with reverberation”, *Proceedings of the* SPECOM 2004, 20-22 October, 2004, St. Petersburg, pp 132-137.

[7] Dennis R. Morgan, "Adaptive algorithms for solving generalized eigenvalue signal enhancement problem", *Signal Processing 84*, pp 957-968, 2004.

[8] Jont B. Allen, David A. Berkley, "Image method for efficiently simulating small-room acoustics", *J. Acoust. Soc. Amer.* Vol.65, no.4, pp 943-950, Apr.1979.

Table 1: Cepstral distortion measures.

Estimation method	Cepstral distortion measure
1. CBF	0.860
2. GSC	0.758
3. GSC-hand labelled pauses	0.607
4. GSC ideal scenario	0.524
5. GEVBF with proposed pause detection	0.519
6. GEVBF-hand labelled speech/pause intervals	0.479
7. GEVBF- ideal scenario	0.453

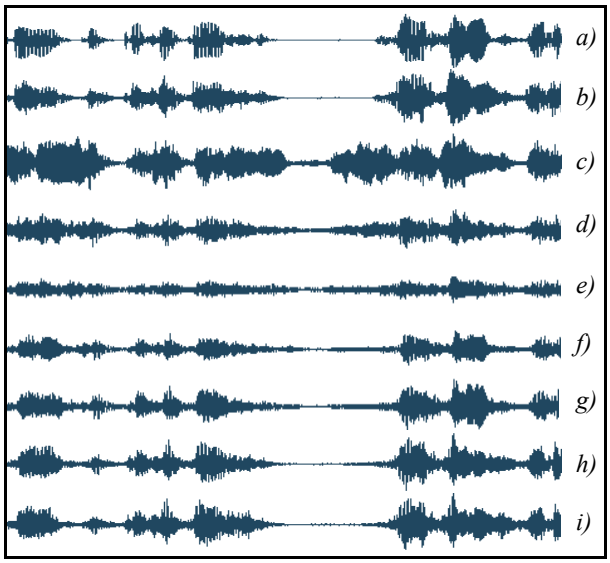

Figure 3. Examples of the time diagrams: *a)* Original signal s_1, *b)* room response of the signal s_1 on microphone 1, *c)* compound of the desired signal s_1 and interference s_2 recorded on microphone 1, *d)* s_1 restored by CBF, *e)* s_1 restored by GSC with full adaptation, *f)* s_1 restored by GSC weights estimated within hand-labelled pauses, *g)* s_1 restored by proposed GEVBF algorithm with proposed pause detection algorithm, *h)* s_1 restored by proposed GEVBF algorithm with covariance matrices $\mathbf{\Phi}_s$ and $\mathbf{\Phi}_u$ estimated within hand labelled time intervals of speech and pause respectively, *i)* proposed GEVBF algorithm with covariance matrices $\mathbf{\Phi}_s$ and $\mathbf{\Phi}_u$ estimated under an ideal scenario where either a speech signal or interference is exclusively present.