interspeech 2008 incorporating SST 08

An Effective Microphone Array Post-filter in Arbitrary Environments

Ning Cheng, Wen-ju Liu, Peng Li, and Bo Xu

National Laboratory of Pattern Recognition,Institute of Automation, Chinese Academy of Sciences
{ncheng,lwj}@nlpr.ia.ac.cn, {pengli,boxu}@hitic.ia.ac.cn

Abstract

The theoretic foundation of traditional microphone array post-filters is the signal model in which the noise between sensors is assumed to be uncorrelated. However, this model is inaccurate in real environments since the correlated noise exists. In this paper, a more generalized signal model which considers both the correlated and uncorrelated noise is introduced. A general expression of the microphone array post-filter is proposed for this model. For better residual noise shaping, the human auditory property is incorporated into the post-filter estimation process. In experiments with real noise microphone array recordings, the proposed technique has shown to produce impressive results in terms of quality measures of the enhanced speech.

Index Terms: post-filter, generalized signal model, human auditory property, speech enhancement

1. Introduction

The problem of using microphone arrays for the task of speech enhancement has received much attention in recent years. So far, a variety of speech enhancement algorithms based on microphone arrays have been proposed [1]-[5]. A recently well studied technique is the post-filter algorithm due to its good noise reduction performance. The commonly used multichannel post-filter, which is based on the Wiener filter, was first introduced by Zelinski [1]. Based on the work of Zelinski, Marro et al. [2] suggested using the auto- and cross-power spectrums of the array inputs to estimate the post-filter transfer function. In this paper, this technique is referred to as the Zelinski post-filter. McCowan [3] provides a more general expression of the post-filter estimation based on a known noise field coherence function.

One problem of the traditional post-filter technique (the Zelinski post-filter) is that it is based on the signal model in which the noise on different channels is assumed to be uncorrelated. In another word, the Zelinski post-filter just considers the uncorrelated noise. However, in real environments, not only the uncorrelated noise exists but also the correlated noise exists.

In this paper, to deal with the problem of suppressing noise in arbitrary environments, we first propose a generalized post-filter based on a generalized signal model which considers both the correlated and uncorrelated noise. Then, the human auditory property has been incorporated into the post-filter estimation process for better residual noise shaping.

The remainder of this paper is organized as follows. Section 2 gives the generalized signal model and the corresponding generalized post-filter. In section 3, the improvement using the human auditory property is presented. In section 4, the performance of the proposed technique is accessed in experiments with microphone array noise recordings. At last, a conclusion is given in section 5.

2. Post-filter based on a generalized signal model

In Figure 1, a linearly and equidistantly distributed microphone array in a noisy environment is considered. A generalized signal model is assumed in which the observed signals consist of three components. The first is the target speech signal coming from a direction. The second is the localized noise arriving from another direction and the third is the non-localized noise, propagating in all directions simultaneously. Obviously, the localized noise is correlated between sensors and the non-localized noise is assumed to be uncorrelated between sensors.

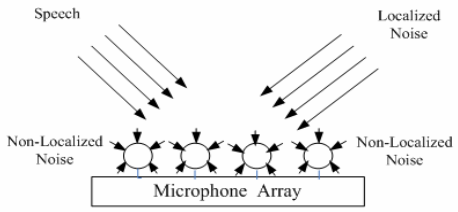


Figure 1: *The signals imposing on the microphone array.*

In Figure 1, the observed signal imposing on the microphone array can be given in the frequency domain as:

$$X = S \cdot d + N \quad (1)$$

$$N = M + V \quad (2)$$

where $X = [X_1, \cdots, X_L]^T$ is the noisy signal vector received by the microphone array, S is the target signal, $d = [d_1, \cdots, d_L]^T$ is the propagation vector of the signal source , $N = [N_1, \cdots, N_L]^T$ is the noise vector, $M = [M_1, \cdots, M_L]^T$ is the localized noise vector , $V = [V_1, \cdots, V_L]^T$ is the non-localized noise vector and L is the number of sensors.

Simmer et al. [4] give the demonstration of expressing the optimal broadband Minimum Mean Square Error (MMSE) filter solution as a classical Minimum Variance Distortionless Response (MVDR) beamformer followed by a single-channel Wiener filter, which is:

$$w_{opt} = \left[\frac{\phi_{SS}}{\phi_{SS} + \phi_{NN}} \right] \frac{\Phi_{NN}^{-1} d}{d^H \Phi_{NN}^{-1} d} \quad (3)$$

where w_{opt} is the optimal filter coefficients vector, ϕ_{SS} and ϕ_{NN} are respectively the (single-channel) target signal and noise auto-power spectrum vectors, d is the propagation vector of the signal source and Φ_{NN} is the (multichannel) noise cross-spectral density matrix.

The bracketed item in the expression (3) is the single-channel Wiener filter part and the remaining item is the well known solution for the MVDR beamformer [5].

According to (3), a multichannel speech enhancement system is constructed as shown in Figure 2, which mainly consists of three parts: the MVDR beamformer to maximize the directivity of the array response, the Wiener post-filter estimator to estimate the post-filter transfer function and the post-filtering part to further enhance the beamformer output.

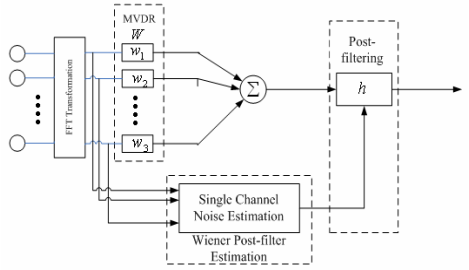


Figure 2: *Diagram of the multichannel speech enhancement system.*

In this paper, we focus on solving the problem of estimating the post-filter term in the expression (3) which is:

$$h = \left[\frac{\phi_{SS}}{\phi_{SS} + \phi_{NN}} \right]. \tag{4}$$

Under the noise field assumptions that:

1) The target signal and noise are uncorrelated.
2) The noise power spectrum is the same on all sensors.
3) The noise is uncorrelated between sensors.

The Zelinski post-filter is given as follows:

$$\hat{h} = \frac{\frac{2}{L(L-1)} \sum_{i=1}^{L-1} \sum_{j=i+1}^{L} \Re\{\phi_{X_i X_j}\}}{\frac{1}{L} \sum_{i=1}^{L} \phi_{X_i X_i}} \tag{5}$$

where $\Re\{\bullet\}$ is the real operator and L is the number of the microphone array sensors.

However, above assumptions are inaccurate in real environments since the localized noise is correlated between sensors.

Considering the practical situations, following assumptions are adopted for our generalized signal model:

1) The target speech signal, the localized noise and the non-localized noise are uncorrelated with each other ($\phi_{S_i M_j} = 0$, $\phi_{S_i V_j} = 0$, $\phi_{M_i V_j} = 0$, $\forall i, j$).
2) The noise power spectrum is the same on all sensors ($\phi_{M_i M_i} = \phi_{MM}, \phi_{V_i V_i} = \phi_{VV}, \forall i$).
3) The localized noise is correlated between sensors $(\phi_{M_i M_j} = \phi_{MM}, \forall i, j)$ and the non-localized noise is uncorrelated between sensors $(\phi_{V_i V_j} = 0, \forall i \neq j)$.

Under these assumptions, the post-filter term (4) can be rewritten as:

$$h = \frac{\phi_{SS}}{\phi_{SS} + \phi_{MM} + \phi_{VV}}. \tag{6}$$

Calculating the auto- and cross-power spectrums of the aligned signals on channels i and j, leads to:

$$\begin{aligned} \phi_{X_i X_i} &= \phi_{S_i S_i} + \phi_{M_i M_i} + \phi_{V_i V_i} + 2\Re\{\phi_{S_i M_i} + \phi_{S_i V_i} + \phi_{M_i V_i}\} \\ &= \phi_{SS} + \phi_{MM} + \phi_{VV} \end{aligned} \tag{7}$$

$$\begin{aligned} \phi_{X_i X_j} &= \phi_{S_i S_j} + \phi_{M_i M_j} + \phi_{V_i V_j} \\ &\quad + \phi_{S_i M_j} + \phi_{S_i V_j} + \phi_{M_i S_j} + \phi_{M_i V_j} + \phi_{V_i S_j} + \phi_{V_i M_j} \\ &= \phi_{SS} + \phi_{MM}. \end{aligned} \tag{8}$$

Obviously, the expression (5) is not the accurate estimation of the expression (6) because under the adopted assumptions, $\frac{2}{L(L-1)} \sum_{i=1}^{L-1} \sum_{j=i+1}^{L} \Re\{\phi_{X_i X_j}\}$ is not the estimate of ϕ_{SS}, but the estimate of $\phi_{SS} + \phi_{MM}$. An accurate expression of ϕ_{SS} is needed to estimate the expression (6). According to (7) and (8), two estimates are given as follows:

$$\phi_{SS} + \phi_{MM} = \frac{2}{L(L-1)} \sum_{i=1}^{L-1} \sum_{j=i+1}^{L} \Re\{\phi_{X_i X_j}\} \tag{9}$$

$$\phi_{SS} + \phi_{MM} + \phi_{VV} = \frac{1}{L} \sum_{i=1}^{L} \phi_{X_i X_i}. \tag{10}$$

ϕ_{SS} can be obtained if the noise power spectrum $\phi_{MM} + \phi_{VV}$ is available. We estimate the noise in each single channel. The computation of the auto- and cross-power spectrums of the noise on channels i and j, results to:

$$\begin{aligned} \phi_{N_i N_i} &= \phi_{M_i M_i} + \phi_{V_i V_i} + 2\Re\{\phi_{M_i V_i}\} \\ &= \phi_{MM} + \phi_{VV} \end{aligned} \tag{11}$$

$$\begin{aligned} \phi_{N_i N_j} &= \phi_{M_i M_j} + \phi_{M_i V_j} + \phi_{V_i M_j} + \phi_{V_i V_j} \\ &= \phi_{MM}. \end{aligned} \tag{12}$$

According to (11), $\phi_{MM} + \phi_{VV}$ can be estimated as follows:

$$\phi_{MM} + \phi_{VV} = \frac{1}{L} \sum_{i=1}^{L} \phi_{N_i N_i}. \tag{13}$$

Combining (10) and (13), we have:

$$\phi_{SS} = \frac{1}{L} \sum_{i=1}^{L} \phi_{X_i X_i} - \frac{1}{L} \sum_{i=1}^{L} \phi_{N_i N_i}. \tag{14}$$

According to (14), an estimate of the expression (6) is obtained as follows:

$$\hat{h} = \frac{\phi_{SS}}{\phi_{SS} + \phi_{MM} + \phi_{VV}} = \frac{\frac{1}{L} \sum_{i=1}^{L} \phi_{X_i X_i} - \frac{1}{L} \sum_{i=1}^{L} \phi_{N_i N_i}}{\frac{1}{L} \sum_{i=1}^{L} \phi_{X_i X_i}}. \tag{15}$$

It should be noted that, if the noise estimation in the single channel is not accurate, the value of $\hat{h}$ may be out of its theoretic range $[0,1]$, caused by the subtraction operation in (15). A simple approach to solve this problem is to bound the $\hat{h}$ estimation as follows:

$$\hat{h} = \begin{cases} 1, & if \quad \hat{h} > 1 \\ 0.1, & if \quad \hat{h} < 0.1 \end{cases} \tag{16}$$

where 0.1 is determined empirically.

3. An improvement by the human auditory property

A potentially important development in noise reduction methods is the incorporation of the psychoacoustic properties of human hearing, namely the so called masking. The masking phenomenon can be explained by the so called critical bands. Within one critical band, one sound (the maskee) becomes inaudible in the presence of another sound (the masker) with a higher intensity. The human auditory frequency range spreads from 0 to 15500 Hz and covers approximately 24 critical bands [6]. In this section, we incorporate this property into the post-filter estimation process to shape the residual noise.

The expression (15) can be expressed as follows:

$$\hat{h} = \frac{\frac{1}{L} tr(\Psi_{SS})}{\frac{1}{L} tr(\Psi_{XX})} = \frac{\frac{1}{L} tr(\Psi_{XX} - \Psi_{NN})}{\frac{1}{L} tr(\Psi_{XX})} \quad (17)$$

where $\Psi_{XX} = diag(\phi_{X_1X_1}, \cdots, \phi_{X_LX_L})$, $\Psi_{NN} = diag(\phi_{N_1N_1}, \cdots, \phi_{N_LN_L})$ and $tr(\bullet)$ is the trace operator.

The eigenvalue decomposition (EVD) of Ψ_{SS} is given by:

$$\Psi_{SS} = U \Lambda U^H \quad (18)$$

where $\Lambda = diag(\lambda_1, \cdots, \lambda_L)$ with the eigenvalues λ_i's in decreasing order, $\lambda_i = \phi_{X_iX_i} - \phi_{N_iN_i}$, $U = [U_1, \cdots, U_L]$ is the unitary eigenvector matrix and U_i is the unitary eigenvector corresponding to λ_i.

Since the noisy signal energy is always larger than the noise energy, λ_i should not be negative. So, we set the negative λ_i to 0 and give the following expression:

$$\hat{\lambda}_i = \max(\lambda_i, 0). \quad (19)$$

The use of (19) is to reduce the impacts of the channels which may have the noise estimation errors.

To use the human auditory property, we need to get the auditory masking energies.

First, we should compute the excitation pattern $C(k)$ which is regarded as an energy distribution along the basilar membrane. $C(k)$ can be calculated by convolving the subband energy $B(k)$ with the spreading function $SF(k)$ [7].

$$C(k) = SF(k) * B(k) \quad (20)$$

where $k = 1, \cdots K$ and $K = 24$ is the critical band number.

An analytical expression for the spreading function is given by:

$$SF(k) = 15.81 + 7.5(k + 0.474) - 17.5\sqrt{1 + (k + 0.474)^2}\,[dB]. \quad (21)$$

The subband energy $B(k)$ is calculated by:

$$B(k) = \sum_{i=1}^{L} \hat{\lambda}_i(k) \left| U_i(k) \right|^2. \quad (22)$$

Or in matrix notation,

$$B = P\hat{\lambda} \quad (23)$$

where $\hat{\lambda} = [\hat{\lambda}_1(k), \cdots, \hat{\lambda}_L(k)]^T$, $P = [p_1(k), \cdots, p_L(k)]$ and $p_i(k) = \left| U_i(k) \right|^2$.

The masking threshold $C_{thr}(k)$ can be calculated as follows [7]:

$$C_{thr}(k) = 10^{\log_{10}|C(k)| - |O(k)/10|} \quad (24)$$

where $O(k)$ is the offset which can be found in [8].

This perceptual information should be mapped to the eigendomain and the mapping method is given by [9]:

$$\theta(k) = P^T C_{thr}(k) \quad (25)$$

where $\theta(k) = [\theta_1(k), \cdots, \theta_L(k)]^T$ and $\theta_i(k)$ is hereafter referred to as the "masking energies".

Making $\hat{\lambda}_i$ less than the "masking energies", $\hat{\lambda}_i$ becomes smaller and more noise suppression is achieved. So, $\hat{\Lambda}$ is proposed to be calculated as:

$$\hat{\Lambda} = diag(\min(\hat{\lambda}_1, \theta_1(k)), \cdots, \min(\hat{\lambda}_L, \theta_L(k))). \quad (26)$$

Then, we have:

$$\hat{\Psi}_{SS} = U \hat{\Lambda} U^H. \quad (27)$$

The expression (17) is rewritten as follows:

$$\hat{h} = \frac{\frac{1}{L'} tr(\hat{\Psi}_{SS})}{\frac{1}{L} tr(\Psi_{XX})} \quad (28)$$

where L' is the rank of the $\hat{\Psi}_{SS}$.

4. Experiments and analysis

To validate the effectiveness of the proposed technique, we compare its performance to other multichannel noise reduction algorithms, including the Beamformer [4], the Zelinski post-filter [1][2] and the McCowan post-filter [3]. The CMU microphone array database [10] is used for the experiments. The recordings were collected by a linear microphone array with fifteen sensors at a sampling rate of 16 kHz. Since this array is not equally spaced, we choose seven sensors of No.5-No.11 to form a linear and equally spaced sub-array. The space between sensors is 4 cm. The adopted recordings have two types: one is collected in a laboratory and the interference is strong computer noise; the other is collected in a conference room and the noise is made by a talk radio. Totally 26 utterances are used. The time aligned noisy inputs of the array are divided in time into frames of 25ms, 15ms overlapped between adjacent frames. At each frame a Hamming window is applied and a STFT analysis takes place. The single channel noise estimation refers to Hasan [11] and the auto- and cross-power spectrums estimation refers to McCowan [3].

The evaluation criteria we adopted are the segmental signal-to-noise ratio enhancement (SSNRE), the PESQ score , the Log-spectral distance (LSD), Log-area ratio (LAR) and Log-likelihood ratio (LLR) [12][13].

High values of the SSNRE and the PESQ score and low values of the LSD, the LAR and the LLR denote high speech quality.

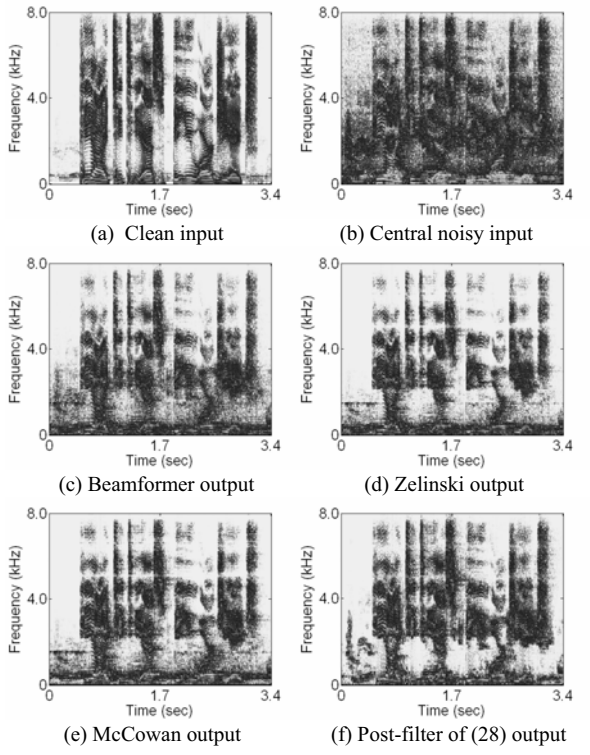
Table 1 displays the average experiment results of the proposed algorithm and the competing algorithms. The "Input" item corresponds to the average value of the microphone array inputs. The "Prop1" and "Prop2" items correspond to the results of the proposed post-filters in the expressions (15) and (28), respectively.

Table 1. *Average experiment results of the proposed algorithm and the competing algorithms.*

	Input	Beamformer	Zelinski	McCowan	Prop1	Prop2
SSNRE(dB):	-	0.02	-0.20	3.25	3.69	4.17
PESQ:	2.26	2.31	2.28	2.33	2.37	2.43
LSD:	7.10	6.73	7.18	6.11	5.79	5.52
LAR:	8.54	9.54	11.62	8.34	7.74	7.48
LLR:	0.81	0.86	1.04	0.94	0.73	0.69

It is easy to find that our post-filter (15) and (28) are better than the competing algorithms under all the five criteria and our post-filter (28) has the best performance in all the test algorithms. Namely the relative % average improvements achieved compared to the best of the reference approaches were 28.3% in SSNRE, 4.3% in PESQ, 9.7% in LSD, 10.3% in LAR and 19.8% in LLR.

Figure 3 shows the spectrograms of the clean speech, the central noisy input and the enhanced results of all the test algorithms for an utterance corresponding to the string of "pittsburgh" for comparison.



(a) Clean input (b) Central noisy input

(c) Beamformer output (d) Zelinski output

(e) McCowan output (f) Post-filter of (28) output

Figure 3: *Signal spectrograms for the utterance of "pittsburgh".*

From Figure 3, we note that the competing algorithms are incapable of removing the noise in the low frequency region. For the MVDR beamformer, this inadequacy is attributed to the fact that the greatest portion of the noise energy is concentrated in the low frequency region, where the beamformer has a low directivity factor. The poor performance of the Zelinski post-filter is due to the reason that this method is based on the assumption of a spatially uncorrelated field. For the McCowan post-filter, the differences between the assumed and actual coherence functions result in its performance significant degradation. Compared with these algorithms, our post-filter reduces more noise at low frequencies and gets better enhanced results at all frequencies since it has a more accurate transfer function estimation and the incorporation of the human auditory property indeed helps the residual noise shaping.

5. Conclusions

In this paper, an effort has been made to develop an effective post-filter to suppress the noise in arbitrary environments. Comparative results have shown that incorporating a more accurate signal model and the human auditory property into the post-filter estimation process result in improved speech enhancement. This intuitively motivates the use of the proposed technique as a general and possible optimum estimation approach.

6. Acknowledgements

This work was supported in part by the China National Nature Science Foundation (No. 60675026，No. 60121302), the 863 China National High Technology Development Projects (No.20060101Z4073, No.2006AA01Z194) and the National Grand Fundamental Research 973 Program of China (No. 2004CB318105).

7. References

[1] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms", in Proc. of ICASSP-88, 1988, Vol. 5, pp. 2578–2581.

[2] C.Claude Marro, Y.Yannick Mahieux, and K. U.K. Uwe Simmer, "Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering", IEEE Trans. Speech Audio Process., Vol. 6, pp. 240–259, May 1998.

[3] Iain A. McCowan, Hervé Bourlard, "Microphone array post-filter based on noise field coherence", IEEE Trans. Speech Audio Process., Vol.11, pp.709-715, Nov. 2003.

[4] K. Uwe Simmer, et al, "Post-filtering techniques", in Microphone Arrays, M. Brandstein and D.Ward, Eds. New York: Springer, ch. 3, pp. 36–60, 2001.

[5] H. Cox, R. Zeskind, and M. Owen, "Robust adaptive beamforming", IEEE Trans. Acoust., Speech, Signal Process., Vol. 35, pp. 1365–1376, Oct. 1987.

[6] E.Zwicker and E.Terhardt, "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency", JASA, Vol.68, no.5, pp.1523-1525, 1980.

[7] R.M.Udrea, N.D.Vizireanu and S.Ciochina, "An Improved Spectral Subtraction Method for Speech Enhancement using a Perceptual Weighting Filter", Digital Signal Processing, 2007, doi:10.1016/j.dsp.2007.08.002.

[8] Virag, N., "Signal Channel Speech Enhancement Based on Masking Properties of the Human Auditory System", IEEE Trans. Speech Audio Process. Vol.7 No. 2, pp.126-137, 1999.

[9] F.Jabloun and B.Champagne, "Incorporating the Human Hearing Properties in the Signal Subspace Approach for Speech Enhancement", IEEE Trans. Speech Audio Process. Vol.11, No.6, pp.700-708, 2003.

[10] Sullivan, T., 1996. CMU microphone array database. http://www.speech.cs.cmu.edu/databases/micarray.

[11] M.Hasan, S.Salahuddin and M.Khan, "A Modified A Priori SNR for Speech Enhancement Using Spectral Subtraction Rules", IEEE Signal Process. Lett., Vol.11, No.4, pp.450-453, 2004.

[12] S. Lefkimmiatis and P. Maragos, "A generalized estimation approach for linear and nonlinear microphone array post-filters", Speech Comm., 49, 657-666, 2007.

[13] K. Manohar and P. Rao, "Speech enhancement in nonstationary noise environments using noise property", Speech Comm., 48, 96-109, 2006