Microphone array speech enhancement based on a generalized post-filter and a novel perceptual filter

Ning Cheng, Wen-Ju Liu, Peng Li, Bo Xu

National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
Beijing 100080, P. R. China
E-mail: ncheng@nlpr.ia.ac.cn

Abstract: The theoretic foundation of traditional microphone array post-filters is the assumption that the noise between sensors is uncorrelated. However, this assumption is inaccurate in real environments since the correlated noise exists. In this paper, a generalized microphone array post-filter is proposed to deal with both the correlated and uncorrelated noise in environments and a novel perceptual filter is proposed to reduce the musical residual noise introduced by the post-filter. Experiments show that the proposed technique produces impressive results in terms of quality measures of the enhanced speech.

Key words: Microphone array, speech enhancement, post-filter, perceptual filter

I. INTRODUCTIONS

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 spectra 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. However, this assumption is inaccurate in real environments since the correlated noise exists. 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 comprehensive signal model including both the correlated and uncorrelated noise. Then, a novel perceptual filter is proposed to reduce the musical residual noise introduced by the post-filter. The proposed method gives a superior performance as compared to the conventional post-filter algorithms.

II. THEORETICAL FRAMEWORK

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.

Fig.1 The signals imposing on the microphone array.

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

\[ X = S \cdot d + N \]  (1)

\[ N = M + V \]  (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:

\[ W_{opt} = \frac{\phi_{SS}}{\phi_{SS} + \phi_{NN}} \Phi_{NN}^{-1} d \]  (3)

where \( W_{opt} \) is the optimal filter coefficients vector, \( \phi_{SS} \) and \( \phi_{NN} \) are respectively the (single-channel) target signal and noise auto-power spectrum vectors, and \( \Phi_{NN} \) 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].

Fig.2 Diagram of the multichannel system.
According to (3), a multichannel speech enhancement system is constructed as shown in Fig.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. In addition, to reduce the musical residual noise made by the post-filter, we add a perceptual filter part in the system.

III. A GENERALIZED POST-FILTER

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

$$\hat{h} = \left[ \begin{array}{c} \phi_{SS} \\ \phi_{SS} + \phi_{NM} \end{array} \right].$$  \hspace{1cm} (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:

$$\phi_{\text{Z}} = \frac{2}{L(L-1)} \sum_{i=1}^{L} \sum_{j \neq i} \Re \{ \phi_{X_iX_j} \}.$$  \hspace{1cm} (5)

The localized noise is correlated between sensors

$$\phi_{NM} = \phi_{M,M_i} \phi_{V_iV_j} + \phi_{M,V_i} \phi_{V_jV_j}.$$  \hspace{1cm} (6)

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

$$\hat{h} = \left[ \begin{array}{c} \phi_{SS} \\ \phi_{SS} + \phi_{NM} \end{array} \right].$$  \hspace{1cm} (4)

Where \( \Re \) 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 comprehensive signal model:
1) The target speech signal, the localized noise and the non-localized noise are uncorrelated with each other \( \phi_{S,M_i} = 0, \phi_{S,V_i} = 0, \phi_{M,V_i} = 0, \forall i, j \).
2) The noise power spectrum is the same on all sensors \( \phi_{M,M_i} = \phi_{M,M_j}, \phi_{V_iV_j} = \phi_{V_jV_j}, \forall i, j \).
3) The localized noise is correlated between sensors \( \phi_{M,M_i} = \phi_{MM}, \forall i, j \) and the non-localized noise is uncorrelated between sensors \( \phi_{V_iV_j} = 0, \forall i \neq j \).

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

$$\hat{h} = \frac{\phi_{SS}}{\phi_{SS} + \phi_{MM} + \phi_{VV}}.$$  \hspace{1cm} (7)

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

$$\phi_{X_iX_i} = \phi_{SS} + \phi_{M,M_i} + \phi_{V_iV_i} + 2 \Re \{ \phi_{S,M_i} + \phi_{S,V_i} + \phi_{M,V_i} \}$$

$$\phi_{X_iX_j} = \phi_{SS} + \phi_{M,M_i} + \phi_{V_iV_j}$$

$$\phi_{X_iX_j} = \phi_{SS} + \phi_{M,M_j} + \phi_{V_jV_j}$$

$$\phi_{X_iX_j} = \phi_{SS} + \phi_{M,M_i} + \phi_{V_iV_j} + \phi_{M,M_j} + \phi_{V_jV_j} + \phi_{M,M_i}$$

$$\phi_{X_iX_j} = \phi_{SS} + \phi_{MM}.$$  \hspace{1cm} (8)

Obviously, the expression (5) is not the accurate estimation of the expression (6) because under the adopted assumptions,

$$\frac{2}{L(L-1)} \sum_{i=1}^{L} \sum_{j \neq i} \Re \{ \phi_{X_iX_j} \}$$

is not the estimate of \( \phi_{SS} \), but the estimate of \( \phi_{SS} + \phi_{MM} \). An accurate expression of \( \phi_{SS} \) is needed to estimate the expression (6). According to (7) and (8), two estimates are given as follows:

$$\phi_{SS} + \phi_{MM} = \frac{2}{L(L-1)} \sum_{i=1}^{L} \sum_{j \neq i} \Re \{ \phi_{X_iX_j} \}.$$  \hspace{1cm} (9)

$$\phi_{SS} + \phi_{MM} + \phi_{VV} = \frac{1}{L} \sum_{i=1}^{L} \phi_{X_iX_i}.$$  \hspace{1cm} (10)

\( \phi_{MS} \) 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:

$$\phi_{N_iN_i} = \phi_{M,M_i} + \phi_{V_iV_i} + 2 \Re \{ \phi_{M,V_i} \}$$

$$\phi_{N_iN_j} = \phi_{M,M_i} + \phi_{M,V_i} + \phi_{M,V_j} + \phi_{V_iV_j}$$

$$\phi_{MM}.$$  \hspace{1cm} (11)

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

$$\phi_{MM} + \phi_{VV} = \frac{1}{L} \sum_{i=1}^{L} \phi_{N_iN_i}.$$  \hspace{1cm} (12)

Combining (10) and (12), we have:

$$\phi_{SS} = \frac{1}{L} \left[ \sum_{i=1}^{L} \phi_{X_iX_i} - \frac{1}{L} \sum_{i=1}^{L} \phi_{N_iN_i} \right].$$  \hspace{1cm} (13)

According to (10) and (14), an estimate of the expression (6) is obtained as follows:

$$\hat{h} = \frac{\phi_{SS}}{\phi_{SS} + \phi_{MM} + \phi_{VV}}$$

$$\phi_{SS} + \phi_{MM} + \phi_{VV} = \frac{1}{L} \left[ \sum_{i=1}^{L} \phi_{X_iX_i} - \frac{1}{L} \sum_{i=1}^{L} \phi_{N_iN_i} \right].$$  \hspace{1cm} (15)

IV. A NOVEL PERCEPTUAL FILTER

While reducing the background noise, the post-filter inevitably introduces some musical residual noise which makes the signal perceptual quality bad. In this section, a novel perceptual filter is proposed to reduce the residual noise. This perceptual filter is based on the human auditory masking phenomenon which can be explained by the so called critical bands. Within one critical band, one sound (the masker) becomes inaudible in the presence of another sound (the maskee) with a higher intensity. The human auditory frequency range spreads from 0 to 15500 Hz and covers approximately 24 critical bands [6].

First, we need to compute the subband energy \( E(k) \):

$$E(k) = \sum_{n=1}^{L} \left| \tilde{S}(e^{jw_n}) \right|^2$$  \hspace{1cm} (16)

Where \( \tilde{S}(e^{jw_n}) \) is the post-filter output, \( b(k) \) is the frequency index depending on the lower and upper frequency boundary of the critical band \( k \), \( k = 1, \cdots, K \) and \( K \) is the critical band number which is decided by the actual frequency range of the data.

An excitation pattern \( C(k) \) can be regarded as an energy
distribution along the basilar membrane. It can be calculated by
convolving the subband energy $E(k)$ with a spreading
function $SF(k)$:
\[ C(k) = SF(k) \ast E(k). \]  
(17)
The analytical expression for the spreading function is given by
[7]:
\[ SF(k) = 15.81 + 7.5(k + 0.474) - 17.5 \sqrt{1 + (k + 0.474)^2}. \]  
(18)
The noise masking threshold (NMT) $T(k)$ is obtained as follows:
\[ T(k) = 10 \log_{10} \left[ \omega(k) \left| \frac{E(k)}{|F(k)|} \right| \right] \]  
(19)
where $O(k)$ is a relative threshold offset indicating whether a
frame is tone-like or noise-like [8].
The estimate of the target signal $\hat{S}(e^{j\omega})$ is expressed as:
\[ \hat{S}(e^{j\omega}) = G(e^{j\omega})\tilde{S}(e^{j\omega}) \]  
(20)
where $G(e^{j\omega})$ is the perceptual gain function and $\tilde{S}(e^{j\omega})$ is
the post-filter output.

The error item is defined as the difference between the clean
signal and the estimated (enhanced) signal, which is:
\[ E(e^{j\omega}) = \hat{S}(e^{j\omega}) - S(e^{j\omega}) = G(e^{j\omega})\tilde{S}(e^{j\omega}) - S(e^{j\omega}) \]  
(21)
where $S(e^{j\omega})$ is the clean signal, $\hat{S}(e^{j\omega})$ is the estimated
signal and $\tilde{N}(e^{j\omega})$ is the musical residual noise.
The first term in equation (21) describes the speech distortion
which can be minimized if the perceptual gain function $G(e^{j\omega}) = 1$. The second term describes the noise
distortion which can be minimized by $G(e^{j\omega}) = 0$. A
perceptual function $G(e^{j\omega})$ can be computed to make the
noise or speech distortions fall below the masking threshold.
This paper chooses the perceptual gain function to minimize the
noise distortion and a variable speech distortion is allowed.
Therefore, the perceptual gain function $G$ is chosen to satisfy the
following criterion:
\[ |G(e^{j\omega})| \leq \left| \tilde{N}(e^{j\omega}) \right|^2 \leq |T| \]  
(22)
where $T$ is the NMT estimated in the expression (19).
An analytical expression of the psychoacoustically motivated
gain function is proposed in the following form:
\[ G(e^{j\omega}) = \frac{T}{\tilde{N}(e^{j\omega})} \]  
(23)

V. EXPERIMENTS AND RESULTS

The CMU microphone array database [9] is used for the
experiments. The recordings were collected in a computer lab
by a linear microphone array with eight sensors spaced 7 cm
apart, at a sampling rate of 16 kHz. The array was placed on a
desk and the speaker was seated directly in front of it at a
distance of 1 m from its center. The room had multiple noise
sources, including several computer fans and overhead air
blowers. The corpus consists of 130 utterances, 10 speakers of
13 utterances each. The time aligned noisy inputs of the array
are divided in time into frames of 25ms with overlap of 15ms
between adjacent frames. At each frame a Hamming window is
applied and a STFT analysis takes place. The critical band
number is $K = 21$. The single channel noise is estimated as in
[8].
The adopted evaluation criteria are the segmental
signal-to-noise ratio enhancement (SSNRE), the Log-spectral
distance (LSD), and the Log-area ratio (LAR) [3][10]. High
values of the SSNRE and low values of the LSD and LAR
denote high speech quality.

<table>
<thead>
<tr>
<th>Input</th>
<th>Beamformer</th>
<th>Zelinski</th>
<th>McCowan</th>
<th>Prop1</th>
<th>Prop2</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSNRE(dB)</td>
<td>-0.33</td>
<td>1.08</td>
<td>5.07</td>
<td>6.90</td>
<td>9.41</td>
</tr>
<tr>
<td>LSD</td>
<td>7.07</td>
<td>6.63</td>
<td>6.78</td>
<td>5.78</td>
<td>5.45</td>
</tr>
<tr>
<td>LAR</td>
<td>8.70</td>
<td>11.24</td>
<td>15.03</td>
<td>11.31</td>
<td>10.08</td>
</tr>
</tbody>
</table>

Table 1 displays the average experiment results. The “Input”
corresponds to the average value of the array inputs. The
“Prop1” and “Prop2” correspond to the outputs of the proposed
post-filter (15) and perceptual filter (23), respectively. Namely
the relative % average improvements achieved compared to the
best of the reference approaches were 85.6% in SSNRE, 9.9% in
LSD and 21.1% in LAR.

Figure 3 shows the spectrograms of the clean speech, the
noisy input and the enhanced speech of all the test algorithms
for an utterance corresponding to the string of “erisckcw1485”
for comparison.

![Fig.3 Spectrograms for the utterance of “erisckcw1485”](image)

From Fig.3, we note that the competing algorithms are incapable
of removing the noise. For the MVDR beamformer, this
inadequacy is attributed to the fact that the beamformer has a
low directivity factor in the low frequency region where the
noise between sensors is significantly correlated. 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, the proposed technique reduces more noise and gets
better enhanced speech since it considers the correlated noise in
the post-filter estimation process and the perceptual filter indeed helps the residual noise shaping.

VI. CONCLUSIONS

This paper has presented an effective microphone array speech enhancement approach. First, a generalized post-filter is formulated to handle a comprehensive noise field including both the correlated and uncorrelated noise. Then, a novel perceptual filter is proposed to reduce the musical residual noise made by the post-filter. Experiment results show that the proposed technique gives significant improvement over the existing array algorithms in terms of objective speech quality measures.

ACKNOWLEDGEMENT

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

REFERENCES