Improved Hybrid Microphone Array Post-filter by Integrating a Robust Speech Absence Probability Estimator for Speech Enhancement

Junfeng Li[1], *Masato Akagi*[2], *and Yôiti Suzuki*[1]

[1] Research Institute of Electrical Communication, Tohoku Univ., 2-1-1, Katahira, Sendai, Japan
[2] School of Information Science, JAIST, 1-1, Asahidai, Nomi-shi, Ishikawa, Japan

Abstract

To improve the performance of multi-channel speech enhancement algorithms, we previously proposed a hybrid Wiener post-filter for microphone arrays under the assumption of a diffuse noise field [4]. In this paper, considering the speech presence uncertainty, we further improve the hybrid post-filter presented before by integrating a novel robust estimator for the *a priori* speech absence probability, which makes full use of the correlation characteristics of the noises on different microphone pairs and hence offers the much more accurate speech absence probability estimates. The effectiveness of this improved hybrid post-filter was finally confirmed by the experiments using multi-channel recordings in various car environments.

Index Terms: Speech enhancement, Microphone array, Post-filtering, Speech absence probability.

1. Introduction

In a noisy environment, speech signals impinging on the distant microphones are severely contaminated, resulting in the significant performance decrease of many applications, such as, mobile phone and hearing aid. Enhancing the desired speech signal on distant microphones is currently one of the important issues for hands-free technologies. Therefore, practically effective and computationally efficient speech enhancement algorithms are greatly called for. Multi-channel algorithms have shown great superiorities in reducing noise signals and enhancing desired speech signal [1].

Among multi-channel speech enhancement algorithms, post-filter is normally used to further improve the performance of microphone arrays in practical environments [1]. A widely used multi-channel post-filter was first presented by Zelinski under the unpractical assumption of a perfectly incoherent noise field [2]. By relaxing this assumption to that of a diffuse noise field, McCowan *et al.* developed a general expression of the Zelinski post-filter based on the *a priori* coherence function of the noise field [3]. The main drawbacks of these post-filters are the inability to suppress spatially correlated noises and the use of the *a priori* coherence function. To overcome these problems, authors have recently presented a hybrid Wiener filter under the assumption of a diffuse noise field which was proven to be an approximate model in many practical environments [4]. This hybrid Wiener filter follows the framework of the multi-channel Wiener filter, and is capable to reduce spatially correlated as well as uncorrelated noise components.

Moreover, considering the fact that no desired speech signal is present in all frequencies and all frames, many researchers proposed to improve the speech enhancement algorithms under speech presence uncertainty by combining the *speech absence probability* (SAP) [5, 6, 7]. Cohen *et al.* suggested an estimator for the *a priori* SAP based on the energy distribution of the signals at the microphone array output [5]. However, the signal energy distribution at microphone array output has been changed due to the effect of the microphone-array (e.g., beamforming) noise reduction algorithms. Therefore, authors proposed an estimator for the *a priori* SAP based on the coherence characteristics of the noise field at microphone array output because the spatial characteristics of the noise field was proven to be more stable than the statistics of the signals and the spatial characteristics were preserved at the microphone array output [6]. However, this estimator did not fully utilize the coherence characteristics on different microphone pairs in multi-microphone scenarios, resulting in some estimation errors caused by the correlated noise components existing on different microphone pairs.

In this paper, we propose an improved hybrid post-filter for microphone arrays under the assumption of a diffuse noise field by integrating the hybrid Wiener post-filter we presented before and a robust estimator of the *a priori* SAP. This robust estimator fully considers and utilizes the spatial correlations between noise signals on different microphone pairs. In particular, we show that the calculation of the original hybrid post-filter and that of this newly derived the *a priori* SAP estimator can be done in an integrated mechanism. Experimental results show that this improved post-filter outperforms the original post-filter in various car environments.

2. Signal Model

Let us consider a M-sensor microphone array in a noisy acoustic environment. The observed signal $x_m(t)$ on the m-th sensor is composed of two components. The first is the desired signal $s_m(t)$ by transforming the desired source signal $s(t)$ with the impulse response $a_m(t)$ between the sound source and the m-th sensor. The second is the additive noise $n_m(t)$. Thus, applying the *short-time Fourier transform* (STFT), the observed signal on the m-th microphone can be represented as

$$X_m(k,\ell) = S_m(k,\ell) + N_m(k,\ell), \quad m = 1, 2, \cdots, M \quad (1)$$

where $X_m(k,\ell)$, $S_m(k,\ell)$ and $N_m(k,\ell)$ are the STFTs of the corresponding signals $x_m(t)$, $s_m(t)$ and $n_m(t)$, respectively. In this paper, we further assume that the microphone array has been steered into the direction of the desired signal beforehand.

3. Analysis of Diffuse Noise Field

To characterize a noise field, a widely used measure is the *magnitude-squared coherence* (MSC) function, also called coherence function, defined as

$$\Gamma_{x_i x_j}(k,\ell) = \frac{|\phi_{x_i x_j}(k,\ell)|^2}{\phi_{x_i x_i}(k,\ell)\phi_{x_j x_j}(k,\ell)}, \quad (2)$$

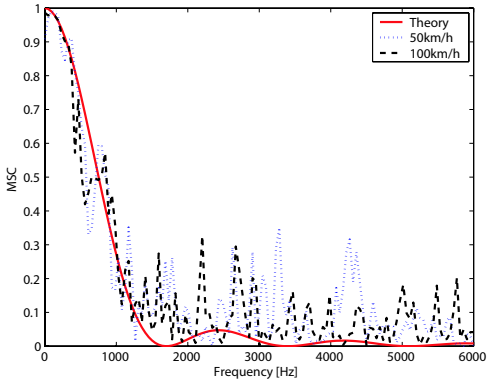

Figure 1: Magnitude-squared coherence function in car environment (d = 10cm).

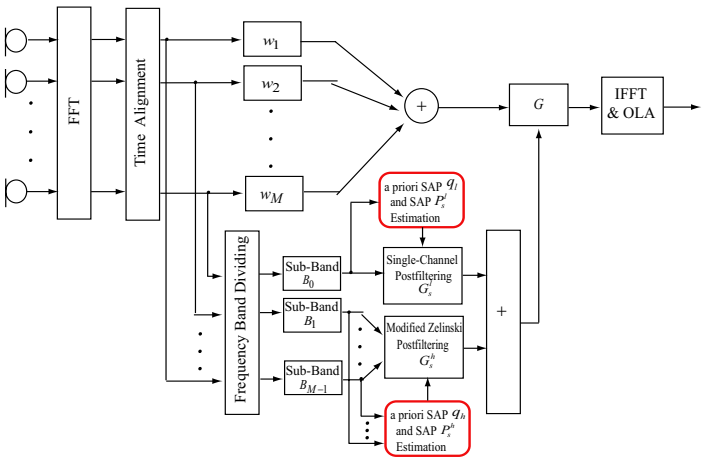

Figure 2: Block diagram of the proposed system.

where $\phi_{x_i x_j}(k,\ell)$ is the cross-spectral density between two signals $x_i(t)$ and $x_j(t)$; $\phi_{x_i x_i}(k,\ell)$ and $\phi_{x_j x_j}(k,\ell)$ are the auto-spectral densities of $x_i(t)$ and $x_j(t)$, respectively.

A diffuse noise field, which is one of the underlying assumptions of this paper, has been shown to be a reasonable model for many practical noise environments [1, 6]. A diffuse noise field is characterized by the following coherence function

$$\Gamma(k) = \left| \frac{\sin(2\pi kd/c)}{2\pi kd/c} \right|^2, \tag{3}$$

where d and c represent the distance between adjacent microphones and the velocity of sound. The coherence function of a perfect diffuse noise field against frequency is plotted in Fig. 1. From Fig. 1 and Eq. 3, some characteristics of the diffuse noise field can be easily observed: (i) coherence function is a frequency-dependent and time-invariant measure; (ii) noises on different microphones are high-correlated in the low frequencies and low-correlated in the high frequencies.

4. An Improved Hybrid Wiener Post-Filer

Under the speech presence uncertainty, we can derive an estimator for two states "speech presence" and "speech absence" as a weighted sum of individual estimators of the speech signal calculated in the two states. The weights are the *a posteriori* probabilities of the two states given the noisy observation. Since the optimal estimator of the desired speech signal given that this signal is absent in the noisy observation is zero, the resulting composite estimator is the product of the estimator of the desired speech signal given that this signal is present in the noisy observation and the *a posteriori* probability of signal presence given the noisy signal. Thus, the improved hybrid Wiener post-filter can be represented as

$$G(k,\ell) = G_s(k,\ell) P_s(k,\ell), \tag{4}$$

where $G_s(k,\ell)$ is the gain function of the hybrid Wiener post-filter when speech is surely present, defined in [4] and briefly formulated in the following subsections; $P_s(k,\ell)$ is the *speech presence probability* given the noisy observation, given by [4, 5]

$$P_s(k,\ell) = \left\{ 1 + \frac{q(k,\ell)}{1-q(k,\ell)} \Big(1+\xi(k,\ell)\Big) \times \exp\left(\frac{\xi(k,\ell)\gamma(k,\ell)}{1+\xi(k,\ell)} \right) \right\}^{-1}, \tag{5}$$

where (i) $q(k,\ell)$ is the *a priori* SAP; (ii) $\xi(k,\ell) = \lambda_s(k,\ell)/\lambda_n(k,\ell)$ and $\gamma(k,\ell) = |X(k,\ell)|^2/\lambda_n(k,\ell)$ are the *a priori* SNR and *a posteriori* SNR as named in [5, 7]; (iii) $\lambda_s(k,\ell)$ and $\lambda_n(k,\ell)$ are the variances of speech signal and noise signal, respectively.

The basic ideas of our estimator of the *a priori* SAP which is based on coherence characteristics are: (i) desired speech signals are strongly correlated on the concerned microphone pairs; (ii) noise signals are weakly correlated on the concerned microphone pairs. Therefore, this technique will fail when speech and noise signals are strongly correlated simultaneously on all microphones. Previously, we presented an estimator for the *a priori* SAP based on coherence characteristics [6], however, it does not sufficiently consider the correlations of the noises on different microphone pairs. It is believed that the discriminating accuracy between speech and noise signals, i.e., the accuracy of the *a priori* SAP estimator, will be improved when sufficiently considering and using all spatial correlation information of noise signals on different microphone pairs.

Considering the spatial coherence characteristics of the noise field, we divide the frequency band into two parts: the low frequency region with high noise coherence and the high frequency region with low noise coherence. In the two regions, the hybrid Wiener post-filter [4] and the robust estimator of the *a priori* SAP are implemented in an integrated way. The block diagram of the proposed post-filter along with a beamformer is plotted in Fig. 2.

4.1. Improved hybrid post-filter in the high frequencies

As Fig. 1 demonstrates, in a diffuse noise field, the spatially low-correlated noise components on different microphones only exist in the frequencies over the transient frequency $f_t = c/(2d)$ [4, 6]. Since the transient frequency is determined by the distance between microphones, microphone pairs with different inter-element spacing are characterized by different transient frequencies. That is, for different microphone pairs with different inter-element spacing, low-correlated noise is found in different frequency regions. Furthermore, for a certain frequency, noise is mutually low-correlated only on limited microphone pairs, generally not on all pairs. This fact motivates us to propose a robust estimator for the *a priori* SAP which makes full use of the spatial correlations of noises on different microphone pairs.

In the high frequency region, the proposed robust estimator of the *a priori* SAP and the hybrid post-filter (i.e., a modified Zelinski post-filter in high frequencies) are implemented in the following steps:

1. Determine the transient frequencies according to the microphone array geometry. Considering a M-sensor array with the equal adjacent-element spacing d and the distance d_{ij} between sensors i and j ($i, j \leq M$), we have $M(M-1)/2$ microphone pairs which determine $M(M-1)/2$ transient frequencies, each of them can be calculated by $f_{t,ij} = c/(2d_{ij})$. Since the inter-element spacings are identical for some microphone pairs, some transient frequencies are identical as well. In principle, among $M(M-1)/2$ microphone pairs, only $M-1$ pairs have different inter-element spacings. Correspondingly, we can determine $M-1$ different transient frequencies, denoted by $f_t^1, f_t^2, \cdots, f_t^{M-1}$. Without loss of generality, we further assume the following relationship between transient frequencies $f_t^1 < f_t^2 < \cdots < f_t^{M-1}$.

2. Determine the microphone pairs on which noise is mutually uncorrelated for each frequency. As a matter of fact, the $M-1$ different transient frequencies, $f_t^1, f_t^2, \cdots, f_t^{M-1}$, divide the full frequency band into M sub-bands, denoted by $B_0, B_1, \cdots, B_{M-1}$. In each sub-band (except B_0), some microphone pairs provide low-correlated noise components on microphones of the pairs. In principle, the $M(M-1)/2$ microphone pairs can be grouped into $M-1$ sets where some microphone pairs are re-used. Each of $M-1$ sets includes the microphone pairs on which noise signals are mutually low-correlated for the individual frequency of interest. Corresponding to the transient frequencies $f_t^1, f_t^2, \cdots, f_t^{M-1}$, the $M-1$ microphone pair sets are represented as: $\Omega_1, \Omega_2, \cdots, \Omega_{M-1}$.

3. Compute the magnitude squared coherence functions. For each frequency in sub-band $B_m (1 \leq m \leq M-1)$, the noise on the microphone pairs of set Ω_m is assumed to be weakly correlated. Therefore, MSC functions calculated in this situation provide much more accurate cues to detect the presence of the desired speech signal. To improve the robustness of the *a priori* SAP estimator against estimation errors, estimates of the auto- and cross-spectral densities are averaged across the microphone pairs in the corresponding pair set Ω_m (not all microphone pairs). The MSC function is thus given by

$$\Gamma_h(k,\ell) = \frac{\frac{1}{|\Omega_m(k)|}\sum_{\{i,j\}\in\Omega_m(k)} \left|\phi_{x_i x_j}(k,\ell)\right|^2}{\frac{1}{|\Omega_m(k)|}\sum_{\{i,j\}\in\Omega_m(k)} \left[\phi_{x_i x_i}(k,\ell)\phi_{x_j x_j}(k,\ell)\right]}. \quad (6)$$

4. Compute the a priori SAP. After calculating the MSCs, we can detect the desired speech signal as follows. If a high coherence (higher than a threshold T_{max}^h) is observed, a speech present state is detected presumably. If a low coherence (lower than a threshold T_{min}^h) is observed, a speech absent state is detected presumably. Note that the *a priori* SAP decreases as the MSC increases. For the MSCs in the range $[T_{min}^h, T_{max}^h]$, the *a priori* SAPs are determined by linear interpolation. Thus, the *a priori* SAPs in the high frequency region $q_h(k,\ell)$ is given by

$$q_h(k,\ell) = \begin{cases} 0, & \Gamma(k,\ell) > T_{max}^h \\ 1, & \Gamma(k,\ell) < T_{min}^h \\ \frac{T_{max}^h - \Gamma(k,\ell)}{T_{max}^h - T_{min}^h}. & \text{otherwise} \end{cases} \quad (7)$$

Moreover, due to the low noise correlation on the microphone pairs of set Ω_m, the original hybrid post-filter (i.e., the modified Zelinski post-filter) in the high frequency region $G_s^h(k,\ell)$ was calculated as [4]

$$G_s^h(k,\ell) = \frac{\frac{1}{|\Omega_m(k)|}\sum_{\{i,j\}\in\Omega_m(k)} \Re\{\phi_{x_i x_j}(k,\ell)\}}{\frac{1}{|\Omega_m(k)|}\sum_{\{i,j\}\in\Omega_m(k)} \left[\frac{1}{2}\left(\phi_{x_i x_i}(k,\ell) + \phi_{x_j x_j}(k,\ell)\right)\right]}. \quad (8)$$

Note that the averages for the auto- and cross- spectral densities are performed on only limited microphone pairs in the corresponding pair set Ω_m on which noises are weakly correlated. Therefore, it avoids the estimation error caused by the correlated noises and offers more accurate and robust estimates for the *a priori* SAPs, further improving the performance of the improved hybrid Wiener post-filter.

4.2. Improved hybrid post-filter in the low frequencies

In the low frequency sub-band (B_0 where $k < f_t^1$), MSCs computed in this region fail to detect the speech signal because both speech and noise are strongly correlated on all microphone pairs. Here, we compute a MSC value that is averaged across the frequencies over the minimum transient frequency, given by

$$\bar{\Gamma}(k,\ell) = \frac{\frac{1}{M-1}\sum_{m=1}^{M-1}\frac{1}{|\Omega_m(k)|}\sum_{\{i,j\}\in\Omega_m(k)} \left|\phi_{x_i x_j}(k,\ell)\right|^2}{\frac{1}{M-1}\sum_{m=1}^{M-1}\frac{1}{|\Omega_m(k)|}\sum_{\{i,j\}\in\Omega_m(k)} \left[\phi_{x_i x_i}(k,\ell)\phi_{x_j x_j}(k,\ell)\right]}. \quad (9)$$

Following the same concept used in the high frequency region, we derive an estimator for the *a priori* SAP in the low frequency region $q_l(k,\ell)$, given by

$$q_l(k,\ell) = \begin{cases} 0, & \bar{\Gamma}(k,\ell) > T_{max} \\ 1, & \bar{\Gamma}(k,\ell) < T_{min} \\ \frac{T_{max} - \bar{\Gamma}(k,\ell)}{T_{max} - T_{min}}. & \text{otherwise} \end{cases} \quad (10)$$

Moreover, due to the high correlations of speech and noise signals on all microphone pairs, the original hybrid post-filter (i.e., the Wiener post-filter) in the low frequency region $G_s^l(k,\ell)$ was calculated by [4]

$$G_s^l(k,\ell) = \frac{\xi(k,\ell)}{1+\xi(k,\ell)}. \quad (11)$$

Consequently, substituting the *a priori* SAP estimates, given by Eqs. (7) and (10) into Eq. (5), we can obtain the speech presence probability estimates. Finally, putting Eqs. (5), (8) and (11) into Eq. (4), we can derive the proposed improved hybrid post-filter which is expected to show more effective in speech enhancement.

5. Experiments and Results

To validate the effectiveness of the improved hybrid post-filter in a diffuse noise field, its performance was investigated and further compared with the original hybrid Wiener post-filter that we proposed before [4] in various car noise environments. A beamformer, implemented by a superdirective beamformer [8], is first applied to the multi-channel noisy signals. Then, the beamformer output is further enhanced by the studied post-filters. The performance is evaluated in terms of objective measures: *segmental SNR* (SEGSNR) and *log-spectral distance* (LSD), defined in [5].

Table 1: Segmental SNR [dB] results of the noisy signal, superdirective beamformer (SDBF), the original hybrid post-filter (ORG-PF) and the improved hybrid post-filter (Imp-PF).

Global SNR	−5	0	5	10	15	20
Condition	50km/h					
Noisy	−9.03	−8.56	−8.14	−7.73	−7.37	−7.07
SDBF	−8.95	−8.46	−8.00	−7.55	−7.12	−6.81
ORG-PF	−7.95	−7.56	−7.15	−6.71	−6.12	−5.22
Imp-PF	**−6.88**	**−6.09**	**−4.30**	**−1.95**	**−0.39**	**0.44**
Condition	100km/h					
Noisy	−9.07	−8.62	−8.18	−7.74	−7.36	−7.07
SDBF	−8.96	−8.48	−8.02	−7.54	−7.13	−6.82
ORG-PF	−7.95	−7.51	−7.09	−6.57	−5.76	−4.59
Imp-PF	**−6.41**	**−5.00**	**−2.87**	**−1.17**	**−0.02**	**0.56**

Table 2: Log spectrum distance [dB] results of the noisy signal, superdirective beamformer (SDBF), the original hybrid post-filter (ORG-PF) and the improved hybrid post-filter (Imp-PF).

Global SNR	−5	0	5	10	15	20
Condition	50km/h					
Noisy	9.92	7.75	5.92	4.38	3.14	2.16
SDBF	9.20	7.25	5.58	4.18	3.03	2.11
ORG-PF	5.33	4.15	3.18	2.43	1.88	1.51
Imp-PF	**2.22**	**1.71**	**1.42**	**1.26**	**1.21**	**1.18**
Condition	100km/h					
Noisy	10.06	7.74	5.86	4.34	3.13	2.19
SDBF	9.01	7.01	5.38	4.04	2.94	2.05
ORG-PF	5.00	3.90	3.02	2.31	1.79	1.45
Imp-PF	**1.82**	**1.54**	**1.40**	**1.29**	**1.22**	**1.18**

5.1. Experimental configurations

To assess the performance of the studied post-filters, an equally-spaced linear array, consisting of three microphones with inter-element spacing of 10 cm, was mounted above the windshield in a car. The array was about 50 cm apart from and in front of the driver. The multi-channel noise recordings were performed across all channels simultaneously when the car was running at the speeds of 50km/h and 100km/h. The multi-channel speech recordings, consisting of 10 Japanese digits, were performed when the car was stationary. Both speech and noise signals were first re-sampled to 12kHz at 16 bit accuracy. We generated the multi-channel noisy signals by artificially mixing the multi-channel speech signals and the multi-channel noise signals at different global SNR levels $[-5, 20]$ dB.

The effectiveness of the diffuse noise field was investigated by comparing the measured coherence function calculated from real noise recordings with the theoretical function, plotted in Fig. 1. Note that the measured coherence function follows the trend of the theoretical function. Therefore, it fulfills the assumption of a diffuse noise field used in the proposed post-filter.

5.2. Objective evaluation results

Experimental results of the average SEGSNR, calculated in two noise conditions at various SNR levels, are listed in Table 1; the results of LSD are presented in Table 2. The performance was evaluated at the first microphone, the beamformer output and the studied post-filter outputs.

As shown in Table 1, the beamformer alone provides only a small degree of SNGSNR improvements compared to the noisy inputs due to its low ability in reducing the noise components in the low-frequency region where car noises mainly concentrate. The original hybrid post-filter gives relatively higher average SEGSNR improvement of about 1.3 dB. The improved hybrid post-filter newly proposed demonstrates the highest average SEGSNR improvement of about 5.17 dB, corresponding the highest speech quality.

Concerning the results of LSD, shown in Table 2, we can readily observe that the beamformer alone decrease the LSD more or less in all conditions. Much lower LSD of about 2.56 dB decrease is observed at the original hybrid post-filter output because of the reduced noise components. Further, the improved post-filter offers the lowest LSD of about 4.1 dB decrease, corresponding to the lowest speech distortion.

6. Conclusion

In this paper, we proposed an improved hybrid Wiener post-filter for microphone arrays by integrating the hybrid post-filter we previously presented and a robust speech presence probability estimator. In this improved post-filter, we fully consider the coherence characteristics of a diffuse noise filed, which offers the more accurate SAP estimates and further improve the noise reduction performance of this post-filter. Experimental results demonstrate the superiority of this improved hybrid Wiener post-filter in reducing noise signals and preserving desired speech signal in various car environments.

7. References

[1] M. S. Brandstein and D. B. Ward (eds.), *Microphone Arrays: Signal Processing Techniques and Applications*, Springer-Verlag, Berlin, 2001.

[2] R. Zelinski, "A Microphone Array With Adaptive Post-Filtering for Noise Reduction in Reverberant Rooms," in *Proc. IEEE Int. Conf. on Acoustic, Speech, Signal Processsing*, vol. 5, pp. 2578-2581, 1988.

[3] I. A. McCowan and H. Bourlard, "Microphone Array Post-Filter Based on Noise Field Coherence," *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 6, pp. 709-716, 2003.

[4] J. Li and M. Akagi, "A hybrid microphone array post-filter in a diffuse noise field," In *Eurospeech2005*, Lisbon, Portugal, pp. 2313-2316, 2005.

[5] I. Cohen, "Multi-channel post-filtering in non-stationary noise environments," *IEEE Trans. Signal Processing*, vol. 52, no. 5, pp. 1149-1160, 2004.

[6] J. Li and M. Akagi, "A noise reduction system based on hybrid noise esitmation techniques and post-filtering in arbitrary noise environments," *Speech Communication*, pp. 111-126, 2006.

[7] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean Square Error Short-Time Spectral Amplitude Estimator," *IEEE Trans. on Acoustic, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109-1121, 1984.

[8] H. Cox, R. Zeskind and M. Owen, "Robust Adaptive Beamforming'" *IEEE Trans. on Acoustic, Speech and Signal Processing*, vol. 35, pp. 1365-1376, 1987.