JAIST Repository
https://dspace.jaist.ac.jp/

Title	A hybrid microphone array post-filter in a diffuse noise field
Author(s)	Li, Junfeng; Akagi, Masato
Citation	Applied Acoustics, 69(6): 546-557
Issue Date	2008-06
Type	Journal Article
Text version	author
URL	http://hdl.handle.net/10119/4892
Rights	NOTICE: This is the author's version of a work accepted for publication by Elsevier. Junfeng Li and Masato Akagi, Applied Acoustics, 69(6), 2008, 546-557, http://dx.doi.org/10.1016/j.apacoust.2007.01.005
Description	

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

A Hybrid Microphone Array Post-filter in a Diffuse Noise Field

Junfeng Li and Masato Akagi
School of Information Science
Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa, 923-1292, Japan

Abstract In this paper, a hybrid post-filter for microphone arrays with the assumption of a diffuse noise field is proposed to suppress correlated as well as uncorrelated noise. In the proposed post-filter, a modified Zelinski post-filter, which is estimated using the signals on the microphone pairs on which noises are uncorrelated by considering the correlation characteristics of noise impinging on different microphone pairs, is applied to the high frequencies to suppress spatially uncorrelated noise; a single-channel Wiener post-filter is applied to the low frequencies for cancellation of spatially correlated noise. In theory, the proposed post-filter is a Wiener post-filter. In practice, experiments using multi-channel recordings were conducted, and experimental results demonstrate the usefulness and superiority of the proposed post-filter compared to other post-filters using speech quality measures and speech recognition rate.

Keyword microphone array; diffuse noise filed; coherence function; hybrid post-filter

1 Introduction

Hands-free technology is desirable for a large number of applications, such as mobile phone and automatic speech recognition system, due to the convenience and flexibility it provides. One main problem associated with this technology is that the signals received by the distant microphones are severely corrupted by various kinds of noises [1]. A potential solution to this problem is the use of microphone arrays due to their spatial filtering capability of suppressing interfering signals arriving from directions other than the look-direction, thus yielding high-quality speech and exhibiting substantial superiority in reducing noise [2].

Recently, Simmer *et al.* [2] reported that: the multi-channel Wiener filter provides the optimal solution to the problem of the multi-channel noise reduction for broadband inputs in *minimum mean square error* (MMSE) sense and can further be decomposed into a *minimum variance distortionless response* (MVDR) beamformer followed by a Wiener

post-filter. Therefore, a post-filter which is based on Wiener theory is normally needed to improve the performance of microphone arrays in practical noise environments [2].

A variety of post-filtering techniques have been reported in the literature [3] [4] [5] [6]. One commonly used multi-channel post-filter, which is based on Wiener filter, was first introduced by Zelinski [3]. The basic assumption behind this post-filter is that noise on different microphones is mutually uncorrelated, corresponding to a perfectly incoherent noise field. This assumption is, however, seldom satisfied in practical environments, especially for closely-spaced microphones and low frequencies which are characterized by the high-correlated noise.

To suppress the high-correlated noise, Fischer *et al.* proposed to combine the *generalized sidelobe canceller* (GSC) with the Zelinski post-filter to suppress the spatially correlated and uncorrelated noise [7]. However, Bitzer *et al.* pointed out that neither the GSC nor the Zelinski post-filter performs well at low frequencies [8]. An alternative solution, presented by Meyer *et al.*, applies the spectral subtraction to suppress the high-correlated noise components [9]. However, this method introduces the artificial "musical noise" and fails to deal with non-stationary noise due to the *voice activity detector* (VAD) based noise estimation technique. Recently, McCowan *et al.* developed a general expression of the Zelinski post-filter based on the *a priori* coherence function of the noise filed [4]. Although this post-filter was shown to achieve improved speech quality and speech recognition accuracy compared to the Zelinski post-filter using the office room recordings, its performance is expected to be significantly degraded when difference between the "actual" and assumed coherence function exists [4].

Recently, a single-channel noise suppression algorithm, referred to as *optimally-modified log-spectral amplitude* (OM-LSA) estimator, was presented for minimizing the log-spectral amplitude distortion in non-stationary noise environments [20]. This OM-LSA estimator was also extended to a multi-channel post-filtering approach when multi-channel inputs are available, which was shown effective in reducing highly non-stationary noise components from the desired source components based on the energy-based speech presence probability estimator [6] [22]. Considering the spatially stable characteristics of noise fields, a speech presence probability estimator based on these spatial characteristics was presented to improve the performance of the OM-LSA post-filter [23] [25]. However, the inherent sensitive implementation parameters involved in the variants of the OM-LSA post-filter [6] [20] [22] greatly degrade their performance in practical environments.

It has been shown that a diffuse noise field provides a reasonable model for a large

number of practical noise environments, such as in reverberant rooms and car environments [2] [4] [9]. For the traditional post-filters, though the Zelinski and McCowan post-filter are based on Wiener theory, they fail to reduce diffuse noise [3] [4]. In contrast, though the OM-LSA post-filters with suitable implementation parameters might be able to deal with diffuse noise, they are not based on Wiener theory, violating the framework of the multi-channel Wiener filter [6] [22]. However, to the authors' knowledge, no existing post-filters in theory is based on Wiener filter and in practice can deal with diffuse noise, which offers the motivation for this research.

In this paper, we propose a novel post-filter with a hybrid structure for microphone arrays under the assumption of a diffuse noise field which is characterized by the low coherence in high frequencies and the high coherence in low frequencies. To suppress the spatially low correlated noise components, a modified Zelinski post-filter is presented and used. In this modified Zelinski post-filter, considering the correlations between noises on different microphone pairs, we further divide the full frequency band into several sub-bands according to the microphone array geometry, in each sub-band the post-filter is estimated using the signals on limited (generally not all) microphone pairs on which noises are low correlated. To suppress the spatially high correlated noise components, a single-channel Wiener post-filter is adopted which produces less "musical noise" due to the use of the decision-directed SNR estimation mechanism. The merits of the proposed post-filter lie in: in theory, it is a Wiener filter, while the OM-LSA post-filters are not based on Wiener theory; in practice, it fully considers and utilizes the correlations between noises on different microphone pairs, resulting in high capability to reduce low-correlated as well as high-correlated noise with minimum speech distortion in a diffuse noise field. The superiorities of the proposed post-filter were verified using the multi-channel recordings in various car environments.

The remainder of this paper is organized as follows. Section 2 formulates the problem to be solved. Section 3 provides a review of the Zelinski post-filter which the proposed post-filter is based on and the McCowan post-filter which is also used for comparison. Section 4 analyzes the spatial characteristics of a diffuse noise field, and then presents the proposed hybrid microphone array post-filter. Section 5 describes some experimental results along with some discussions to validate the advantages of the proposed post-filter. Section 6 draws some conclusions.

2 Problem Formulation

Let consider a M-sensor microphone array in a noisy environment. The observed signal $x_i(t)$ on the i-th sensor is composed of two components: the desired signal $s(t)$ and the additive noise $n_i(t)$, $i = 1, 2, \ldots, M$. *Time delay compensation* (TDC) for the desired speech signal on each microphone can be done using the coherence based time delay estimation technique [13]. In this paper, for the explanation simplicity, we assume that the TDC has been performed perfectly in advance. Hence, applying the *short-time Fourier transform* (STFT), the TDC output signal $X_i(k, \ell)$ on the i-th microphone in the time-frequency domain can be represented as:

$$X_i(k, \ell) = S(k, \ell) + N_i(k, \ell), \quad i = 1, 2, \cdots M \tag{1}$$

where k and ℓ are the frequency index and frame index, respectively; $X_i(k, \ell)$, $S(k, \ell)$ and $N_i(k, \ell)$ are the STFTs of the corresponding signals.

This paper focuses on addressing the problem of estimating the Wiener post-filter with the assumption of a diffuse noise field. The Zelinski post-filter [3] and the McCowan post-filter [4] have been presented, however, the associated drawbacks are: the incapability in reducing high-correlated noise and the required *a priori* knowledge of the noise field.

3 Review of Related Work

In this section, we briefly review two post-filters, referred to as the Zelinski post-filter and the McCowan post-filter. Our proposed post-filter is based on the Zelinski post-filter and compared to both.

3.1 Zelinski Post-Filter

The Zelinski post-filter approaches a Wiener filter in a perfectly incoherent noise field based on the estimates of the auto- and cross- spectral densities. With the assumptions that the desired signal and noise signal are uncorrelated and that noise on different microphones is also uncorrelated and of identical power spectral density, the auto- and cross- spectral densities of multi-channel inputs, $\phi_{x_i x_i}(k, \ell)$ and $\phi_{x_i x_j}(k, \ell)$, can be simplified as:

$$\phi_{x_i x_i}(k, \ell) = \phi_{ss}(k, \ell) + \phi_{nn}(k, \ell), \tag{2}$$

$$\phi_{x_i x_j}(k, \ell) = \phi_{ss}(k, \ell). \tag{3}$$

Based on the simplified expressions of the auto- and cross- spectral densities, the

Zelinski post-filter can be formulated as [3]:

$$G_z(k,\ell) = \frac{\frac{2}{M(M-1)} \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} \Re\{\phi_{x_i x_j}(k,\ell)\}}{\frac{1}{M} \sum_{i=1}^{M} \phi_{x_i x_i}(k,\ell)}, \tag{4}$$

where $\Re\{.\}$ is the real operator. Note, to improve the robustness of this post-filter, the averaging operation is performed across all sensor pairs. Moreover, it is of interest to note that the auto- and cross- spectral densities are estimated from the multi-channel inputs. This estimation technique slightly over-estimates the noise spectral density [15]. However, it has been proven to give a high noise reduction performance and also is widely used [3] [4] [14] [15]. Therefore, this estimation technique is also employed in our proposed hybrid Wiener post-filter, detailed in the following.

3.2 McCowan Post-Filter

As a matter of fact, the basic assumption of the Zelinski post-filter, that noise on each microphone is uncorrelated, is seldom satisfied in practical environments. Considering this fact, McCowan relaxed this practically unreasonable assumption to the one that noise on each microphone, of identical power spectral densities, is correlated through the coherence function [4].

With the assumption of zero correlation between the desired speech signal and noise signal and the relaxed assumption, the auto- and cross- spectral densities of multi-channel inputs, $\phi_{x_i x_i}(k,\ell)$ and $\phi_{x_i x_j}(k,\ell)$, can be simplified as:

$$\phi_{x_i x_i}(k,\ell) = \phi_{ss}(k,\ell) + \phi_{nn}(k,\ell), \tag{5}$$

$$\phi_{x_i x_j}(k,\ell) = \phi_{ss}(k,\ell) + \Gamma_{n_i n_j}(k,\ell)\phi_{nn}(k,\ell), \tag{6}$$

where $\Gamma_{n_i n_j}(k,\ell)$ is the complex coherence function, defined as:

$$\Gamma_{n_i n_j}(k,\ell) = \frac{\phi_{n_i n_j}(k,\ell)}{\sqrt{\phi_{n_i n_i}(k,\ell)\phi_{n_j n_j}(k,\ell)}}. \tag{7}$$

Based on these expressions of auto- and cross- spectral densities, the speech power spectral density, which is the numerator term of the Wiener post-filter, can be represented as:

$$\hat{\phi}_{ss}^{(ij)}(k,\ell) = \frac{\Re\{\phi_{x_i x_j}(k,\ell)\} - \frac{1}{2}\Re\{\Gamma_{n_i n_j}(k,\ell)\}\left(\phi_{x_i x_i}(k,\ell) + \phi_{x_j x_j}(k,\ell)\right)}{1 - \Re\{\Gamma_{n_i n_j}(k,\ell)\}}. \tag{8}$$

Then the McCowan post-filter can be derived as [4]:

$$G_M(k,\ell) = \frac{\frac{2}{M(M-1)} \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} \hat{\phi}_{ss}^{(ij)}(k,\ell)}{\frac{1}{M} \sum_{i=1}^{M} \phi_{x_i x_i}(k,\ell)}. \tag{9}$$

Although the McCowan post-filter has been shown to achieve improved performance compared to the Zelinski post-filter using multi-channel recordings in an office, a significant performance degradation is expected when difference between the actual and assumed coherence functions exists. The performance dependence of the McCowan post-filter on the assumed coherence function was also analyzed in [4].

4 Proposed Microphone Array Post-Filter

In this section, we first describe the coherence function and its application in analyzing a noise field. Then a hybrid post-filter with the assumption of a diffuse noise field is proposed. Finally, advantages of the proposed post-filter are presented qualitatively.

4.1 Analysis of a Noise Field

To characterize a noise field, a widely used measure is the *magnitude-squared coherence* (MSC) function, simply called coherence function, defined as the magnitude square of the complex coherence function and given by:

$$\mathrm{MSC}_{n_i n_j}(k,\ell) = \frac{|\phi_{n_i n_j}(k,\ell)|^2}{\phi_{n_i n_i}(k,\ell)\phi_{n_j n_j}(k,\ell)}, \tag{10}$$

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 [2]. A diffuse noise field is characterized by the following MSC function:

$$\mathrm{MSC}(k) = \left| \frac{\sin(2\pi k d/c)}{2\pi k d/c} \right|^2, \tag{11}$$

where d and c represent the distance between adjacent microphones and the velocity of sound. The MSC function of a perfect diffuse noise field against frequency is plotted in Fig. 1. From Fig. 1, some characteristics of a diffuse noise field can be easily observed:

1. The MSC function is a frequency-dependent and time-invariant measure;

2. Noise on different microphones is high-correlated in the low frequencies and low-correlated in the high frequencies.

These observations motivate us to divide the spectrum into the low-correlated and high-correlated parts, the transient frequency f_t between two regions is chosen as the first minimum, given by $f_t = c/(2d)$ [9] [15]. Since the velocity of sound c is considered as a constant, the transient frequency is merely determined by the distance d between two microphones, which is a key point for our proposed post-filter.

4.2 Proposed Post-Filter

To formulate the proposed post-filter, let us first give some assumptions on which it is based:

1. Desired speech signal and noise signal are uncorrelated on each microphone;
2. Noise power spectral density is identical on each microphone;
3. Noise on different microphones is diffuse noise.

As a matter of fact, assumption (1) is normally made in speech signal processing, and assumptions (2) and (3) were verified to be fulfilled in a large number of practical noise environments.

In the following discussion, we propose a hybrid post-filter, which applies a modified Zelinski post-filter in the high frequency region and a single-channel Wiener post-filter in the low frequency region, with the hope of enhancing its noise reduction performance. The block diagram of the proposed post-filter along with beamformer is plotted in Fig. 2.

4.2.1 A Modified Zelinski Post-Filer in the high frequencies

Based on the assumption that noise on each microphone is mutually uncorrelated, the Zelinski post-filter provides a solution for minimizing the mean-square error between speech and its estimate in an incoherent noise field. As mentioned above, its performance is often significantly degraded when the correlated noise components are involved in estimating the cross-spectral densities of multi-channel inputs. It is, therefore, believed that the performance degradation would be eliminated if the noise, used to estimate the cross-spectral densities of multi-channel inputs, is sufficiently uncorrelated.

As Fig. 1 demonstrates, in a diffuse noise field, the spatially weakly correlated noise components on different microphones only exist in the frequencies over the transient frequency f_t. 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 modified Zelinski post-filter by calculating the cross-spectral densities of multi-channel inputs on corresponding microphone pairs, not on all sensor pairs (as used in the Zelinski and the McCowan post-filters).

The modified Zelinski post-filter is implemented in the following steps:

1. *Determine the transient frequencies according to the microphone array geometry.* Considering a M-sensor array with inter-element spacing 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, if the equidistant microphones are assumed, 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 (expect 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 weakly 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 spectral densities of the desired speech signal and noisy signal.* 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 weakly correlated. Thus, the spectral densities of noisy signal and desired speech signal can be estimated from the auto- and cross- spectral densities

of multi-channel inputs, that is:

$$\hat{\phi}_{x_i x_i}(k,\ell) = \phi_{ss}(k,\ell) + \phi_{nn}(k,\ell), \tag{12}$$

$$\hat{\phi}_{x_i x_j}(k,\ell) = \phi_{ss}(k,\ell). \tag{13}$$

4. *Compute the gain function of the modified Zelinski post-filter.* To improve the robustness of the proposed post-filter, estimates of the auto- and cross- spectral densities are averaged across the microphone pairs in the corresponding pair set Ω_m, generally not all microphone pairs. The gain function of the modified Zelinski post-filter is given by:

$$G_{mz}(k,\ell) = \frac{\frac{1}{|\Omega_m(k)|} \sum_{\{i,j\}\in\Omega_m(k)} \Re\{\hat{\phi}_{x_i x_j}(k,\ell)\}}{\frac{1}{|\Omega_m(k)|} \sum_{\{i,j\}\in\Omega_m(k)} \left[\frac{1}{2}\left(\hat{\phi}_{x_i x_i}(k,\ell) + \hat{\phi}_{x_j x_j}(k,\ell)\right)\right]}. \tag{14}$$

Note that in this modified Zelinski post-filter, the average for the auto- and cross-spectral densities is performed on only limited microphone pairs in the corresponding pair set Ω_m determined in step 2. Since noise on microphones in set Ω_m is weakly correlated, the estimation error caused by the correlated noise components should be mitigated, improving the accuracy and robustness of this modified Zelinski post-filter. While in other post-filters including the Zelinski post-filter and the McCowan post-filter, the average was done across all microphone pairs, involving the correlated noise components in estimating the spectral densities, which introduces the estimation error and further degrades the noise reduction performance.

Moreover, it should be noted that the first two steps should be done in advance, since they are only dependent on the microphone array geometry and independent of input signals. The limited microphone pairs, involved in the estimation procedure of the auto- and cross- spectral densities, contribute to the decrease of computational cost of this modified Zelinski post-filter.

4.2.2 A Single-Channel Technique in the low frequencies

In the low frequency sub-band (B_0 where $k < f_t^1$), noise on all microphone pairs is high-correlated, indicating that the auto-spectral density of the desired speech signal can not be estimated from the cross-spectral density of multi-channel inputs. Thus, no post-filter that calculates the auto- and cross- spectral densities can perform well in these frequencies.

In the low frequencies ($k < f_t^1$), therefore, we turn to a single-channel technique to

estimate a Wiener filter. The gain function of the Wiener filter is rewritten here:

$$G_s(k,\ell) = \frac{E\left[|S(k,\ell)|^2\right]}{E\left[|S(k,\ell)|^2\right] + E\left[|N(k,\ell)|^2\right]} = \frac{SNR_{priori}(k,\ell)}{1 + SNR_{priori}(k,\ell)}, \tag{15}$$

where $E[.]$ denotes the expectation operator, and $SNR_{priori}(k,\ell)$ the *a priori* SNR, as named in [17], defined by $SNR_{priori}(k,\ell) = E\left[|S(k,\ell)|^2\right] / E\left[|N(k,\ell)|^2\right]$. The estimate of the *a priori* SNR, $SNR_{priori}(k,\ell)$, is updated in a decision-directed scheme, as follows [17]:

$$SNR_{priori}(k,\ell) = \alpha \frac{|S(k,\ell-1)|^2}{E\left[|N(k,\ell-1)|^2\right]} + (1-\alpha)\max\left[SNR_{post}(k,\ell) - 1, 0\right], \tag{16}$$

where α $(0 < \alpha < 1)$ is a forgetting factor and $SNR_{post}(k,\ell)$ is the *a posteriori* SNR, as named in [17], defined by $SNR_{post}(k,\ell) = |X(k,\ell)|^2 / E\left[|N(k,\ell)|^2\right]$. This decision-directed estimation mechanism for the *a priori* SNR significantly decreases the residual "musical noise", as detailed in [19].

To improve the performance of this single-channel Wiener filter, a crucial point is to estimate noise power spectral density $E\left[|N(k,\ell)|^2\right]$ with high accuracy. Here, it is implemented by a soft-decision based approach, given by:

$$E\left[|N(k,\ell)|^2\right] = \beta E\left[|N(k,\ell-1)|^2\right] + (1-\beta) E\left[|N(k,\ell)|^2 \middle| \mathbf{X}(k,\ell)\right], \tag{17}$$

where β $(0 < \beta < 1)$ is a forgetting factor controlling the update rate of noise estimation. Under speech presence uncertainty, the second term in the right side of Eq. (17) can be estimated as:

$$E\left[|N(k,\ell)|^2 \middle| \mathbf{X}(k,\ell)\right] = q(k,\ell)\overline{|\mathbf{X}(k,\ell)|^2} + (1 - q(k,\ell)) E\left[|N(k,\ell-1)|^2\right], \tag{18}$$

where $q(k,\ell)$ denotes the speech absence probability, $\overline{|\mathbf{X}(k,\ell)|^2} = \frac{1}{M}\sum_{m=1}^{M} |X_m(k,\ell)|^2$ the average of the individual power spectral density on each sensor. The reason for calculating this average is that considering only one sensor may yield a biased measurement. With the assumption of a complex Gaussian statistic model and applying the Bayes rule and total probability theorem, the speech absence probability conditioned on the observations can be given [17]:

$$q(k,\ell) = \left(1 + \frac{1 - q'(k,\ell)}{q'(k,\ell)} \frac{1}{1 + SNR_{priori}(k,\ell)} \exp\left(\frac{SNR_{post}(k,\ell) SNR_{priori}(k,\ell)}{1 + SNR_{priori}(k,\ell)}\right)\right)^{-1}. \tag{19}$$

where $q'(k,\ell)$ is the *a priori* speech absence probability. In the experiments, $q'(k,\ell)$ is set to 0.5 as in [21].

Here, it is of interest to note that the post-filter described above given by Eq. (15) is a Wiener filter exactly. This post-filter, which minimizes the mean square error of spectrum,

is also different from the Ephraim-Malah algorithm which is based on the MSE of spectral amplitude [17]. In comparison of the traditional post-filters, this proposed Wiener filter show some advantages: (i) it is able to greatly reduce the "musical noise" due to the use of the decision-directed *a priori* SNR estimation technique [19]; (ii) it is able to deal with the non-stationary noise due to the soft-decision based noise estimation technique [20].

4.3 Analysis of Proposed Post-Filter

In theory, the proposed post-filter is a Wiener post-filter. In the low frequency region, the single-channel post-filter given by Eq. (15) is obviously a Wiener filter. In the high frequency region, since noise used to formulate the modified Zelinski expression are weakly correlated, the cross-spectral density of multi-channel signals provides more accurate speech auto-spectral density estimate. Therefore, the modified Zelinski post-filter used in the high frequency region approaches a Wiener filter. Comparatively, although the original Zelinski post-filter and the McCowan post-filter have a Wiener-filter structure, performance degradation is expected due to the correlated noise components involved in estimating cross-spectral densities.

It also should note that the proposed post-filter provides a more general expression for the microphone array post-filter. In a perfectly incoherent noise field, the proposed post-filter will reduce to the Zelinski post-filter, just by setting the transient frequencies to zero. And in a perfectly coherent noise field, the proposed post-filter will reduce to the single-channel Wiener post-filter, just by setting the transient frequencies to the highest frequency of interest.

5 Experiments and Results

To validate the effectiveness of the proposed hybrid post-filter in a diffuse noise field, its performance was investigated and further compared to other conventional post-filters, including the Zelinski post-filter [3], the McCowan post-filter [4], the single-channel Wiener post-filter alone [16] and the *noise coherence based optimally modified log-spectral amplitude* (Coh-OM-LSA) estimator [25], in various car noise environments. A beamformer was first applied to the multi-channel noisy signals. Then, the beamformer output was further enhanced by the studied post-filters. The performance was evaluated using speech enhancement and speech recognition experiments.

5.1 Experimental Configurations

The performance of the studied post-filters was assessed in real car noise environments. For this purpose, an equally-spaced linear array, consisting of three microphones with inter-element spacing of 10cm, was mounted on the roof near driver's sun-visor in a car. The array was just about 50cm apart from and directly in front of the driver. Multi-channel noise recordings were performed across all channels when the car was running in two conditions: (1) at speed of 50km/h without air-condition noise (the air condition is off), (2) at speed of 100km/h with high-level air-condition noise (the air condition is on). The effectiveness of the diffuse noise field was investigated by comparing the measured MSC function calculated from real noise recordings with the theoretical function, plotted in Fig. 1. It can be seen from Fig. 1 that the measured MSC function follows the trend of the theoretical function, which fulfills the assumption of a diffuse noise field used in the proposed post-filter.

5.2 Speech Enhancement Experiments

For speech enhancement experiments, multi-channel speech recordings were performed across all channels when the car is stopped. The speech signals, consisting of 100 Japanese city names, were uttered by two speakers (one male and one female) at the driver's position. 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 multi-channel speech recordings and multi-channel car noise recordings in two noise conditions (50km/h and 100km/h) at different global SNR levels [-5, 15] dB. (The calculation of global SNR is detailed in [26].)

In speech enhancement experiments, the beamforming filter was implemented by a superdirective beamformer [12]. Note that with the consideration of robustness, the white noise gain constraining procedure was applied during the implementation [12]. The *directivity index* (DI) of this superdirective beamformer is shown in Fig. 3, which illustrates its low noise reduction performance for the low-frequency noise components.

5.2.1 Objective Evaluation Measures

To evaluate the studied post-filters, two objective speech quality measures were used: segmental SNR and *mel-frequency cepstral coefficient* (MFCC) distance.

The first, *segmental SNR* (SEGSNR), is a widely used objective evaluation measure for speech enhancement and noise reduction algorithms [26]. SEGSNR is defined as the

ratio of the power of clean speech to that of noise signal embedded in a noisy signal or an enhanced signal by the studied algorithms, given by:

$$\text{SEGSNR} = \frac{1}{L}\sum_{\ell=0}^{L-1} 10\log\left(\frac{\sum_{k=0}^{K-1}\left[s(\ell K+k)\right]^2}{\sum_{k=0}^{K-1}\left[\hat{s}(\ell K+k)-s(\ell K+k)\right]^2}\right), \tag{20}$$

where (i) $s(.)$ is the reference speech signal, $\hat{s}(.)$ is the noisy signal or the enhanced signals processed by the tested algorithms; (ii) L and K represent the number of frames in the signal and the number of samples per frame (equal to the length of STFT). Note that a higher SEGSNR means the higher speech quality of enhanced signal [26].

A second evaluation measure, MFCC distance, is defined as the distance between MFCCs of a clean speech signal and those of a noisy signal or enhanced signal, which is given by:

$$d_{\text{mfcc}} = \frac{1}{|\Phi|}\sum_{\ell\in\Phi}\sum_{i}\left(c_i - c_i^{'}\right)^2, \tag{21}$$

where Φ represents the set of frames in which speech is present and $|\Phi|$ its cardinality; c_i and $c_i^{'}$ are the 12-order MFCCs of the clean speech signal and noisy signal or enhanced speech signals, respectively. Note that a lower MFCC distance level indicates lower speech distortion [26].

5.2.2 Objective Evaluation Results

Experimental results of the average SEGSNR and MFCC distance calculated in two noise conditions at various SNR levels, are plotted in Figs. 4 and 5. The values were averaged across all sentences in each noise condition. The performance was evaluated at the first microphone, the beamformer output and the studied post-filter outputs.

As illustrated in Fig. 4, the beamformer shows low SEGSNR improvement due to its little directivity (Fig. 3) in the low frequencies. The Zelinski post-filter also only offers limited performance improvement. By integrating an appropriate coherence function of the noise field into the post-filter formulation, the McCowan post-filter shows a great SEGSNR improvement. The single-channel Wiener post-filter shows further SEGSNR improvements compared with the Zelinski and the McCowan post-filters in all noise conditions. The proposed post-filter demonstrates highest performance improvements in SEGSNR sense among the studied post-filters in all tested conditions.

Concerning the results of MFCC distance, plotted in Fig. 5, we can readily observe that the beamformer alone and the Zelinski post-filter decrease MFCC distances in all

conditions with regard to noisy inputs. Moreover, the single-channel Wiener post-filter shows the lower MFCC distances, especially at low SNRs. The proposed post-filter and the McCowan post-filter offer the lowest speech distortion to an almost same degree at all SNRs, with regard to other post-filters in all noise conditions.

Taking account of noise reduction and speech distortion, the proposed post-filter demonstrates the great superiority, the highest speech quality and the lowest speech distortion, with regard to other comparative post-filters in all noise conditions.

5.2.3 Subjective Evaluation Results

Subjective evaluation of the studied post-filters was performed using speech spectrograms and validated by informal listening tests. Typical examples of speech spectrograms, corresponding to Japanese sentence "hatinohe kesennuma yukuhasi", are presented in Fig. 6, in the car environment with a speed of 100km/h. Fig. 6(d) shows that the output of beamformer is characterized by high-level low-frequency noise due to its weakness in the low frequencies, as shown in Fig. 3. The Zelinski post-filter also offers very limited performance in the low frequencies because of the high-coherence characteristics of noise in this region. Fig. 6(f) illustrates that the McCowan post-filter does suppress a large amount of noise, even in the low frequency region, and the residual noise exists due to the difference between the assumed and actual coherence values at instantaneous time. The single-channel Wiener post-filter results in speech distortion, as shown in Fig. 6(g). Fig. 6(h) illustrates the proposed post-filter is able to further suppress the correlated and uncorrelated noises simultaneously, without additional speech distortion. Informal listening tests proved the superiority of the proposed post-filter compared to others.

5.3 Speech Recognition Experiments

In speech recognition experiments, the beamforming filter was implemented using the subtractive beamformer which is based on the hybrid noise estimation technique we proposed previously [25]. The performance the proposed hybrid Wiener post-filter is then evaluated and further compared to that of the *noise coherence based optimally-modified log-spectral amplitude* (Coh-OM-LSA) estimator presented before in [25] in terms of speech recognition rate.

For speech recognition experiments, the speech data were selected from AURORA-2J database for training and testing [27]. The acoustic model was trained using 8440 sentences, uttered by 55 persons. For testing, the multi-channel car noise recordings (with

the speed of 100km/h) were first re-sampled to 12kHz at 16 bit accuracy. The noise-corrupted data were then generated by adding the randomly selected segments of the multi-channel car noise across 1001 test sentences in AURORA-2J at different SNR levels from 0dB to 20dB in 5dB steps.

The signals were pre-emphasized with a coefficient 0.97. A hamming window of 32ms length with 16ms frame rate was used. The first 12 dimensions of de-correlated log compressed Mel energy spectrum was chosen (the zero-th order coefficient was discarded). Combining with the log power energy, we got 13 dimensional static feature vector. Together with their first and second order dynamic values, 39 dimensional feature vectors were formed. The acoustic models consist of ten digits, one silence and short pause models. Each distribution of digit has 18 states with 16 output distributions. Silence model has 5 states with 3 distributions, and short pause model has 3 states with one distribution. Each distribution of digit has 20 Gaussians while that of silence and short pause has 36 Gaussians. Each model was trained as a left-to-right topology with three states (without skip among states) by using Baum-Welch algorithm with a flat-starting embedded training. Standard Viterbi decoding technique was used for recognition.

5.3.1 Speech Recognition Results

The speech recognition results in terms of speech recognition rate are shown in Fig. 7. Both noise reduction algorithms provide some degree of performance improvement in speech recognition rate compared with noisy inputs. The average recognition rate improvement achieved by the Coh-OM-LSA based noise reduction system amounts to about 13.6%. In contrast, the hybrid post-filter based noise reduction system provides an average recognition rate improvement of about 18.6%. The recognition rate drastically decreases as the noise level increases (i.e., the signal-to-noise ratio decreases). Moreover, in very high SNR conditions, all the tested algorithms provide just slight performance improvement compared with the noisy inputs, which is reasonable since the inputs are "clean" enough and a relatively high recognition rate is achievable in these conditions. In comparison of the Coh-OM-LSA based algorithm, the proposed hybrid post-filter based algorithm provides much higher speech recognition rate in all noise conditions.

The speech recognition results in terms of error reduction rate are shown in Fig. 8. Fig. 8 also illustrates that the proposed hybrid post-filter based noise reduction system gives much higher error reduction rate (about 56.0% on average) than the Coh-OM-LSA based noise reduction system (about 39.7% on average). From the careful observation of

Fig. 8, we especially notice that the recognition error reduction rates achieved by the hybrid post-filter based system amount to 80.5% at 15 dB SNR, 72.9% at 10 dB SNR and 42.2% at 5 dB SNR. These error reduction rate improvements are extensively large for improving the performance of the state-of-the-art speech recognizers in noisy environments.

Based on the speech recognition results above mentioned, we can see that the proposed hybrid post-filter provides better noise reduction performance than the Coh-OM-LSA based post-filter for speech recognition application. This superiority is caused by the high noise reduction performance and the low speech distortion introduced by the hybrid post-filter with regard to the Coh-OM-LSA estimator.

5.4 Discussions

Compared to the other post-filters, the advantages of the proposed post-filter are discussed in this section from the standpoint of practice based on the experimental results.

The proposed hybrid post-filter is superior to the Zelinski post-filter since the basic assumption of the proposed post-filter (diffuse noise field) is more reasonable than that of the Zelinski post-filter (incoherent noise field) in practical environments. In addition, the Zelinski post-filter fail to reduce the low-frequency (high-correlated) noise components, while the proposed hybrid post-filter is successful for these noise components.

The proposed hybrid post-filter is superior to the McCowan post-filter. The McCowan post-filter is determined based on the coherence function of the noise field itself. Thus, its performance is greatly dependent on the accuracy of the assumed coherence function. The differences between the assumed and actual coherence functions result in its performance significant degradation. However, the proposed hybrid post-filter utilizes the transient frequency only to distinguish correlated and uncorrelated noises, independent of the actual instantaneous values of the coherence function, alleviating the effect caused by the difference between the assumed and actual coherence functions in some sense.

The proposed hybrid post-filter should be superior to the single-channel Wiener filter which is used in the whole frequency band. The single-channel Wiener filter, which is based on measurements of noise characteristics, can hardly be applied for highly non-stationary noise sources even if the soft-decision mechanism is adopted. However, multi-channel technique based on the estimates of the auto- and cross- spectral densities theoretically provides good performance for the highly non-stationary noise. Our proposed modified Zelinski post-filter utilizes this attractiveness fully in each frequency bin in the high frequency region. Additionally, in the low frequency region, both the proposed and the

single-channel Wiener post-filters have the same problem in dealing with the highly non-stationary noises.

The proposed hybrid post-filter might be superior to the Coh-OM-LSA estimator for speech recognition application. The speech quality was improved by the Coh-OM-LSA estimator [25], however, the inherent sensitive implementation parameters are not easy to determine and the non-suitable parameters significantly deteriorate the performance of the Coh-OM-LSA based noise reduction system in practical conditions. In contrast, the proposed hybrid post-filter avoids the problem of the sensitive implementation parameters and offers a robust solution for implementing the Wiener-theory-based post-filter which is capable to suppress both low correlated and high correlated noise components in a diffuse noise field.

6 Conclusions

In this paper, we proposed a hybrid post-filter for microphone arrays with the assumption of a diffuse noise field. The proposed post-filter applies a modified Zelinski post-filter in the high frequency region and a single-channel Wiener post-filter in the low frequency region. Compared to other algorithms, the proposed post-filter has the following advantages: (1) in theory, the proposed post-filter is a Wiener filter, comparatively, the OM-LSA based post-filters are not based on Wiener theory; (2) in practice, the proposed post-filter is successful in reducing uncorrelated as well as correlated noise components and preserving the desired speech components, which is attributed to the fact that only signals on microphone pairs where noises are exactly low-correlated are used to calculate the post-filter in the high frequencies. That is, the correlation characteristics of noises on different microphone pairs are fully considered and utilized in this proposed post-filter. Its superiority was verified by the experiments using multi-channel recordings in various car environments.

References

[1] M. Omologo, P. Svaizer and M. Matassoni, "Environmental conditions and acoustic transduction in hands-free speech recognition", *Speech Communication*, vol. 25, pp. 75-95, 1998.

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

[3] 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.

[4] 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.

[5] I. Cohen and B. Berdugo, "Microphone array post-filtering for non-stationary noise suppression," in *Proc. IEEE Int. Conf. Acoustic Speech Signal Processing*, pp. 901-904, May 2002.

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

[7] S. Fischer, K. D. Kammeyer, and K. U. Simmer, "Adaptive microphone arrays for speech enhancement in coherent and incoherent noise fields," in *Proc 3rd joint meeting of the Acoustical Society of America and the Acoustical Society of Japan*, Honolulu, Hawaii, 1996.

[8] J. Bitzer, K.U. Simmer and K.-D. Kammeyer, "Multichannel noise reduction - algorithms and theoretical limits," in *Proc. European Signal Processing Conference*, Rhodes, Greece, pp. 105-108, 1998.

[9] J. Meyer and K. U. Simmer, "Multi-channel speech enhancement in a car environment using Wiener filtering and spectral subtraction," in *Proc. IEEE Int. Conf. on Acoustic, Speech, Signal Processsing*, Munich, Germany, pp. 21-24, 1997.

[10] O. L. Frost, "An algorithm for linearly constrained adaptive array processing," in *Proc. IEEE*, vol. 60, pp. 926-935, 1972.

[11] L.J. Griffiths and C.W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. on Antennas Propagat.*, vol. AP-30, pp. 27-34, 1982.

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

[13] G.C. Carter, "Coherence and time delay estimation," In *Proc. of the IEEE*, vol. 75, no. 2, pp. 236-255, 1987.

[14] C. Marro, Y. Mahieux and K. U. Simmer, "Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering," *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 3, May, 1998.

[15] K. U. Simmer and A. Wasiljeff, "Adaptive microphone arrays for noise suppression in the frequency domain," *Second cost 229 Workshop on Adaptive Algorithm in communication*, pp. 185-194, 1992.

[16] A. A. Azirani, R. L. B. Jeannes and G. Faucon, "Speech enhancement using a Wiener filtering under signal presence uncertainty," in *European Signal Processing Conference*, pp. 971-974, 1996.

[17] 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.

[18] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator", *IEEE Trans. on Acoustic, Speech and Signal Processing*, vol. 33, no. 2, pp. 443-445, 1985.

[19] O. Cappe, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," in *Proc. IEEE Int. Conf. on Acoustic, Speech, Signal Processsing*, vol. 2, no. 2, pp. 345-349, 1994.

[20] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, no. 11, pp. 2403-2418, 2001.

[21] I. Y. Soon, S. N. Koh, and C. K. Yeo, "Improved noise suppression filter using self-adaptive estimator of probability of speech absence", *Signal Processing*, vol. 75, pp. 151-159, 1999.

[22] S. Gannot and I. Cohen, "Speech enhancement based on the General Transfer Function GSC and postfiltering," *IEEE Trans. on Acoustic, Speech and Signal Processing*, vol. 12, no. 6, pp. 561-571, 2004.

[23] J. Li, X. Lu and M. Akagi, "A noise reductin system in arbitrary noise environments and its applications to speech enhancement and speech recognition," in *IEEE Int. Conference on Acoustic, Speech and Signal Processing*, pp.277-280, March, 2005.

[24] J. Li and M. Akagi, "A hybrid microphone array post-filter in a diffuse noise field", In *9-th European Conference on Speech Communication and Technology, Eurospeech2005*, pp. 2313-2316, September, 2005.

[25] J. Li and M. Akagi, "A noise reduction system based on hybrid noise estimation technique and postfiltering in arbirtary noise environments", *Speech Communication*, vol. 48, pp. 111-126, 2006.

[26] S. R. Quackenbush, T. P. Barnwell, M. A. Clements, Objective Measures of Speech Quality. Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1988.

[27] http://sp.shinshu-u.ac.jp/CENSREC.

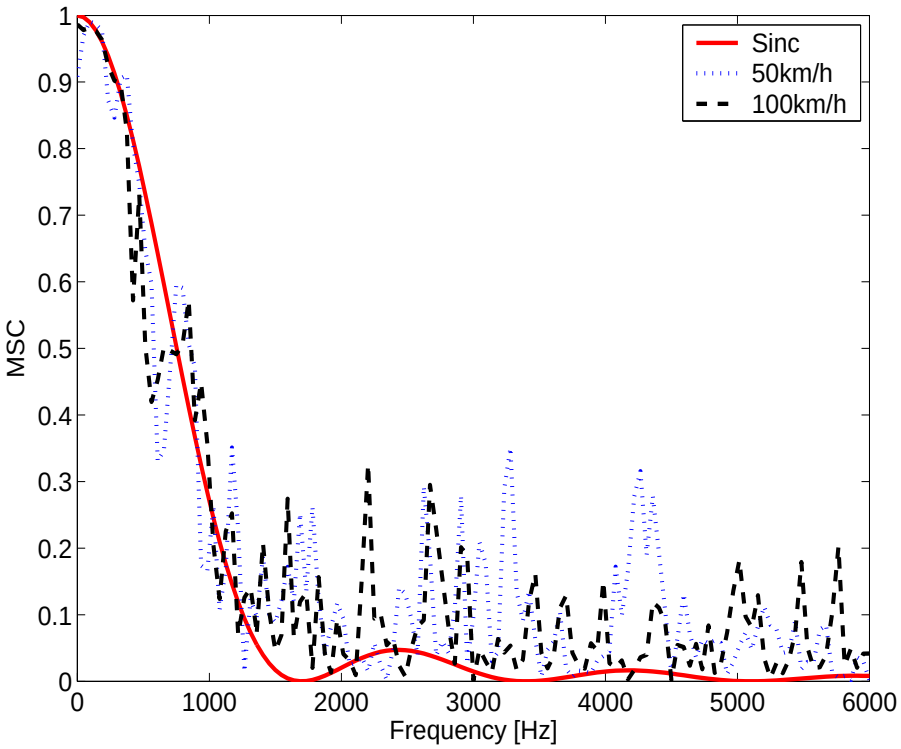


Figure 1: Magnitude-squared coherence function in car environment ($d = 10$cm).

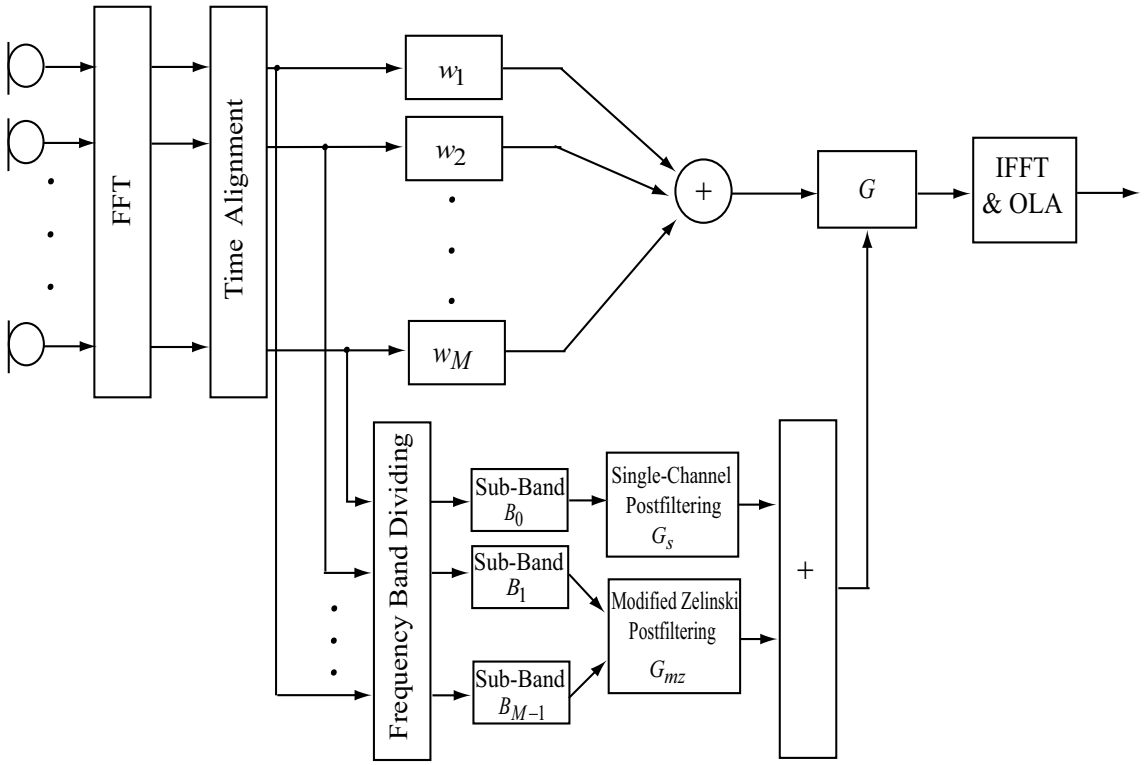


Figure 2: Block diagram of the proposed system.

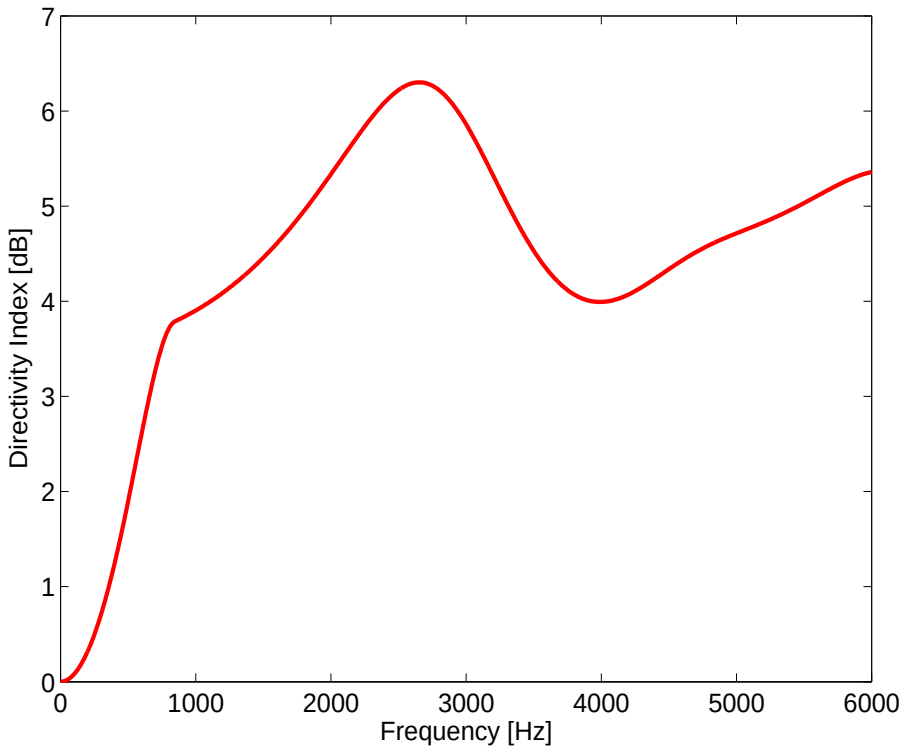
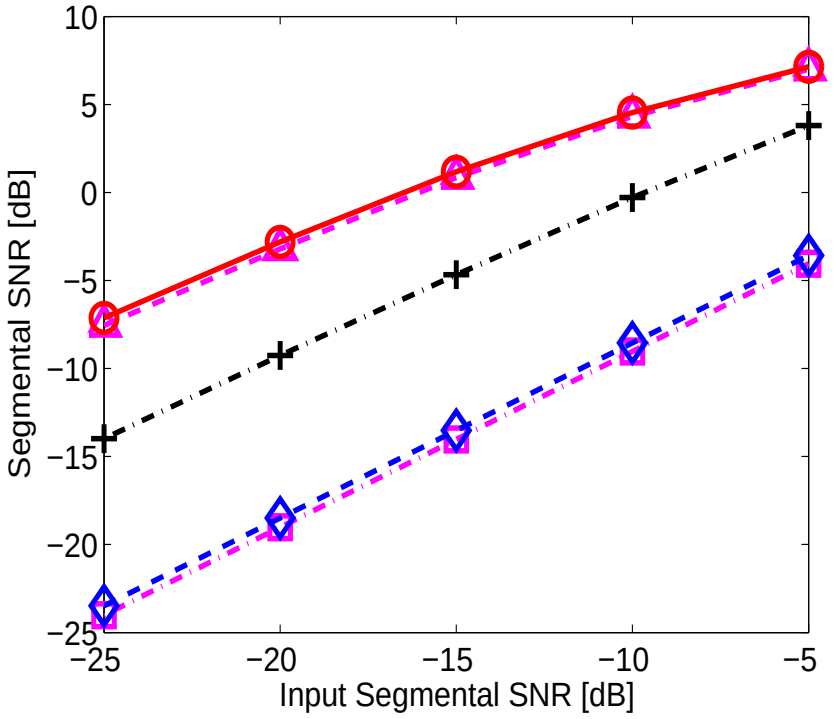
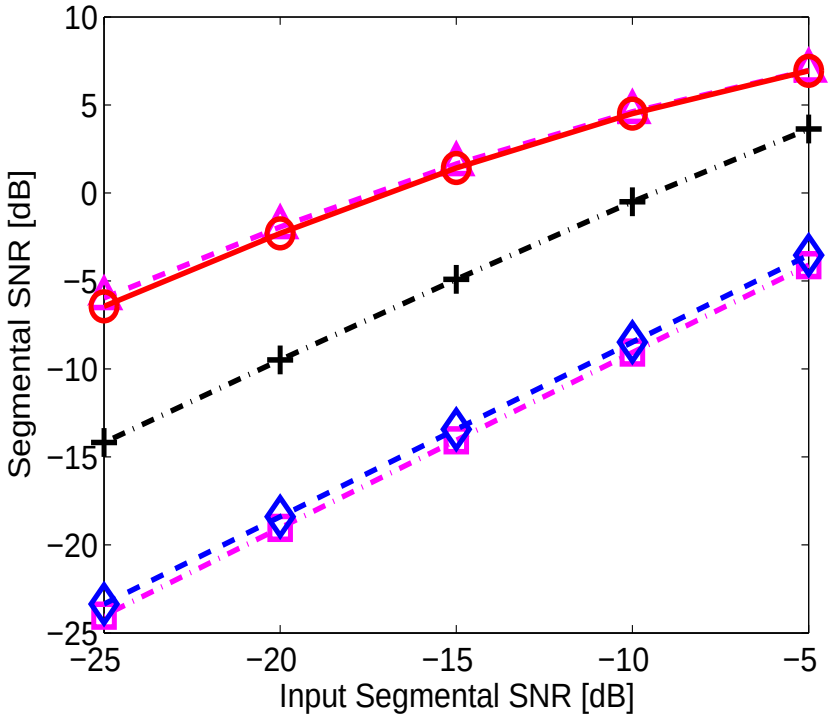


Figure 3: Directivity index of the superdirective beamformer (d=10cm).

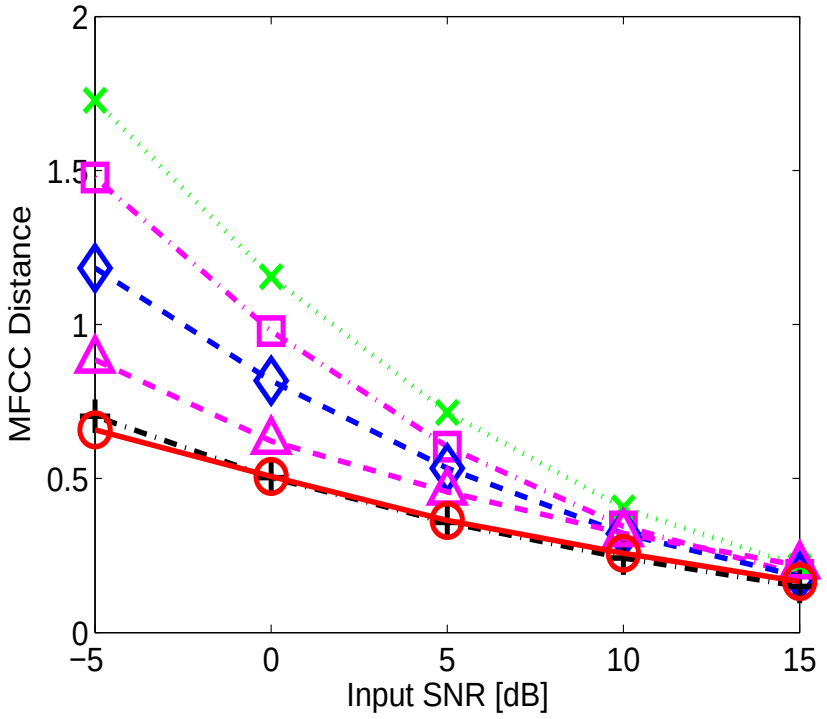


(a)

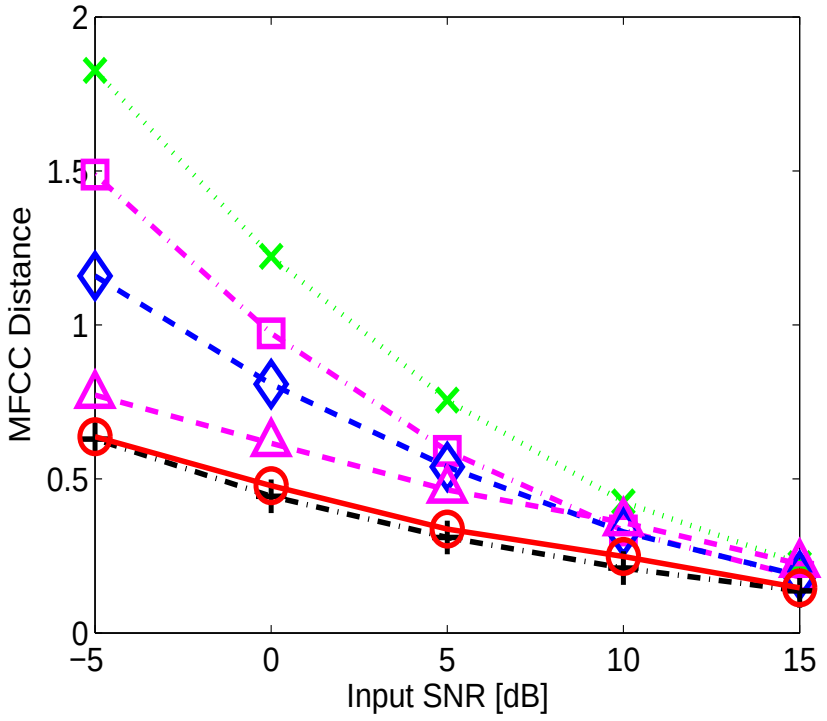


(b)

Figure 4: Average segmental SNR (SEGSNR) at beamformer output (□), Zelinski post-filter output (◊), McCowan post-filter output (+), single-channel Wiener filter output (△), proposed post-filter output(○), in various noise conditions: 50km/h (a) and 100km/h (b).



(a)



(b)

Figure 5: MFCC distance at the first microphone (×), beamformer output (□), Zelinski post-filter output(◊), McCowan post-filter output(+), single-channel Wiener filter output(△), proposed post-filter output(○), in various noise conditions: 50km/h (a) and 100km/h (b).

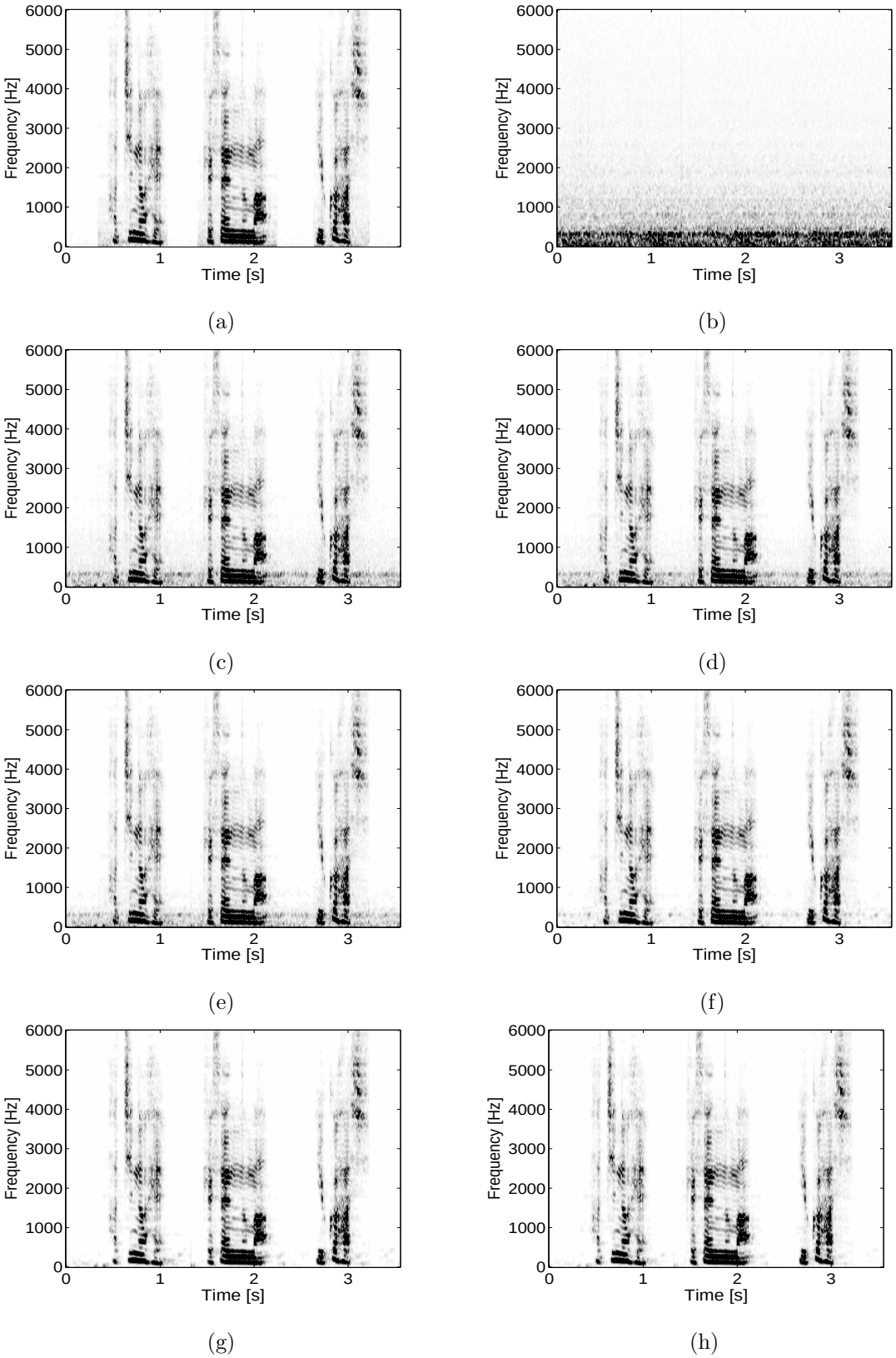


Figure 6: Speech spectrograms. (a) Original clean speech signal at the first microphone: "hatinohe kesennuma yukuhasi"; (b) Noise signal at the first microphone; (c) Noisy signal at the first microphone (SNR = 10 dB); (d) Beamformer output; (e) Zelinski post-filter output; (f) McCowan post-filter output; (g) Single-channel Wiener post-filter output; (h) Proposed post-filter output.

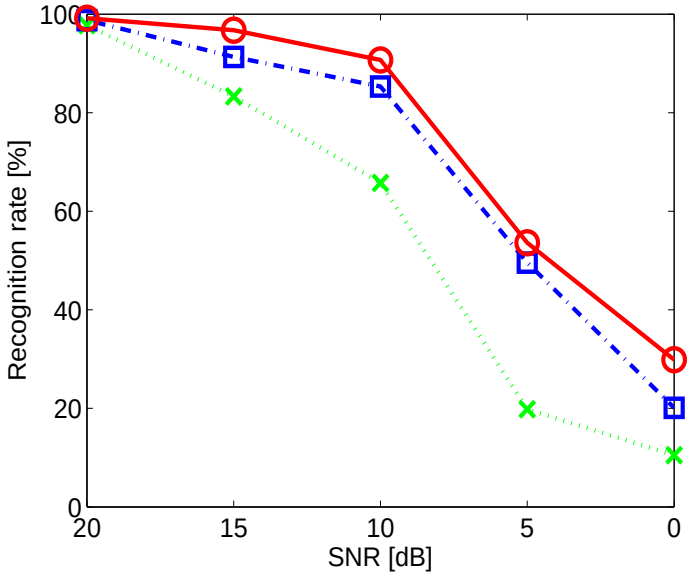


Figure 7: Speech recognition results of the noisy signal (×), the noise coherence based optimally-modified log-spectral amplitude (Coh-OM-LSA) estimator (□) and the proposed hybrid Wiener post-filter (○) in the car noise condition with the speed of 100km/h.

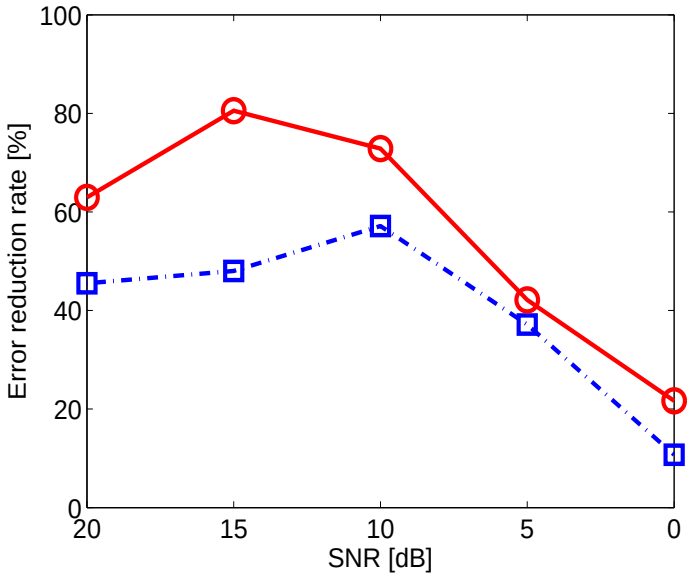


Figure 8: Recognition error reduction rates of the noise coherence based optimally-modified log-spectral amplitude (Coh-OM-LSA) estimator (□) and the proposed hybrid Wiener post-filter (○) in the car noise condition with the speed of 100km/h.