A Noise Reduction Method Based on Hybrid Noise Estimation Technique and Post-Filtering in Arbitrary Noise Environments

Junfeng Li and Masato Akagi
School of Information Science
Japan Advanced Institute of Science and Technology
1-1 Asahidai, Tatsunokuchi, Nomigun, Ishikawa, 923-1292, Japan
{junfeng,akagi}@jaist.ac.jp

Abstract: In this paper, we propose a novel noise reduction system using a hybrid noise estimation technique and post-filtering which is effective in dealing with localized and non-localized noise components. To suppress localized noises, we develop a hybrid noise estimation technique that combines a reformulated multi-channel estimation technique and a soft-decision single-channel estimation technique. Final estimation accuracy is significantly improved by a speech absence probability estimator which considers the strong correlations of speech presence uncertainty between adjacent frequency components and consecutive frames. The estimated spectra of localized noises are subtracted from those of noisy observations by spectral subtraction. The non-localized noises are further reduced by a multi-channel post-filter which minimizes the mean squared error of log-spectra. An estimator for the *a priori* speech absence probability is derived from the coherence characteristic of the noise field at spectral subtraction output, improving the spectral enhancement of the desired speech signal. Experimental results demonstrated that the hybrid estimation technique produces more accurate spectral estimates for localized noises and the proposed noise reduction system results in significant improvements in terms of objective speech quality measures: segmental SNR (SEGSNR) and Mel-Frequency Cepstral Coefficient (MFCC) distance and subjective speech quality measures: spectrograms and listening tests in various noise conditions.

Keywords: Hybrid Noise Estimation; Post-Filtering; Coherence Function; Speech Presence Uncertainty

1 Introduction

Recent years, noise reduction systems have been in great demand for an increasing number of speech applications, such as automatic speech recognition (ASR) systems and celluar telephony. Although the ASR systems have achieved high recognition accuracy in laboratory environments where the training conditions match the testing conditions to a large degree, their performance seriously degrades in the real-world environments where various kinds of noises result in a mismatch between training conditions and testing conditions for the ASR systems [1]. Adverse environments also deteriorate the quality of speech that is transmitted in speech communication systems [2]. One direct solution to this problem is to use a head-set or hand-held equipment, which is inconvenient for users. One potential solution is to construct a noise reduction system as a front-end processor for these systems.

So far, a variety of noise reduction algorithms have been proposed in the literature [2]-[10]. Generally speaking, all of these algorithms can be classified into two categories: single-channel techniques and multi-channel techniques according to the number of sensors they need. Compared to the single-channel technique, the multi-channel technique is substantially superior in reducing noise and enhancing speech, due to its spatial filtering capability of suppressing the interfering signals arriving from directions other than the specified look-direction [3]. Therefore, multi-channel noise reduction approaches have attracted increasing research interests.

The linearly constrained adaptive beamformer, first presented by Frost, keeps the signals arriving from the desired look-direction intact while suppressing the signals from other directions by minimizing the output power of the beamformer [6]. A generalized sidelobe canceller (GSC) beamformer, as an alternative implementation structure of Frost beamformer, has been widely researched [7]. In the GSC beamformer, adaptive signal processing is normally used to avoid cancellation of the desired speech signal [7]. However, adaptive signal processing decreases the stability of the noise reduction system under practical conditions. A small-scale subtractive beamformer-based noise reduction system has recently been proposed by Akagi *et al.* [9][10][11]. Its superiority lies in its high noise suppression capability, especially for sudden noises, with only small number of microphones and the analytical noise estimation scheme. And its weakness lies in the assumption that only localized noise components exist in the environment.

Among multi-channel noise reduction systems, post-filtering is normally needed to improve the entire performance in practical applications [14]. A multi-channel post-filter, first presented by Zelinski, is based on the auto- and cross- spectral densities of the input

signals, assuming zero cross-correlation between noise signals on different microphones [15]. Recently, it has been extended to a generalized expression based on *a priori* knowledge of the noise field [16]. Moreover, Bitzer *et al.* showed that neither GSC nor Wiener post-filter can work well at low frequencies in diffuse noise field [18][19]. An alternative solution, proposed by Meyer [20], applies spectral subtraction in low frequencies and a Wiener filter in high frequencies at the beamformer output. The two main drawbacks of these filtering approaches lie in their inability to suppress spatially correlated noises, and the required voice activity detector (VAD) or the *a priori* knowledge of the noise field.

In this paper, the problem of suppressing both localized and non-localized noises is dealt with by a novel noise reduction system based on a hybrid noise estimation technique and post-filtering. Based on a generalized signal model and a robust and accurate speech absence probability estimator, a hybrid noise estimation technique is presented, dealing with the inherent grating sidelobes of small size microphone arrays and the inability of single-channel techniques to estimate non-stationary noises. This hybrid estimation technique produces much more accurate spectral estimates for the localized noise components, which are then reduced by spectral subtraction. Presumably, the desired signals are strongly correlated in the frequencies of interest at different microphones and the non-localized noise signals are modelled as diffuse noises at spectral subtraction output. A novel estimator for the *a priori* speech absence probability is derived based on the coherence characteristic of the noise filed at spectral subtraction output. Subsequently, the residual non-localized noises are further suppressed by a multi-channel post-filter of which the gain function is improved by the *a priori* speech absence probability estimator. The performance of the proposed noise estimation/reduction systems is evaluated and compared to that of other conventional estimation/reduction systems in various noise conditions.

The remainder of this paper is organized as follows. In section 2, a generalized signal model is introduced along with an overview of the proposed noise reduction system. In section 3, the localized noises are first estimated by a hybrid estimation technique and are subsequently subtracted from observations. In section 4, an estimator for the *a priori* speech absence probability is presented which improves the suppression capability for the non-localized noises of the multi-channel post-filter. The superiorities of the proposed noise estimation/reduction systems are verified in various noise conditions in section 5. Finally, some conclusions are drawn in section 6.

2 An Overview of the Proposed Noise Reduction System

In this section, a generalized signal model is introduced and an overview of the proposed noise reduction system is given.

Considering a microphone array with three linearly and equidistant distributed omni-directional microphones in a noisy environment, shown in Fig. 1, a generalized signal model is assumed in which the observed signals consist of three components. The first is the desired speech signal arriving from a direction such that the difference in arrival time between the two main microphones is 2ζ. The second is localized noise signals arriving from directions such that the time differences are $2\delta_k$ $(k = 1, 2, ..., K)$ and the third is non-localized noise signal, modelled as diffuse noise, propagating in all directions simultaneously. Thus, the observed signals imposing on three microphones (left, center and right), denoted by $l(t)$, $c(t)$ and $r(t)$, can be given by:

$$l(t) = s(t-\zeta) + \sum_{k=1}^{K} n_k^c(t-\delta_k) + n_l^{uc}(t), \quad (1)$$

$$c(t) = s(t) + \sum_{k=1}^{K} n_k^c(t) + n_c^{uc}(t), \quad (2)$$

$$r(t) = s(t+\zeta) + \sum_{k=1}^{K} n_k^c(t+\delta_k) + n_r^{uc}(t), \quad (3)$$

where n_k^c and $n_\cdot^{uc}$ denote k-th localized noise signal and non-localized noise signal, respectively.

Based on this generalized signal model, the intention of our work is to reduce both localized and non-localized noises while keeping the desired signal distortionless. To implement this idea, a noise reduction system is constructed as shown in Fig. 2, which mainly consists of the following modules:

- Time delay compensation: This module compensates for the effect of propagation between speech source and microphones on the desired speech signal. It is assumed that the output signals of the module are perfectly time aligned and represented by the same symbols, only setting $\zeta = 0$ in Eqs. (1)-(3), for notational simplicity.

- Spectral analysis: The time aligned signals are analyzed by Short Time Fourier Transform (STFT), outputting the amplitude spectra and phase spectra.

- Localized noise suppression: To suppress localized noise signals, their spectra are first estimated by a hybrid noise estimation approach which effectively combines

a multi-channel technique and a single-channel technique in a parallel structure. For the multi-channel technique, an estimation approach based on a reformulated subtractive beamformer is adopted. For the single-channel technique, a soft-decision based approach is used. Moreover, considering the strong correlations of speech presence uncertainty between adjacent frequency components and consecutive frames, we derive a robust and accurate speech absence probability estimator, which greatly improves the estimation accuracy of the hybrid estimation technique for localized noises. The final spectral estimates are subtracted from those of noisy observations by spectral subtraction.

- Non-localized noise suppression: To further suppress the residual non-localized diffuse noise, a multi-channel post-filter, minimizing the mean square error of the log-spectra, is adopted. Moreover, an estimator for the *a priori* speech absence probability is derived based on the coherence characteristic of the noise field at spectral subtraction output. The *a priori* speech absence probability further enhances the noise suppression performance of the multi-channel post-filter.

- Spectral synthesis: The enhanced speech signal is synthesized by using inverse STFT and the overlap-and-add technique.

The two main modules mentioned above, localized noise suppression and non-localized noise suppression, will be discussed in detail in sections 3 and 4.

3 Localized Noise Suppression Using a Hybrid Noise Estimation Technique and Spectral Subtraction

In this section, we focus on suppressing localized noise components. To do this, we first propose a hybrid estimation technique to obtain the spectral estimates of localized noises which are then subtracted from those of noisy observations.

3.1 A Hybrid Noise Estimation Technique

A hybrid noise estimation technique which combines a reformulated multi-channel estimation approach and a single-channel estimation approach is described in this subsection.

3.1.1 The Multi-Channel Noise Estimation Approach

Based on the generalized signal model, the multi-channel noise estimation approach we previously proposed can be reformulated as follows.

The time-aligned signals $l(t)$, $c(t)$ and $r(t)$ are shifted $\pm\tau$ in the time domain ($\tau \neq 0$), and two subtractive beamformers in the time domain are constructed as [11]:

$$g_{lr}(t) = \frac{1}{4}\{[l(t+\tau) - l(t-\tau)] - [r(t+\tau) - r(t-\tau)]\}, \quad (4)$$

$$g_{cr}(t) = \frac{1}{4}\{[c(t+\tau) - c(t-\tau)] - [r(t+\tau) - r(t-\tau)]\}. \quad (5)$$

In order to simplify implementation, the differences of non-localized noises at different microphones are assumed to be small enough to be ignored. Then, the two beamformers in the frequency domain can be derived as [11]:

$$G_{lr}(\lambda,\omega) = \sin\omega\tau N^c(\lambda,\omega)\sin\omega\delta, \quad \omega_{i-1} \leq \omega < \omega_i, \quad (6)$$

$$G_{cr}(\lambda,\omega) = \sin\omega\tau N^c(\lambda,\omega)e^{j\omega\frac{\delta}{2}}\sin\omega\frac{\delta}{2}, \quad \omega_{i-1} \leq \omega < \omega_i, \quad (7)$$

$$(i = 1, 2, ..., I, \omega_0 = 0, \omega_I = \pi,)$$

where (i) $N^c(\lambda,\omega)$ and δ denote the spectrum and the virtual direction of arrival (DOA) of the integrated localized noise signal in sub-band ($\omega_{i-1} \leq \omega < \omega_i$), respectively; (ii) λ and ω are the frame index and the frequency index. Note that the outputs of two subtractive beamformers do not contain any desired speech signal components which have been blocked successfully. Given the virtual DOA of the integrated noise signal, the spectra of localized noise can be easily estimated from the outputs of the beamformers, represented as ($\tau = \delta$) [11]:

$$\hat{N}_m^c(\lambda,\omega) = \begin{cases} G_{lr}(\lambda,\omega)/\sin^2\omega\delta, & \sin^2\omega\delta > \varepsilon_1 \ , \\ G_{cr}(\lambda,\omega)e^{-j\omega\frac{\delta}{2}}/\sin^2\omega\frac{\delta}{2}, & \sin^2\omega\delta < \varepsilon_1 \text{ and } \sin^2\omega\frac{\delta}{2} > \varepsilon_2, \\ G_{cr}(\lambda,\omega)/\varepsilon_2, & \text{otherwise}, \end{cases} \quad (8)$$

where (i) the subscript m in $\hat{N}_m^c(\lambda,\omega)$ indicates the spectral values are estimated by this multi-channel technique; (ii) ε_1 and ε_2 are two threshold values determined experimentally ($\varepsilon_1 = 0.5$ and $\varepsilon_2 = 0.1$ in this work). And the virtual direction of the integrated localized noise signal can be computed from the outputs of two beamformers by the algorithm in [11].

3.1.2 Analysis and Improvement of the Multi-channel Noise Estimation Approach

The multi-channel noise estimation approach has a great ability to estimate the spectra for localized noises arriving from determinable directions. However, when the condition $\omega\delta = 2k\pi$ holds, the localized noise spectra can not be accurately estimated since the beamformers do not output any signal. In this case, estimation accuracy of this multi-channel approach was degraded since an approximation was used in [11]. This phenomenon corresponds to the grating sidelobes of the microphone arrays with small physical size.

To deal with this problem, we propose a hybrid noise estimation technique in which a single-channel estimation approach is incorporated when the multi-channel estimation approach fails. In this hybrid estimation technique, the values of $\sin^2\omega\delta$ and $\sin^2\omega\frac{\delta}{2}$, in Eq. (8), determine whether the output of the single-channel approach or that of the multi-channel approach should be the final output of this hybrid structure. When the maximum of $\sin^2\omega\delta$ and $\sin^2\omega\frac{\delta}{2}$ is larger than a threshold ε (an empirical constant), the output of the multi-channel approach is more accurate and preferable as the final output. Otherwise, the output of the single-channel approach is more accurate and preferable as the final output. Thus, the final spectral estimates of the localized noises by the proposed hybrid estimation technique can be given by:

$$\hat{N}^c(\lambda,\omega) = \begin{cases} \hat{N}^c_m(\lambda,\omega), & \max\left(\sin^2\omega\delta, \sin^2\omega\frac{\delta}{2}\right) > \varepsilon, \\ \hat{N}^c_s(\lambda,\omega), & \text{otherwise}, \end{cases} \quad (\omega_{i-1} \le w < \omega_i, i = 1, 2, ..., I) \tag{9}$$

where $\hat{N}^c_m(\lambda,\omega)$ and $\hat{N}^c_s(\lambda,\omega)$ represent the spectral estimates of localized noises by the multi-channel technique, given by Eq. (8), and by the single-channel technique detailed in the following subsection.

3.1.3 The Single-Channel Noise Estimation Approach

As Eq. (9) shows, the spectra of localized noises, $\hat{N}^c_s(\lambda,\omega)$, should be computed by a single-channel estimation approach. In this work, a soft-decision single-channel approach is adopted since it can adaptively update the noise spectra, given by:

$$\begin{aligned} |\hat{N}^c_s(\lambda,\omega)|^2 &= \eta_n(\lambda,\omega) \\ &= \alpha_\eta\eta_n(\lambda-1,\omega) + (1-\alpha_\eta)E\left[|N(\lambda,\omega)|^2|C(\lambda,\omega)\right], \end{aligned} \tag{10}$$

where (i) $C(\lambda,\omega)$ denotes the STFT of $c(t)$ in Eq. (2); (ii) $\eta_n(\lambda,\omega)$=$E\left[|N(\lambda,\omega)|^2\right]$ is the variance of the noise signal; (iii) α_η $(0 < \alpha_\eta < 1)$ is a forgetting factor controlling the

update rate of noise estimation. Under the speech presence uncertainty, the second term in the right side of Eq. (10) can be estimated as the spectra of observed signals during speech pauses or it can hold the values obtained in the previous pauses during speech active periods, given by:

$$E\left[|N(\lambda,\omega)|^2|C(\lambda,\omega)\right] = q(\lambda,\omega)|C(\lambda,\omega)|^2+(1-q(\lambda,\omega))\,\eta_n(\lambda-1,\omega), \quad (11)$$

where $q(\lambda,\omega)$ denotes the speech absence probability. As Eq. (11) shows, the spectral estimation capability of the single-channel approach can be significantly enhanced through a speech absence probability estimator detailed in the following subsection, improving the final estimation accuracy of the hybrid technique.

3.1.4 A robust and accurate speech absence probability estimator

Under the assumption of complex Gaussian statistic model and applying the Bayes rule and total probability theorem, the speech absence probability conditioned on the observations can be given by [24]:

$$q(\lambda,\omega) = \left(1+\frac{1-q^{'}(\lambda,\omega)}{q^{'}(\lambda,\omega)}\frac{1}{1+\xi(\lambda,\omega)}\exp\left(\frac{\xi(\lambda,\omega)\gamma(\lambda,\omega)}{1+\gamma(\lambda,\omega)}\right)\right)^{-1}, \quad (12)$$

where (i) $q^{'}(\lambda,\omega)$ is the *a priori* speech absence probability; (ii) $\xi(\lambda,\omega)=\eta_s(\lambda,\omega)/\eta_n(\lambda,\omega)$ and $\gamma(\lambda,\omega)=|C(\lambda,\omega)|^2/\eta_n(\lambda,\omega)$ are the *a priori* SNR and *a posteriori* SNR, as named in [24], and $\eta_s(\lambda,\omega)$ represents the variance of speech signal.

Eq. (12) demonstrates that for the given *a priori* speech absence probability $q^{'}(\lambda,\omega)$, the speech absence probability $q(\lambda,\omega)$ is greatly dependent on the *a priori* SNR $\xi(\lambda,\omega)$ and *a posteriori* SNR $\gamma(\lambda,\omega)$. It is believed that accurate and robust speech absence probability estimates can be obtained only when $\xi(\lambda,\omega)$ and $\gamma(\lambda,\omega)$ are accurate and robust enough. Consequently, we now turn to the issue of improving the accuracy and robustness of the *a priori* SNR $\xi(\lambda,\omega)$ and *a posteriori* SNR $\gamma(\lambda,\omega)$ estimates.

Taking into account the strong correlation of speech presence uncertainty in adjacent frequency bins and consecutive frames, we propose two estimators for the *a priori* SNR $\xi(\lambda,\omega)$ and *a posteriori* SNR $\gamma(\lambda,\omega)$.

- In the frequency domain, the estimates of $\xi(\lambda,\omega)$ and $\gamma(\lambda,\omega)$ are smoothed by applying a normalized window b of size $2D+1$, given by:

$$\tilde{\xi}(\lambda,\omega) = \sum_{k=w-D}^{k=w+D} b(k)\xi(\lambda,k), \quad (13)$$

$$\tilde{\gamma}(\lambda,\omega) = \sum_{k=w-D}^{k=w+D} b(k)\gamma(\lambda,k). \quad (14)$$

Estimation accuracy of $\xi(\lambda,\omega)$ and $\gamma(\lambda,\omega)$ is improved due to the fact the noise spectra in adjacent frequencies are likely to be estimated by the multi-channel estimation technique with high accuracy. Furthermore, this frequency-smoothing procedure eliminates fluctuations of the *a priori* SNR $\xi(\lambda,\omega)$ and *a posteriori* SNR $\gamma(\lambda,\omega)$ along the frequency axis on the time-frequency plane, which results in more robust SNRs estimates.

- In the time domain, the frequency-smoothed estimates of the *a priori* SNR $\tilde{\xi}(\lambda,\omega)$ and *a posteriori* SNR $\tilde{\gamma}(\lambda,\omega)$ are further processed based on the previous values, given by:

$$\bar{\xi}(\lambda,\omega) = \alpha_\xi \frac{\tilde{S}^2(\lambda-1,\omega)}{\eta_n(\lambda-1,\omega)} + (1-\alpha_\xi)\max[\tilde{\gamma}(\lambda,\omega)-1,0], \quad (15)$$

$$\bar{\gamma}(\lambda,\omega) = \tilde{\gamma}(\lambda,\omega), \quad (16)$$

where α_ξ $(0<\alpha_\xi<1)$ is a forgetting factor and $\tilde{S}(\lambda-1,\omega)$ is the enhanced signal by spectral subtraction in the previous frame, given by Eq. (18). Actually, Eq. (15) is just the decision-directed scheme detailed in [24]. It is of interest to note that the smoothing operation in the time domain is not carried out for the *a posteriori* SNR, since it should be calculated from the current observations and independent on the previous observations.

Based on the time-frequency smoothed *a priori* SNR $\bar{\xi}(\lambda,\omega)$ and *a posteriori* SNR $\bar{\gamma}(\lambda,\omega)$, the speech absence probability $q(\lambda,\omega)$ can be obtained as:

$$q(\lambda,\omega) = \left(1+\frac{1-q^{'}(\lambda,\omega)}{q^{'}(\lambda,\omega)}\frac{1}{1+\bar{\xi}(\lambda,\omega)}\exp\left(\frac{\bar{\xi}(\lambda,\omega)\bar{\gamma}(\lambda,\omega)}{1+\bar{\gamma}(\lambda,\omega)}\right)\right)^{-1}, \quad (17)$$

where $q^{'}(\lambda,\omega)$ is given by: $q^{'}(\lambda,\omega)=1-P_{local}(\lambda,\omega)P_{global}(\lambda,\omega)P_{frame}(\lambda,\omega)$. And $P_{local}(\lambda,\omega)$, $P_{global}(\lambda,\omega)$ and $P_{frame}(\lambda,\omega)$ correspond to the speech presence probabilities which are estimated based on the speech energy distribution in a local frequency window, a larger frequency window and neighboring frames in the time-frequency domain respectively, detailed in [26].

It should be pointed out that high spectral estimation accuracy of the proposed hybrid estimation technique is ensured by the good combination of the multi-channel estimation approach and the single-channel estimation approach from several aspects. Firstly, as shown in Eq. (9), the final spectral estimates are computed by the multi-channel approach, which produces the real spectral estimates, in most cases. Secondly, estimation accuracy of the single-channel approach is significantly improved, which is attributed to the multi-channel estimation approach and the speech absence probability estimator. Since accurate

spectral estimates by the multi-channel approach are likely to be distributed around those determined by the single-channel approach, accuracy and robustness of the speech absence probability estimator can be ensured by applying the time-frequency smoothed *a priori* SNR and *a posteriori* SNR which are more accurate. Furthermore, the accurate and robust speech absence probability produces more accurate spectral estimates by the single-channel estimation approach, improving the accuracy by final spectral estimates by the proposed hybrid noise estimation technique. Thus, to a certain degree, the hybrid estimation technique deals with the inherent grating sidelobes of small-size microphone arrays and the inability of single-channel approach to estimate the non-stationary noises.

3.2 Suppress Localized Noise With Spectral Subtraction

The proposed hybrid noise estimation technique gives accurate spectral estimates for localized noises. Subsequently, the estimated spectra are subtracted from those of the observed noisy signals by the spectral subtraction method, given by [5]:

$$\tilde{S}(\lambda,\omega) \quad = \quad \begin{cases} C(\lambda,\omega) - \alpha\hat{N}(\lambda,\omega), & C(\lambda,\omega) > \alpha\hat{N}(\lambda,\omega), \\ \beta C(\lambda,\omega), & \text{otherwise}, \end{cases} \tag{18}$$

where α and β are the overestimation factor and spectral floor factor. In our case, since the spectral estimates of localized noises are of high accuracy, $\alpha = 1$ is set to avoid distorting the speech signal. And β is determined experimentally.

4 Non-localized Noises Suppression With Post-Filtering

In this section, we address the problem of suppressing the residual non-localized noises by employing a multi-channel post-filter. In our proposed noise reduction system, the well-known *optimally-modified log-spectral amplitude* (OM-LSA) estimator is adopted as post-filter due to: its superiority in eliminating "musical noises"[26] resulting from the spectra subtraction given by Eq. (18). An estimator for the *a priori* speech absence probability is derived based on the coherence characteristic of the noise field at spectral subtraction output. The further spectral enhancement of the desired speech signal is achieved by modifying the gain function of the post-filter with the newly obtained *a priori* speech absence probability estimator.

4.1 OM-LSA Estimator

The basic idea of the OM-LSA estimator is to minimize the mean square error between the log-spectra of the desired speech signals and their optimal estimates. Under speech presence uncertainty and based on the assumption of complex Gaussian statistic model, the OM-LSA estimator is given by [26]:

$$G(\lambda,\omega) \;=\; G_{H_1}(\lambda,\omega)^{1-q_p(\lambda,\omega)} G_{\min}^{q_p(\lambda,\omega)}, \tag{19}$$

where (i) $G_{\min}$ is an empirical constraint constant; (ii) $q_p(\lambda,\omega)$ is the speech absence probability (different from $q(\lambda,\omega)$) which is calculated at spectral subtraction output. (iii) $G_{H_1}(\lambda,\omega)$ is the gain function of the traditional MMSE-LSA estimator when speech is surely present, defined by [25]:

$$G_{H_1}(\lambda,\omega) \;=\; \frac{\xi_p(\lambda,\omega)}{1+\xi_p(\lambda,\omega)} \exp\left(\frac{1}{2}\int_{v_p(\lambda,\omega)}^{\infty} \frac{e^{-t}}{t} dt\right), \tag{20}$$

where (i) $v_p(\lambda,\omega) = \gamma_p(\lambda,\omega)\xi_p(\lambda,\omega)/(1+\xi_p(\lambda,\omega))$; (ii) $\xi_p(\lambda,\omega)$ and $\gamma_p(\lambda,\omega)$ are the *a priori* SNR and *a posteriori* SNR that are computed at spectral subtraction output.

It is of interest to note that the speech absence probability $q_p(\lambda,\omega)$, in Eq. (19), should be calculated again with Eq. (17) at spectral subtraction output. And, a novel estimator for the *a priori* speech absence probability $q_p^{'}(\lambda,\omega)$ should be designed, since the energy distribution of the spectral subtraction output has been changed which results in the failure of the energy-based scheme in [26].

4.2 Coherence Function of Noise Fields

To characterize the noise field, a widely used measure is the *magnitude-squared coherence* (MSC), more commonly referred to as coherence, defined as:

$$\Gamma_{xy}(\lambda,\omega) \;=\; \frac{|\phi_{xy}(\lambda,\omega)|^2}{\phi_{xx}(\lambda,\omega)\phi_{yy}(\lambda,\omega)}, \tag{21}$$

where (i) $\phi_{xy}(\lambda,\omega)$ is the cross-spectral density between two signals $x(t)$ and $y(t)$; (ii) $\phi_{xx}(\lambda,\omega)$ and $\phi_{yy}(\lambda,\omega)$ are the auto-spectral densities of $x(t)$ and $y(t)$, respectively. The auto- and cross-spectral densities of the signals, $\phi_{xx}(\lambda,\omega)$ and $\phi_{xy}(\lambda,\omega)$, can be estimated in a recursive way as:

$$\phi_{xx}(\lambda,\omega) \;=\; \alpha_\phi \phi_{xx}(\lambda-1,\omega) + (1-\alpha_\phi) X(\lambda,\omega) X^*(\lambda,\omega), \tag{22}$$

$$\phi_{xy}(\lambda,\omega) \;=\; \alpha_\phi \phi_{xy}(\lambda-1,\omega) + (1-\alpha_\phi) X(\lambda,\omega) Y^*(\lambda,\omega), \tag{23}$$

where α_ϕ $(0<\alpha_\phi<1)$ is a forgetting factor controlling the update of auto- and cross-spectral densities.

In the generalized signal model given by Eqs. (1)-(3), non-localized noises are modelled as diffuse noise characterized by the following MSC function:

$$\Gamma(\omega) \quad = \quad \left| \frac{\sin\left(\omega \mathrm{d}/\upsilon\right)}{\omega d/\upsilon} \right|^2, \tag{24}$$

where d and υ are the distance between microphones and the velocity of sound.

Fig. 3 illustrates the MSCs computed with the noise signals recorded in a car environment and the theoretical MSC of diffuse noise field. Obviously, the measured MSCs follow the trend of the theoretical MSC of the diffuse noise field with some relative variances. Furthermore, at the output of spectral subtraction, the diffuse characteristic of non-localized noises is kept, since the hybrid estimation technique based noise suppression algorithm aims to reduce only localized noises with little influence on non-localized noises. This fact is confirmed by the MSC functions, shown in Fig. 4, computed with the input signals and the output signals of spectral subtraction when only the diffuse noises are feed into the system.

4.3 A Priori Speech Absence Probability Estimator

This subsection deals with the problem of detecting the desired speech signal at spectral subtraction output based on the MSC function computed with the output signals.

Fig. 5 shows an example of the average MSC value over all frequencies at spectral subtraction output when the speech signals are spatially strong correlated and the noises are spatially weak correlated. It is obvious to know that the MSC can be used to detect speech signal due to its capability in discriminating the desired speech signals and interfering noises. This fact has been used under the assumption that the noises are spatially uncorrelated over all frequencies at different microphones [27]. A novel estimator for the *a priori* speech absence probability is designed at spectral subtraction output based on diffuse noise field assumption in the following.

Let us first give some assumptions upon which the *a priori* speech absence probability estimator is based:

- The desired speech signal and interfering noises are statistically independent;
- The desired speech signal is strong correlated over all frequencies of interest at different microphones;
- The residual non-localized noises at the output of the spectral subtraction are modelled as diffuse noises.

At spectral subtraction output, the MSC function $\Gamma(\lambda,\omega)$, is first calculated based on the output signals which are obtained by subtracting the estimated localized noise spectra from those of noisy observations on left, center and right microphones, as shown in Eq. (18). The output signals at different frequencies are characterized by different coherence values, as shown in Fig. 4. This observation motivates that the MSC spectra at spectral subtraction output are divided into two parts: the low frequency region with high noise coherence and the high frequency region with low noise coherence. The transient frequency between two regions is the first minimum frequency of the MSC function of diffuse noise field, given by $f = \upsilon/(2d)$, where d and υ are the distance between adjacent microphones and sound velocity.

To determine the *a priori* speech absence probability in the high frequency region and in the low frequency region, we proposed two different schemes described as follows:

- In the high frequency region, the MSC spectra are divided into E sub-bands with a feasible bandwidth, BW, and are averaged over the frequencies in each sub-band, obtaining the averaged MSC $\bar{\Gamma}_e(\lambda,\omega)(e = 1, 2, ..., E)$ in e-th sub-band. The averaged MSC $\bar{\Gamma}_e(\lambda,\omega)$ provides an informative clue for determining the *a priori* speech absence probability in e-th sub-band. If a high coherence (higher than a threshold $T\max_e$) is detected, a speech present state is detected presumably. If a low coherence (lower than a threshold $T\min_e$) is detected, a speech absent state is detected presumably. Note that the *a priori* speech absence probability decreases as the MSC at spectral subtraction output increases. For the MSC values in the range $[T\min_e, T\max_e]$, the *a priori* speech absence probabilities are derived by interpolation between 1 and 0. Thus, the *a priori* speech absence probability in the high frequency region, $q'_{p,h}(\lambda,\omega)$, is given by:

$$q'_{p,h}(\lambda,\omega) = \begin{cases} 0, & \bar{\Gamma}_e(\lambda,\omega) > T\max_e \ , \\ 1, & \bar{\Gamma}_e(\lambda,\omega) < T\min_e \ , \ \omega \in [\omega_e^{low}, \omega_e^{high}], \\ \frac{T\max_e - \bar{\Gamma}_e(\lambda,\omega)}{T\max_e - T\min_e}, & \text{otherwise}, \end{cases} \quad (25)$$

  where ω_e^{low} and ω_e^{high} are the low and high boundaries of e-th sub-band.

- In the low frequency region, the MSC spectra computed with the output signals in the low frequency region fail to discriminate the speech signal and interfering diffuse noises, since both of them are strongly correlated. Based on the MSC value $\bar{\Gamma}(\lambda,\omega)$ computed and averaged over the frequencies higher than the transient frequency and

Table 1: parameters set in the experiments

$M = 256$	ε_1=0.5	ε_2 =0.1	ε=0.1	D=2	$I = 32$
α_ξ =0.98	$\alpha = 1.0$	β=0.001	BW =32	α_ϕ =0.7	Θ=12
$T\max_e$=0.7	$T\min_e$=0.2	$T\max$ =0.7	$T\min$ =0.2	$G_{\min}$ =0.005	$\upsilon = 340m/s$

following the same concept used in high frequency region, we derive an estimator for the *a priori* speech absence probability in the low frequency region, $q'_{p,l}(\lambda,\omega)$, given by:

$$q'_{p,l}(\lambda,\omega) = \begin{cases} 0, & \bar{\Gamma}(\lambda,\omega) > T\max\ , \\ 1, & \bar{\Gamma}(\lambda,\omega) < T\min\ , \\ \frac{T\max-\bar{\Gamma}(\lambda,\omega)}{T\max-T\min}, & \text{otherwise,} \end{cases} \tag{26}$$

and

$$\bar{\Gamma}(\lambda,\omega) = \frac{1}{E}\sum_{e=1}^{E}\bar{\Gamma}_e(\lambda,\omega). \tag{27}$$

The *a priori* speech absence probability estimates, given by Eqs. (25) and (26), are then incorporated into the post-filtering, further improving the spectral enhancement of the desired speech signal.

5 Experiments and Discussions

In this section, two sets of experiments were conducted to evaluate and compare the performance of the proposed noise estimation/reduction methods with that of other traditional noise estimation/reduction methods. The first set of experiments was devoted to show the pure contribution of the hybrid noise estimation technique. The second set was to evaluate the performance of the proposed noise reduction approach as a complete system, compared to several conventional noise reduction systems, in the simulated and real-world noise environments.

All parameters set in our experiments are listed in Table 1.

5.1 Evaluation of The Hybrid Noise Estimation Approach

In the first set of experiments, we concentrated on the improvement in estimation accuracy of the proposed hybrid noise estimation technique for localized noise signals compared to the corresponding single-channel and multi-channel estimation techniques.

5.1.1 Sound Data

To objectively evaluate the performance of the proposed hybrid noise estimation technique, 54 clean speech sentences, selected from ATR database and uttered by 3 male and 3 female speakers, were used. The tested noises consisted of synthesized noises (white Gaussian noise and pink noise) and real-world car noise. The speech data and noise data were first resampled to 12 kHz and linearly quantized at 16 bits. The noisy signals were generated by mixing the clean speech signals with the directional tested noises with directions of arrival (DOAs) 10-80 degrees to the right.

5.1.2 Evaluation Measure

The performance of the proposed hybrid noise estimation technique was evaluated and compared to the corresponding single-channel and multi-channel techniques in terms of Normalized Estimation Error (NEE), defined as:

$$\text{NEE} = \frac{1}{L}\sum_{\lambda=1}^{L} 20\log_{10}\left(\frac{\sum_{\omega=0}^{M-1}\left(\left|\hat{N}^c(\lambda,\omega) - N^c(\lambda,\omega)\right|\right)}{\sum_{\omega=0}^{M-1} N^c(\lambda,\omega)}\right), \tag{28}$$

where $\hat{N}^c(\lambda,\omega)$ and $N^c(\lambda,\omega)$ are the estimated noise spectrum and "ideal" noise spectrum respectively; M and L are the length of STFT and the number of frames. It should be noted that the smaller NEE represents the more accurate noise estimate obtained by the tested estimation technique.

5.1.3 Evaluation Results

The average NEEs over the noise signals with different DOAs in the tested noise conditions are listed in Table 2. Table 2 demonstrates that the normalized noise estimation error is consistently decreased for all the tested noise conditions, especially for directional car noise, when the proposed hybrid noise estimation technique is used. This improvement amounts to 3 dB compared to the single-channel estimation technique alone and 5 dB compared to the multi-channel estimation technique alone in localized car noise environment. Fig. 6 illustrates typical examples of the NEE comparisons of the single-channel, multi-channel and hybrid noise estimation techniques in localized white and car noise environments. All the observations obtained from Table 2 and Fig. 6 verify the superiority of our proposed hybrid noise estimation technique compared to the multi-channel estimation alone technique and the signal-channel estimation alone technique.

Table 2: Average NEEs in various noise conditions

	white	pink	car
single-channel	-5.2842	-4.6423,	-6.7910
multi-channel	-12.8905	-10.2486,	-4.5205
hybrid	-13.3378	-11.1818	-9.7014

5.2 Evaluation of The Noise Suppression System

The second set of experiments was conducted to evaluate the performance of the proposed noise reduction method as a complete system in the simulated and real-world noise environments. Furthermore, its performance was compared with the performance of other conventional noise reduction systems, including Delay-and-Sum Beamformer, Delay-and-Sum Beamformer with multi-channel Wiener post-filter [28], single-channel OM-LSA estimator [26] and the subtractive beamformer alone system [11], under various noise conditions in terms of objective and subjective evaluation measures.

5.2.1 Sound Data

In the simulated noise environment, the input noise signals consisted of a localized noise signal (directional car noise with DOA of 40 degrees to the right) and a non-localized noise signal (diffuse noise). The diffuse noise field was generated by placing 18 independent pink noise sources around the microphone array. The noisy data were obtained by adding the recorded input noise signals to the clean speech signals which were same as those used in the first set of experiments at various global SNR levels in the range [-5, 20] dB.

In the real-world noise environment, an equi-spaced linear array, consisting of three microphones with inter-element interval of 10cm, was mounted above the windshield in the car. Car noise signals were recorded while the car was running on the highway at speed of 50 km/h and 100 km/h with the air-condition on. Speech data were same as those used in first set of experiments selected from ATR database. The first set of input microphone signals were generated by mixing the clean speech signals and car noise signals at various global SNR levels in the range [-5, 20] dB. The second set of input noisy signals was generated by mixing the speech signal, car noises with the speed of 100 km/h and another interfering speech voice (localized noise) with DOA of 60 degree to the right. This interfering speech voice was used to imitate the passenger's voice in car environment.

5.2.2 Objective Evaluation Measures

The objective evaluation measures used in our experiments include Segmental SNR (SEGSNR) and Mel-Frequency Cepstral Coefficient (MFCC) Distance.

Segmental SNR (SEGSNR) is a widely used objective evaluation criterion for speech enhancement or noise reduction algorithms since it is more correlated to subjective results [35]. SEGSNR is defined as the ratio of the power of "ideal" clean speech to that of the noise signal embedded in a noisy signal or in an enhanced speech signal by tested algorithms over all frames, given by:

$$\text{SEGSNR} = \frac{1}{L}\sum_{\lambda=0}^{L-1} 10\log_{10}\left(\frac{\sum_{j=0}^{M-1}[s(\lambda M+j)]^2}{\sum_{j=0}^{M-1}[\hat{s}(\lambda M+j)-s(\lambda M+j)]^2}\right), \tag{29}$$

where (i) $s(.)$ and $\hat{s}(.)$ are the reference speech signal and noisy signal or enhanced signals processed by the tested algorithms; (ii) L and M represent the number of frames in the signal and the number of samples per frame (equal to the length of STFT).

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 can be represented as:

$$d_{\text{mfcc}} = \frac{1}{|\Phi|}\sum_{\lambda\in\Phi}\sum_{\theta=0}^{\Theta}\left(\psi_\theta-\psi_\theta^{'}\right)^2, \tag{30}$$

where Φ represents the set of frames in which speech is present and $|\Phi|$ its cardinality; ψ_θ and $\psi_\theta^{'}$ are the MFCCs of the clean speech signal and noisy signal or enhanced speech signals, respectively; $\Theta = 12$ denotes the order of the MFCCs.

5.2.3 Objective Evaluation Results

Fig. 7 shows the experimental results of the tested noise reduction systems in terms of SEGSNR in the simulated and real-world noise environments at various noise levels. Significant noise suppression performance is achieved consistently by employing the proposed noise reduction system in the tested conditions. Compared to the noisy inputs and the enhanced signals by the Delay-and-Sum beamformer with/without post-filter, the SEGSNE improvements of the proposed system amount to about 20 dB in low SNRs and about 15 dB in high SNRs for all the tested noise conditions. These are larger than the enhanced signals from other traditional systems, especially when the passenger is speaking. The

significant SEGSNR improvement of the proposed noise reduction system is attributed to its sufficient capability in suppressing both the localized and non-localized noises simultaneously.

Fig. 8 shows the experimental results of the tested noise reduction systems in terms of MFCC distance in the simulated and real-world environments at various noise levels. The MFCC distances of the enhanced signals are decreased by all the test systems, especially in low SNRs, compared to those of the noisy input signal. Among the tested noise reduction systems, the lowest MFCC distances are achieved by the proposed system in all noise environments at various noise levels. The improvement in MFCC distance sense of the proposed noise reduction system can be attributed to its spatial filtering capability in discriminating desired speech from interfering noises.

5.2.4 Subjective Evaluations

Subjective evaluations and comparisons of the tested noise reduction systems were performed using speech spectrograms and listening tests. Typical examples of speech spectrograms are illustrated in Fig. 9 for the real-world car noise environment with the speed of 100 km/h and interfering passenger's voice. Figs. 9 (c) and (d) demonstrate that the outputs of the Delay-and-Sum beamformer without/with post-filter are characterized by the high level of noise due to its small physical size (3ch in this work) and incapability in reducing the correlated noise in low frequencies. The single-channel OM-LSA estimator suppresses a large amount of diffuse noise components while there is less suppression for the spatially correlated interfering passenger's voice, as shown in Fig. 9 (e). Fig. 9 (f) illustrates that the hybrid estimation technique based noise suppression system alone succeeds in suppressing the localized passenger's interfering voice and fails in suppressing the non-localized diffuse noise components. The further spectral enhancement, shown in Fig. 9 (g), is achieved by the proposed noise reduction system which is effective in suppressing both the localized and non-localized noise components simultaneously. This improvement is attributed to the fact that not only the statistic characteristic of the signals but also the spatial characteristics of the noise field are taken into account in the proposed noise reduction system, which thus provides more possibilities in distinguishing the desired speech and undesired signals including localized and non-localized signals.

For the listening tests, the sounds data (clean, noisy and enhanced speech sounds), in car environment with speed of 100 km/h and interfering passenger's voice, were first grouped into various pairs and each of which was then randomly presented to 8 subjects

through binaural earphones at a comfortable loudness level. Scheffe's paired comparison method was used to evaluate the preference of enhanced speeches in terms of seven-grade scores [-3, 3]. Then, preference of enhanced speeches at various SNRs was calculated as the score normalized with the number of subjects and sentences in our experiments. The experimental results at various SNRs, illustrated in Fig. 10, shows that our proposed noise reduction system also yields the highest scores compared to other traditional noise reduction systems, which further verifies the superiority of the proposed noise reduction in listening sense as well.

6 Conclusions

In this paper, we presented a novel noise reduction system using a hybrid noise estimation technique and post-filtering based on a generalized signal model. The final estimation accuracy of the hybrid estimation technique for localized noise components was improved by a robust and accurate speech absence probability estimator. The estimated localized noise spectra were then suppressed by spectral subtraction. At spectral subtraction output, a novel estimator for the *a priori* speech absence probability was derived based on the coherence characteristic of the noise field and it further enhanced the noise reduction performance of the post-filter for non-localized noises.

The proposed hybrid noise estimation technique was evaluated and its performance was further compared to that of the conventional noise estimation techniques in various localized noise environments. The experimental results showed the hybrid estimation technique yielded the smallest estimation error among the tested estimation techniques. And, the proposed noise reduction system was evaluated and its performance was further compared to that of the conventional noise reduction systems in various noise conditions in terms of the objective and subjective measures. The experimental results showed that among the tested noise reduction systems the proposed system was able to reduce both localized and non-localized noises and consistently achieved the best noise reduction performance without additional distortion of the desired speech signal in the tested noise environments.

7 Acknowledgement

The authors would like to thank all the subjects who attended the listening tests, especially thank Mr. Takeshi Saitou for his help in the listening tests.

References

[1] Rabiner L. and Juang B.-H., 1993. Speech Recognition System Design and Implementation Issues. Fundamental of Speech Recognition. Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

[2] Gary W. Elko, 1996. Microphone Array systems for hands-free telecommunication. Speech Communication. vol.20, no.3-4, pp. 229-240.

[3] Bitzer J., Simmer K.U., 2001. Superdirective Microphone Arrays. Microphone Arrays Signal Processing Techniques and Applications, Berlin: Springer, pp.19-38.

[4] Boll S.F., 1979. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. on Acoustic, Speech and Signal Processing, vol. ASSP-27, no. 2, pp. 113-120.

[5] Berouti M., Schwartz R.and Makhoul J., 1979. Enhancement of Speech Corrupted by Additive Noise. In Proc. ICASSAP'79, pp. 208-211.

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

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

[8] Bitzer J., Simmer K.U. and Kammeyer K.-D. 1998. Multichannel Noise Reduction - Algorithms and Theoretical Limits. In Proc. European Signal Processing Conference, Rhodes, Greece, pp. 105-108.

[9] Akagi M. and Mizumachi M., 1997. Noise Reduction by Paired Microphones. In Proc. EUROSPEECH97, pp. 335-338.

[10] Mizumachi M. and Akagi M., 1999. Noise reduction method that is equipped for a robust direction finder in adverse environments. In Proc. IEEE Workshop on Robust Method for Speech Recognition in Adverse Conditions, Tampere, Finland, pp. 179-182.

[11] Akagi M. and Kago T., 2002. Noise Reduction Using a Small-Scale Microphone Array in Multi Noise Source Environment. In Proc. IEEE International Conference on Acoustic, Speech Signal Processing, ICSSAP-2002, pp. 909-912.

[12] Li J. and Akagi M., 2004. Noise Reduction Using Hybrid Noise Estimation Technique and Post-Filtering. Accepted by ICSLP2004.

[13] Van veen B.D. and Buckley K.M., 1988. Beamforming: A Versatile Approach to Spatial Filtering. IEEE ASSP Magazin, vol.5, no. 2, pp. 4-24.

[14] Simmer K.U., Bitzer J., Marro C., 2001. Post-Filtering Techniques. Microphone Arrays Signal Processing Techniques and Applications, Berlin: Springer, pp.39-60.

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

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

[17] Bitzer J., Simmer K.U. and Kammeyer K.-D., 2001. Multi-Microphone Noise Reduction Techniques as Front-end Devices for Speech Recognition", Speech Communication, vol. 34, pp. 3-12.

[18] Bitzer J. Simmer K.U. and Kammeyer K.-D., 1999. Multi-microphone noise reduction by post-filter and superdirective beamformer. In International Workshop on Acoustic Echo and Noise Control, Pocono Manor, US, pp. 27-30.

[19] Bitzer J., Simmer K.U. and Kammeyer K.-D., 1999. Theoretical noise reduction limits of the generalized sidelobe canceller (GSC) for speech enhancement. In Proc. 24th IEEE International Conference on Acoustic, Speech Signal Processing, ICASSP-99, Phoenix, Arizona, pp. 2965-2968.

[20] Meyer J. and Simmer K. U., 1997. Multi-channel speech enhancement in a car enviornment using Wiener filtering and spectral subtraction. In Proc. 22th IEEE International Conference on Acoustic, Speech Signal Processsing, ICASSP-97, Munich, Germany, pp. 21-24.

[21] Marro C., Mahieux Y. and Simmer K.U. 1998. Analysis of Noise Reduction and Dereverberation Techniques Based on Microphone Arrays with Postfiltering. IEEE Trans. on Speech and Audio Processing, vol. 6, no. 3, pp. 240-259.

[22] Meyer J., Simmer K.U. and Kammeyer K.D., 1997. Comparison of one- and two-channel noise estimation techniques. In Proc. 5th International Workshop on Acoustic Echo and Noise Control, IWAENC-97, London, UK, pp. 17-20.

[23] Zhang X.X. and Hansen John H.L., 2003. CSA-BF: A Constrained Switched Adaptive Beamformer for Speech Enhancement and Recognition in Real Car Enviornment. IEEE Trans. on Speech and Audio Processing, vol. 11, no. 6, pp. 733-745.

[24] Ephraim Y. and Malah D., 1984. 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.

[25] Ephraim Y. and Malah D., 1985. 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.

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

[27] Bouquin-Jeannes R. L. and Faucon G., 1995. Study of a voice activity detector and its influence on a noise reduction system. Speech Communication, vol. 16, no. 3, pp. 245-254.

[28] Simmer K.U. and Wasiljeff A., 1992. Adaptive Microphone Arrays for Noise Suppression in the Frequency Domain. In Proc. Workshop on Adaptive Algorithms in Communications, Bordeaux, France, pp. 185-194.

[29] Spriet A., Moonen M. and Wouters J., 2003. The impact of speech detection errors on the noise reduction performance of multi-channel Wiener filter. In Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, ICASSP, Hong Kong, China, pp. 501-504.

[30] Omologo M. and Svaizer P., 1997. Use of the Crosspower-Spectrum Phase in Acoustic Event Location. IEEE Trans. on Speech and Audio Processing, vol. 5, no. 3, pp. 288-292.

[31] Bouquin-Jeannes R. L., Azirani A.A. and Faucon G., 1997. Enhancement of Speech Degraded by Coherent and Incoherent Noise Using a Cross-Spectral Estimator. IEEE Trans. on Speech and Audio Prcoessing, vol. 5, no. 5, pp. 484-487.

[32] Malah D., Cox R.V. and Accardi A.J., 1999. Tracking speech-presence uncertainty to improve speech enhancement in nonstationary noise environments. in Proc. 24th IEEE ICASSP'99, Phoenix, pp. 789-792.

[33] Gannot S., Burshtein D. and Weinstein E., 2001. Signal Enhancement Using Beamforming and Nonstationary with Application to Speech. IEEE Trans. on Signal Processing, vol. 49, no. 8, pp. 1614-1626.

[34] Cohen I., 2003. Analysis of Two-Channel Generalized Sidelobe Canceller (GSC) With Post-Filtering. IEEE Trans. on Speech and Audio Processing, vol. 11, no. 6, pp. 684-698.

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

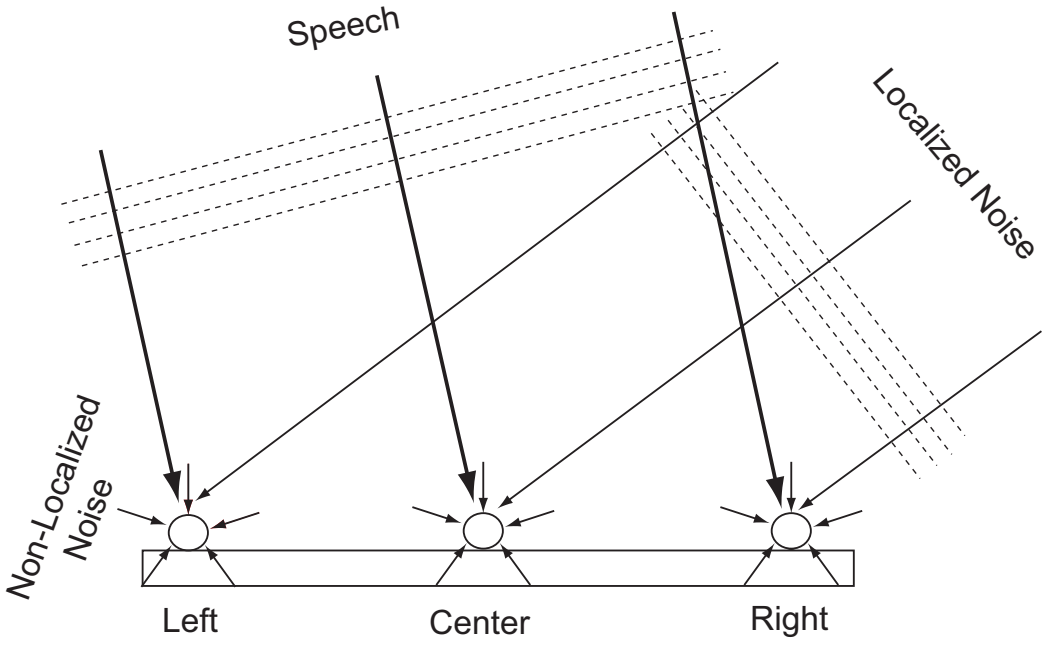

Figure 1: Relationship between microphone array and acoustic signals

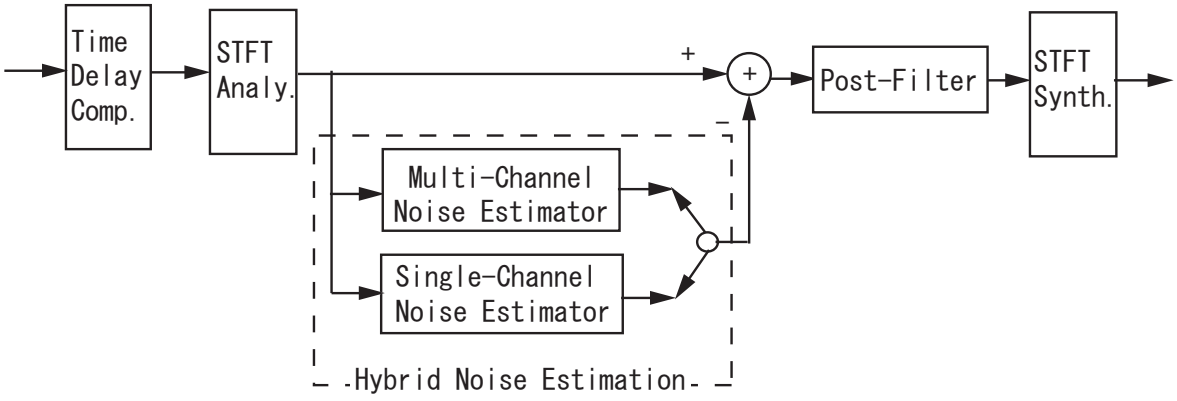

Figure 2: Block diagram of the proposed noise reduction system

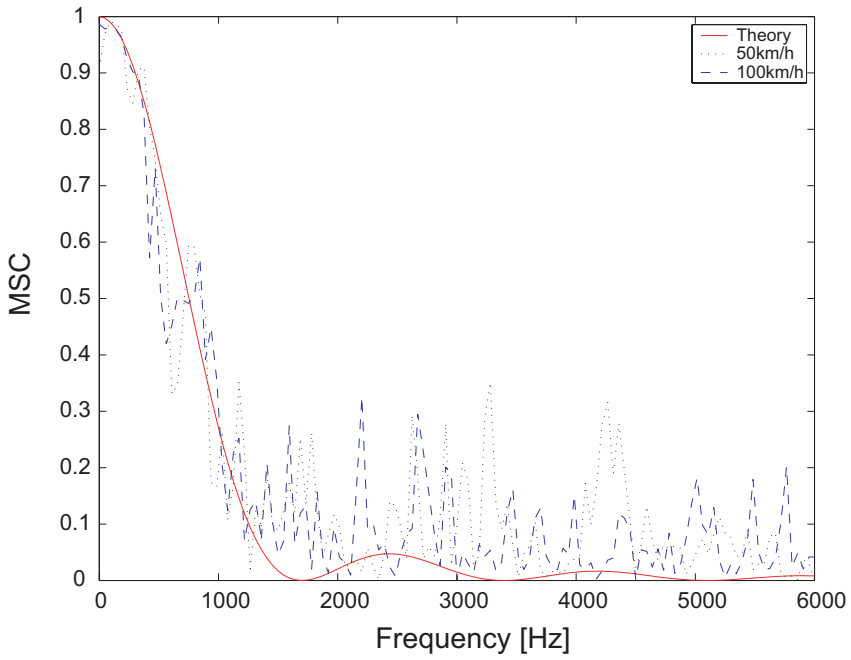

Figure 3: Magnitude-squared coherence in car environments: Theoretical MSC (solid line) and measured MSCs in a car environment with speeds of 50 km/h (dotted line) and 100 km/h (dashed line). The distance between adjacent microphones is 10 cm.

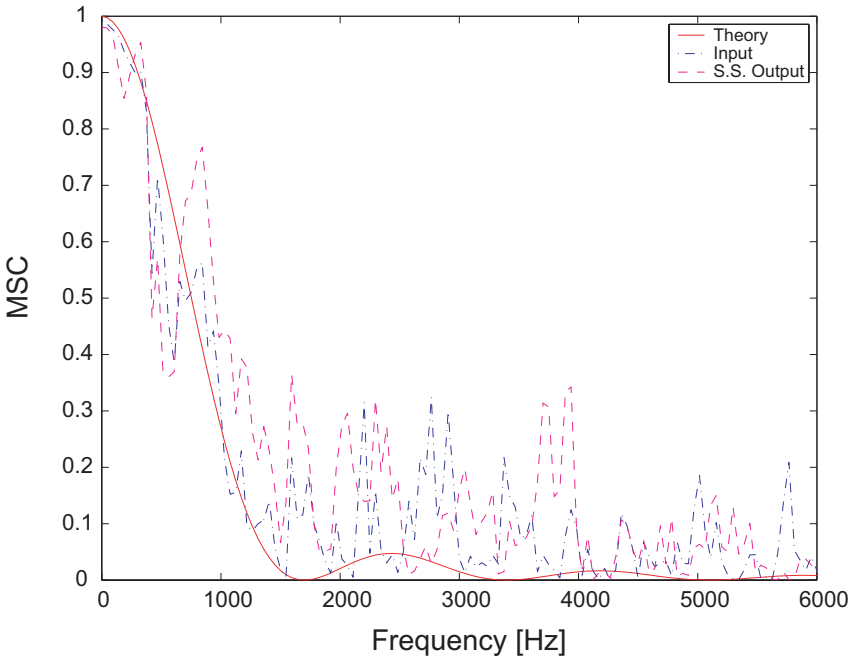

Figure 4: Magnitude-squared coherence in car environments: Theoretical MSC (solid line) and measured MSCs at the input of the system (dashdot line) and at output of the spectral subtraction (dashed line). The distance between adjacent microphones is 10 cm.

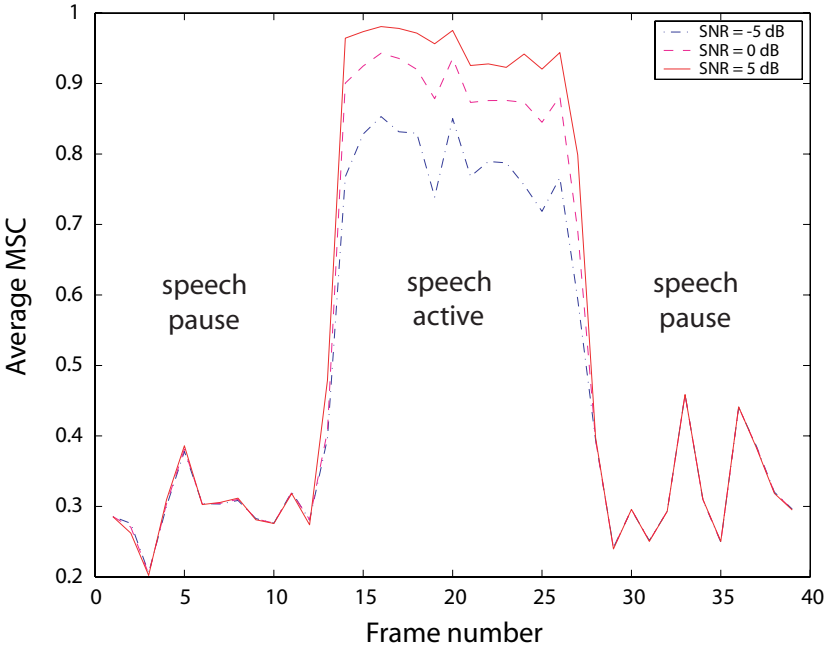

Figure 5: Average MSCs over all frequencies in a car environment at various SNRs: SNR = -5 dB (dashdot line); SNR = 0 dB (dashed line); SNR = 5 dB (solid line). Noise condition: car environment with the speed of 100 km/h.

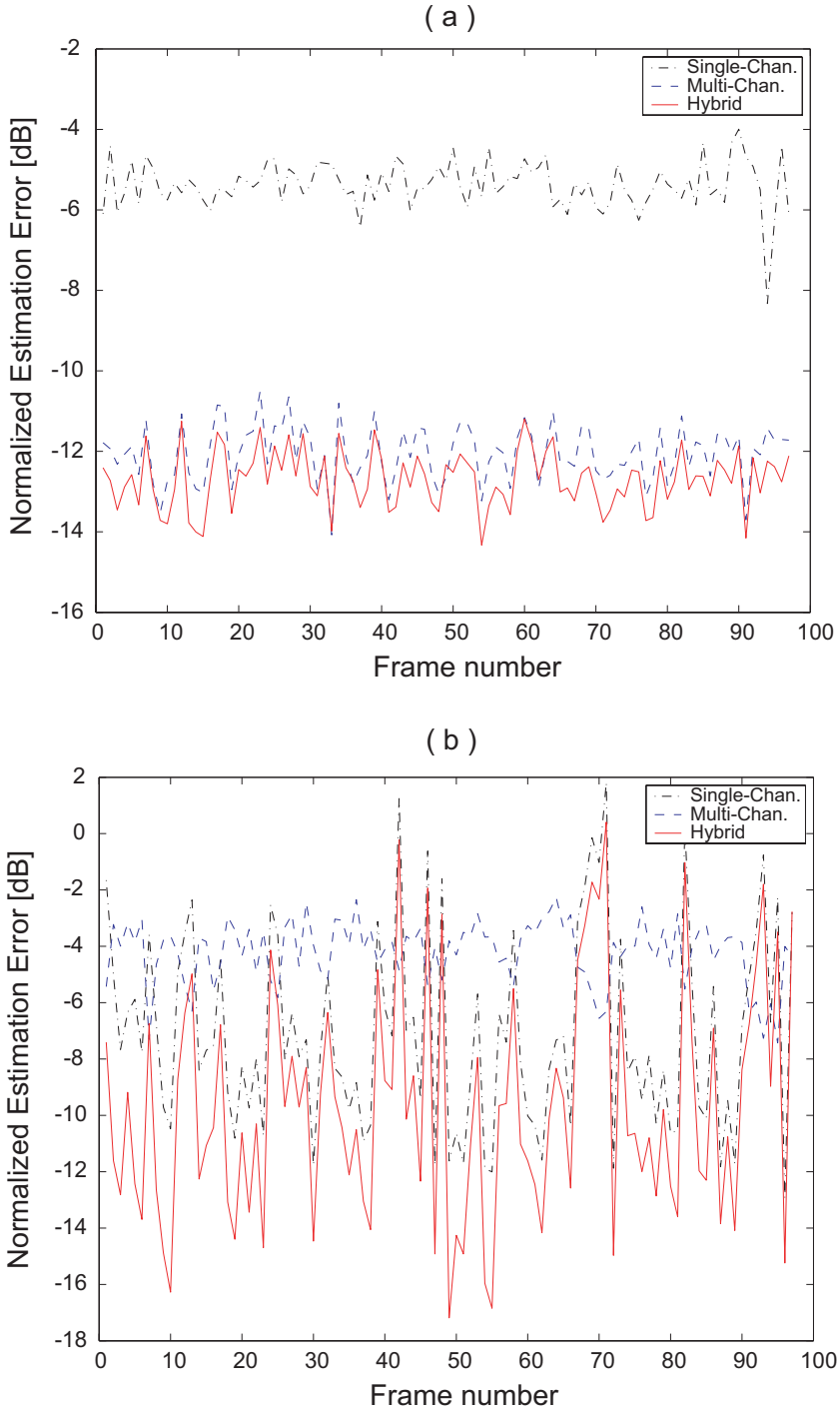

Figure 6: Normalized Noise Estimation Error (dB) for signals processed by single-channel technique (dashdot), multi-channel technique (dashed) and hybrid technique (solid) under white noise conditions (a) and car noise conditions (b).

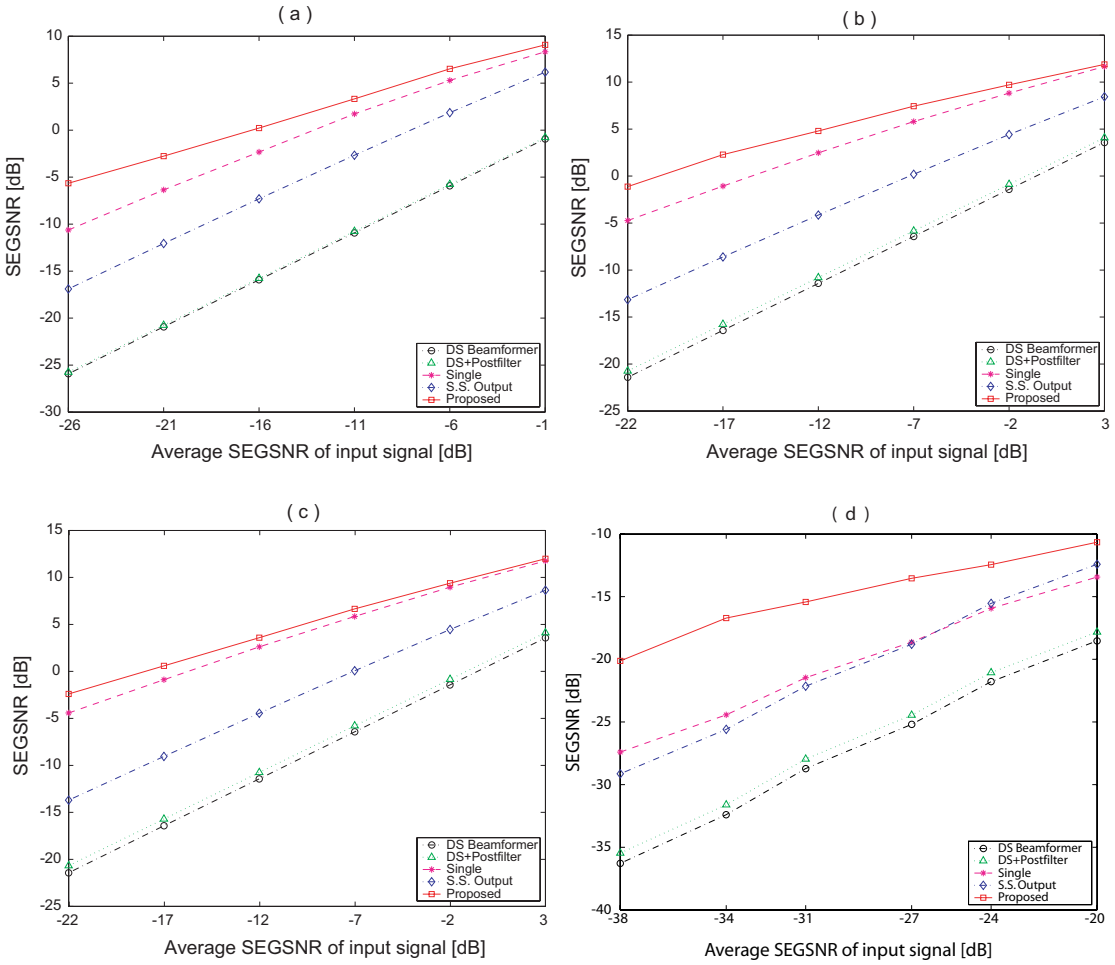

Figure 7: Average segmental SNR (dB) at delay-and-sum beamformer output (○), delay-and-sum beamformer with postfilter output (△), single-channel OM-LSA output (*), spectral subtraction output (◇) and proposed system output (□), in various noise conditions: Simulated condition (a); Car environment with a speed of 50 km/h (b); Car environment with a speed of 100 km/h (c); Car environment with the speech of 100 km/h and passenger's interfering voice (d).

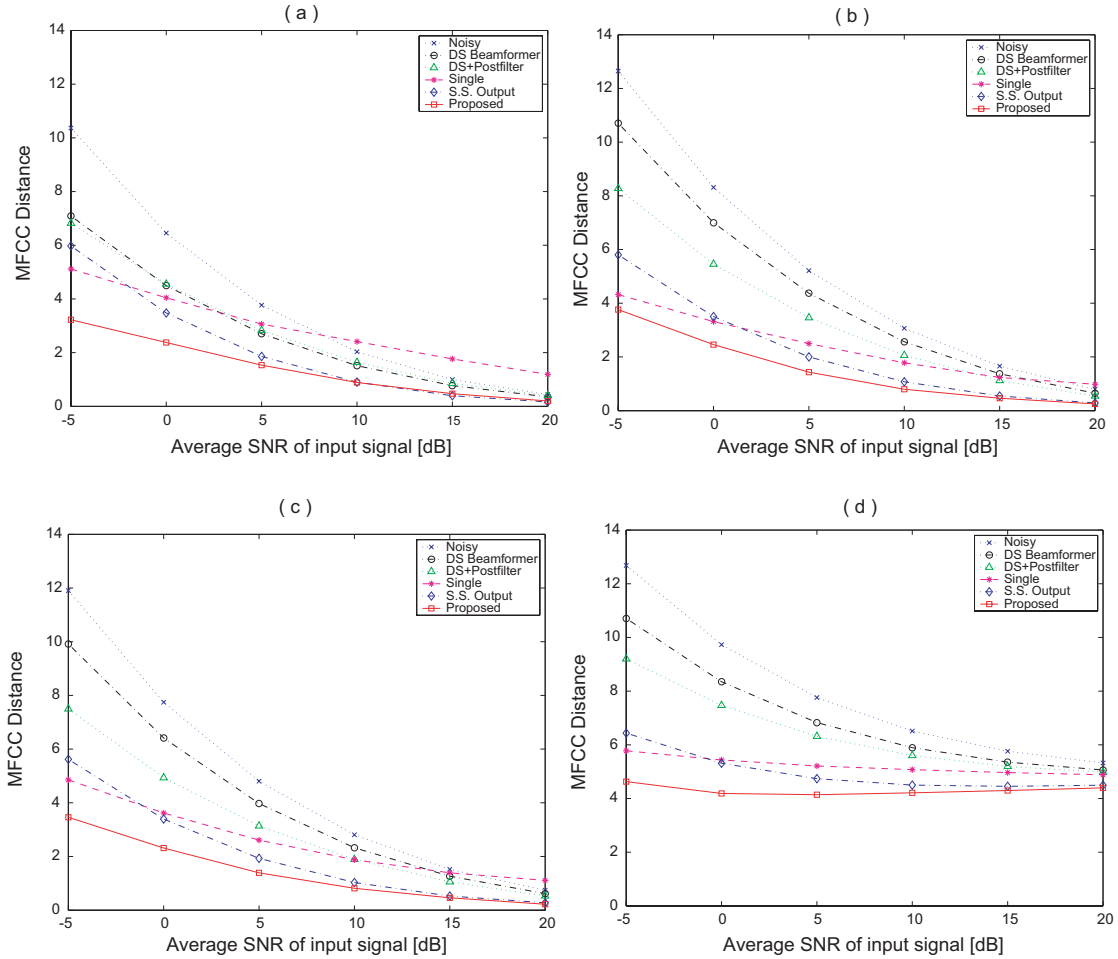

Figure 8: Average MFCC distances at center microphone ($\times$), delay-and-sum beamformer output ($\circ$), delay-and-sum beamformer with postfilter output ($\triangle$), single-channel OM-LSA output (*), spectral subtraction output ($\diamond$) and proposed system output ($\square$), in various noise conditions: Simulated condition (a); Car environment with a speed of 50 km/h (b); Car environment with a speed of 100 km/h (c); Car environment with a speech of 100 km/h and passenger's interfering voice (d).

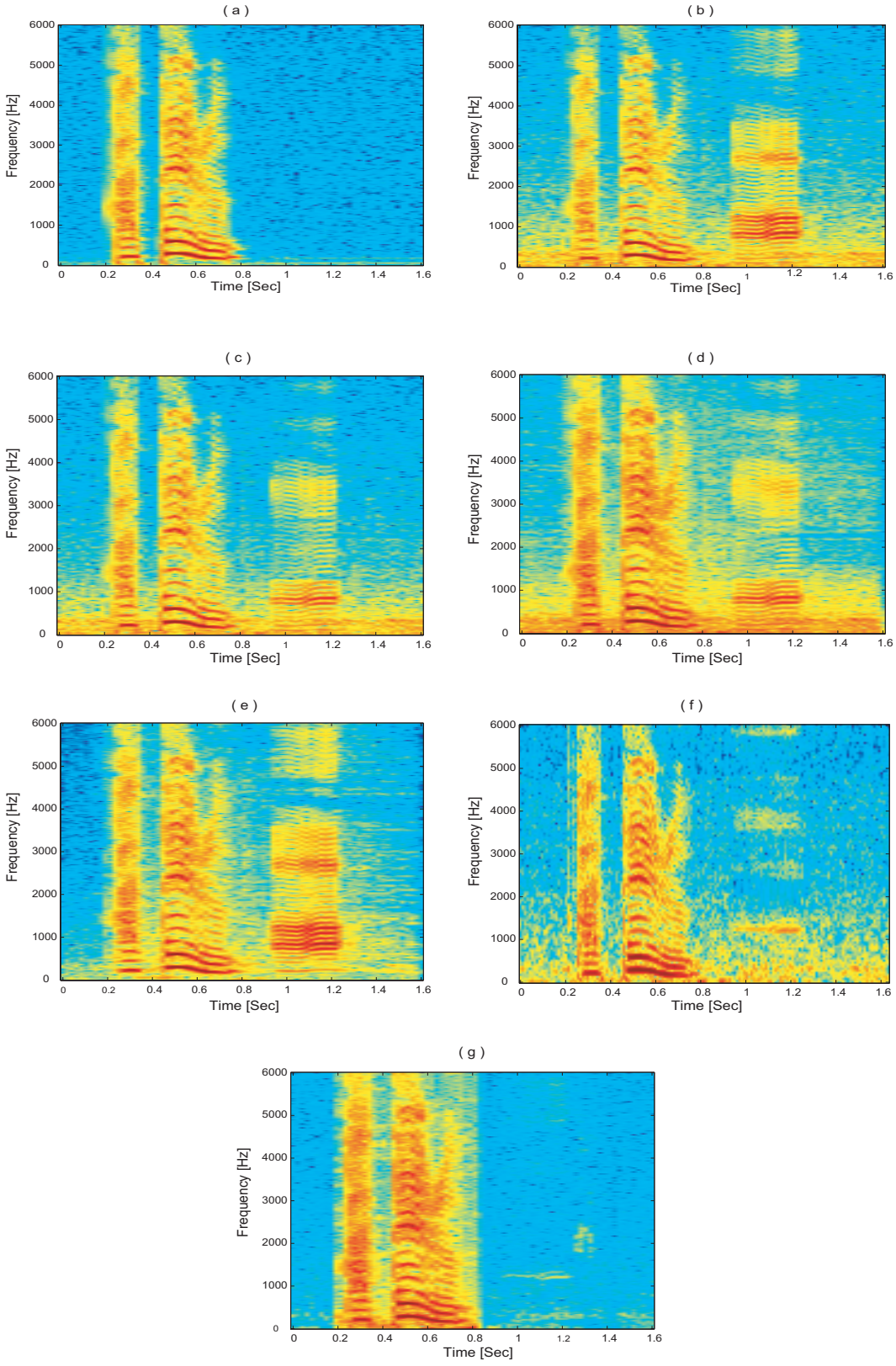

Figure 9: Speech spectrograms. Clean signal at center microphone (a); Noisy signal at center microphone (b); Delay-and-sum beamformer output (c); Delay-and-sum beamformer with postfilter output (d); Single-channel OM-LSA output (e); Spectral subtraction output (f); Proposed system output (g). Noise condition: car environment with a speed of 100 km/h and passenger's interfering voice.

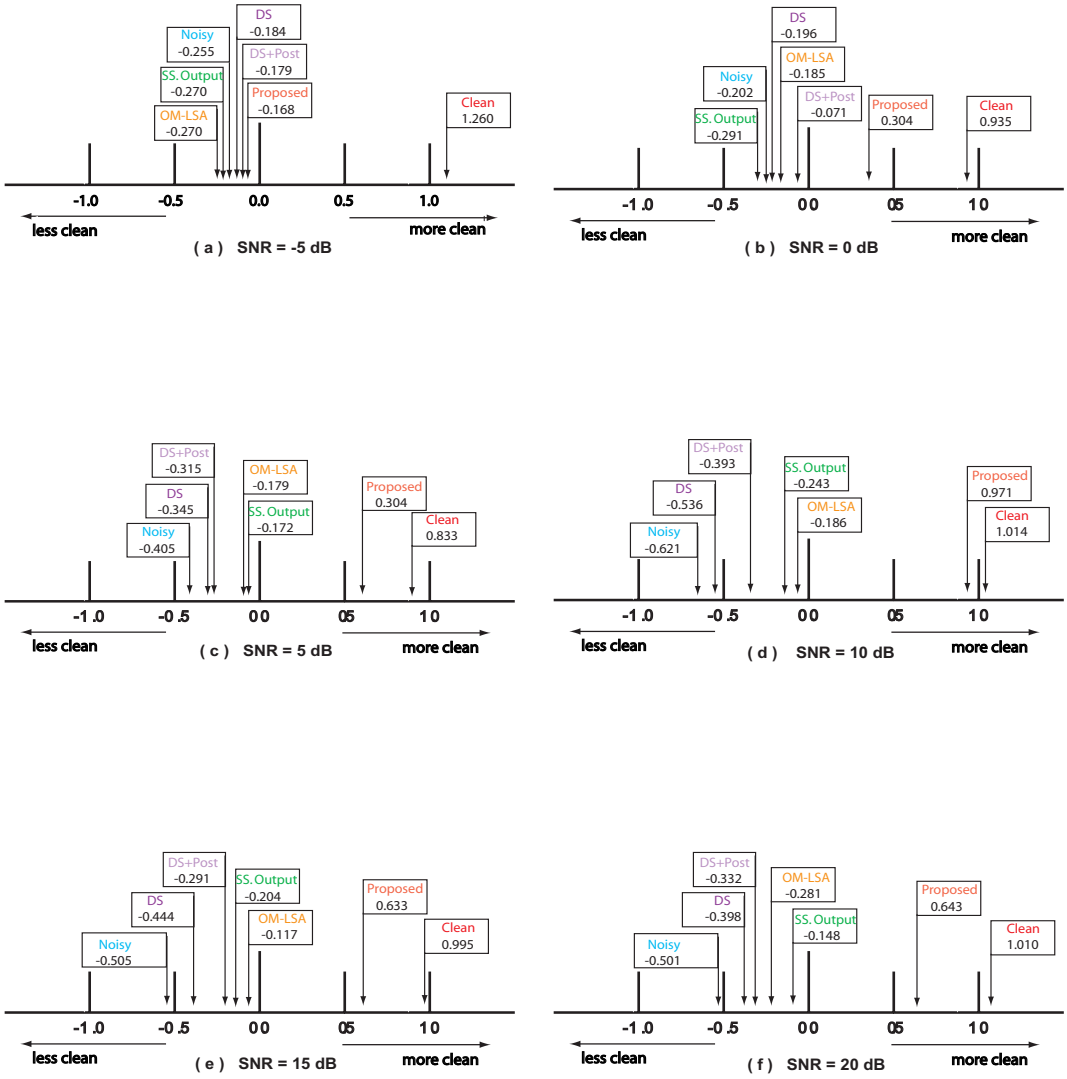

Figure 10: Listening test results. Comparisons of clean speech (Clean), noisy speech (Noisy) and enhanced speeches by delay-and-sum beamformer (DS), delay-and-sum beamformer with postfilter (DS+Post), single-channel OM-LSA (OM-LSA), the subtractive beamformer based system (S.S.) and proposed noise reduction system (Proposed) for listening in various SNRs. SNR = -5 dB (a); SNR = 0 dB (b); SNR = 5 dB (c); SNR = 10 dB (d); SNR = 15 dB (e); SNR = 20 dB (f). Noise condition: car environment with the speed of 100 km/h and passenger's interfering voice.