A NOISE REDUCTION SYSTEM IN ARBITRARY NOISE ENVIRONMENTS AND ITS APPLICATIONS TO SPEECH ENHANCEMENT AND SPEECH RECOGNITION

Junfeng Li, Xugang Lu and Masato Akagi

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

ABSTRACT

This paper proposes a novel noise reduction system in arbitrary noise environments, consisting of localized and non-localized noises, where few existing systems work well. In the proposed system, localized noises are estimated and reduced by the hybrid noise estimation technique we previously proposed and spectral subtraction. And non-localized noises are reduced by a post-filter of which the performance is further improved by a novel estimator for the *a priori* speech absence probability calculated under the assumption of diffuse noise field. Experimental results show that the proposed system results in significant improvements in terms of speech quality measures and speech recognition performance in various noise conditions.

1. INTRODUCTION

In recent years, much research has been undertaken into noise reduction to improve the performance of speech communication and recognition systems. Noise reduction suppress environmental noises, improving the speech quality and increasing the recognition accuracy. Thus, noise reduction has been of increased interests for many researchers.

Although a variety of noise reduction systems have been proposed, few of them can reduce both localized and non-localized noises simultaneously in arbitrary noise environments [1]-[6]. To suppress localized noises, a lot of algorithms based on beamforming techniques have been presented with the drawback of large physical size (delay-and-sum beamformer) or adaptive signal processing (GSC beamformer) [1]. To suppress non-localized noises, post-filtering is normally needed. A commonly used post-filter is the one, first proposed by Zelinski, based on the assumption of incoherent noise field [4]. This assumption is, however, seldom satisfied in practical environments, especially for closely-spaced microphones and low frequencies. Recently, a generalized expression for Zelinski post-filter has been derived based on the *a priori* knowledge of noise field [5].

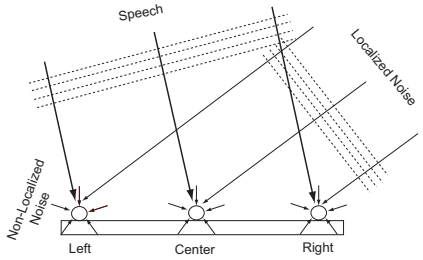


Fig. 1. Microphone array and signal model.

In this paper, we propose a novel noise reduction system which consists of localized noises suppression previously presented, and non-localized noises suppression based on the *optimally-modified log-spectral amplitude* (OM-LSA) estimator [6]. Under the assumption of diffuse noise field which was confirmed to be more accurate in a number of realistic noise environments [5], we propose a novel estimator for the *a priori* speech absence probability (SAP), further improving the noise reduction ability of the post-filter. The performance of proposed system is investigated and is shown to result in significant improvement over the comparative systems in various noise environments.

2. PROPOSED NOISE REDUCTION SYSTEM

Considering a three-sensor microphone array in a noisy environment, shown in Fig. 1, the observed signal on each microphone is composed of desired speech signal, localized noises arriving from determinable directions and non-localized noises propagating in all directions. The aim of our task is to reduce both localized and non-localized noises simultaneously while keeping the desired speech distortionless. To implement this idea, we construct a noise reduction system, shown in Fig. 2, which consists of two main parts: localized noise suppression and non-localized noise suppression, detailed in the following.

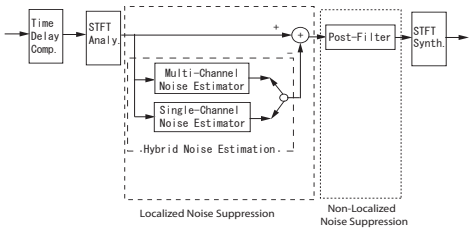


Fig. 2. Block diagram of the proposed system

2.1. Localized Noises Suppression

To suppress localized noises, authors have proposed a hybrid noise estimation technique, which combines a subtractive beamformer based multi-channel estimation technique and a soft-decision based single-channel estimation technique, yielding more accurate spectral estimates for localized noises [3]. The spectrum of localized noise, $\hat{N}^c(\lambda,\omega)$, calculated by the hybrid technique, can be given by [3]:

$$\hat{N}^c(\lambda,\omega) = \begin{cases} \hat{N}^c_m(\lambda,\omega), & \text{not array sidelobes,} \\ \hat{N}^c_s(\lambda,\omega), & \text{array sidelobes,} \end{cases} \quad (1)$$

where λ and ω are the frame index and the frequency index; $\hat{N}^c_m(\lambda,\omega)$ and $\hat{N}^c_s(\lambda,\omega)$ are the estimated spectrum for localized noise by the multi-channel technique [2] and the single-channel technique [3], respectively. Estimation accuracy of the hybrid technique is further improved by considering the strong correlation of speech presence uncertainty between adjacent frequencies and consecutive frames [3]. The estimated spectra of localized noises are then subtracted from those of three noisy signals.

2.2. Non-Localized Noises Suppression

The residual non-localized noises are further suppressed by a post-filter which is based on the OM-LSA estimator characterized by the following gain function [6]:

$$G(\lambda,\omega) = G_{H_1}(\lambda,\omega)^{1-q(\lambda,\omega)} G_{\min}^{q(\lambda,\omega)}, \quad (2)$$

where $G_{\min}$, $q(\lambda,\omega)$ and $G_{H_1}(\lambda,\omega)$ are a constraint constant, the SAP at spectral subtraction output and the gain function of the traditional MMSE-LSA estimator when speech is surely present given in [7], respectively.

As Eq. 2 shows, the performance of this post-filter is greatly dependent on the SAP $q(\lambda,\omega)$ which is further closely related to the *a priori* SAP $q^{'}(\lambda,\omega)$ according to Bayes' rule [6]. Therefore, the performance of this post-filter is believed to be significantly dependent on the *a priori* SAP. However, the problem of how to accurately calculate the *a priori* SAP at spectral subtraction output has not been solved so far. In the following, we propose a new estimator for the *a priori* SAP based on the coherence characteristic of the noise field at spectral subtraction output.

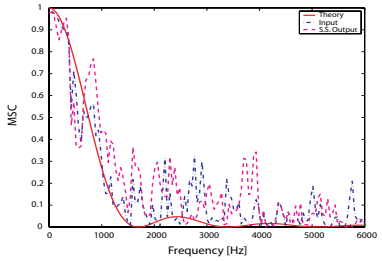


Fig. 3. Magnitude-squared coherence in car environment: Theoretical MSC (solid line) and measured MSCs at the input of the system (dashdot line) and at the output of spectral subtraction (dashed line). The distance between microphones is 10 cm.

2.2.1. *Noise field analysis at spectral subtraction output*

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

$$\Gamma(\lambda,\omega) = \frac{|\phi_{ij}(\lambda,\omega)|^2}{\phi_{ii}(\lambda,\omega)\phi_{jj}(\lambda,\omega)}, \quad (i,j=1,2,3), \quad (3)$$

where $\phi_{ij}(\lambda,\omega)$ is the cross-spectral density between two signals with auto-spectral densities of $\phi_{ii}(\lambda,\omega)$ and $\phi_{jj}(\lambda,\omega)$.

The MSC of theoretical diffuse noise field is shown in Fig. 3 along with the measured MSCs using the input car noises and the outputs of spectral subtraction. Fig. 3 suggests that: (i) car noise environment is characterized by diffuse noise; (ii) diffuse characteristic of non-localized noises does not change at spectral subtraction output; (iii) noises at spectral subtraction output are weakly correlated in high frequencies and strongly correlated in low frequencies.

2.2.2. *An estimator for the a priori SAP*

Based on the observations mentioned above, the MSC spectra are divided into two parts: high frequency region with low MSCs and low frequency region with high MSCs. And the transient frequency between two regions is the first minimum frequency of the MSC of theoretical diffuse noise field, given by $f = c/(2d)$, where d and c are the distance between neighboring microphones and the velocity of sound propagation, respectively. Furthermore, two different schemes are proposed to determine the *a priori* SAP in the high frequency region and the low frequency region, as follows:

- In the high frequency region: The MSC spectra are further divided into E sub-bands and averaged across the frequencies in each sub-band, obtaining the average MSC $\bar{\Gamma}_e(\lambda,\omega)$ $(e = 1,2,...,E)$ in e-th sub-band. If a high coherence $(\bar{\Gamma}_e(\lambda,\omega) > T\max_e)$ is detected, a speech present state is detected presumably. If a low coherence $(\bar{\Gamma}_e(\lambda,\omega) < T\min_e)$ is detected, a speech absent state is detected presumably. For

$\bar{\Gamma}_e(\lambda,\omega) \in [T\min_e, T\max_e]$, the *a priori* SAP is determined by the linear interpolation. Thus, the *a priori* SAP in the high frequency region, $q_h^{'}(\lambda,\omega)$, can be given by:

$$q_h^{'}(\lambda,\omega) = \begin{cases} 0, & \bar{\Gamma}_e(\lambda,\omega) > T\max_e, \\ 1, & \bar{\Gamma}_e(\lambda,\omega) < T\min_e, \\ \frac{T\max_e - \bar{\Gamma}_e(\lambda,\omega)}{T\max_e - T\min_e}, & \text{otherwise}, \end{cases}$$
$$\omega \in [\omega_e^{low}, \omega_e^{high}], \quad (4)$$

where ω_e^{low} and ω_e^{high} are the low and high boundaries of e-th sub-band, and $T\min_e$ and $T\max_e$ are two empirical constants.

- In the low frequency region: The speech is assumed to be present when it is present in the high frequencies. Thus, an average MSC $\bar{\Gamma}(\lambda,\omega)$, obtained by averaging the MSCs across the frequencies over the transient frequency, provides a useful measure to detect speech. Based on the MSC $\bar{\Gamma}(\lambda,\omega)$ and following the same ideas in the high frequency region, the *a priori* SAP in the low frequency region, $q_l^{'}(\lambda,\omega)$, is estimated as:

$$q_l^{'}(\lambda,\omega) = \begin{cases} 0, & \bar{\Gamma}(\lambda,\omega) > T\max, \\ 1, & \bar{\Gamma}(\lambda,\omega) < T\min, \\ \frac{T\max - \bar{\Gamma}(\lambda,\omega)}{T\max - T\min}, & \text{otherwise}, \end{cases} \quad (5)$$

where $T\max$ and $T\min$ are another two empirical constants, and

$$\bar{\Gamma}(\lambda,\omega) = \frac{1}{E}\sum_{e=1}^{E} \bar{\Gamma}_e(\lambda,\omega). \quad (6)$$

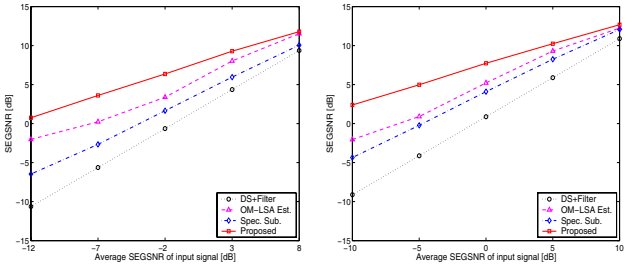
The estimated *a priori* SAP is then incorporated into the post-filter with the purpose of improving the noise reduction performance of this post-filter.

3. EXPERIMENTS AND RESULTS

3.1. Experimental Configurations

An equally-spaced linear microphone array, consisting of three sensors with a inter-element spacing of 10 cm, shown in Fig. 1, was mounted above the windshield in a car. The array was about 50 cm apart from and directly in front of the driver. The recording was performed across all channels simultaneously, which were mainly composed of engine noise, high air-condition noise and the noise coming from frication between tyres and road. Clean speech data, taken from NTT database, consists of 350 phoneme-balanced sentences. Both speech and noise data were first re-sampled to 12 kHz at 16 bit accuracy.

To examine the performance of proposed system in various noise conditions, two sets of noise-corrupted data were generated. The first data set (set A) involved the addition of a randomly selected segment of the multi-channel car noise at different global SNR levels 0-20 dB across 50 speech sentences. The second data set (set B) involved the addition of the multi-channel car noise and a secondary speaker's speech (passenger's interference), which was Japanese vowel /a/, with DOA of 60 degree to the right. Data set B corresponds to a realistic context for a typical car environment. To compare the performance of proposed system, other three systems were chosen, that is, delay-and-sum beamformer with Wiener post-filter, the OM-LSA single-channel algorithm and the localized noise suppression algorithm alone (the output of spectral subtraction).



(a) segmental SNR for data set A (b) segmental SNR for data set B

Fig. 4. Segmental SNR results for data sets A (a) and B (b). Delay-And-Sum beamformer output (dotted); Single-channel OM-LSA estimator output (dashed); Spectral subtraction output (dashdot); Proposed system output (solid).

3.2. Speech Enhancement Experiments

A first set of experiments was conducted using data sets A and B. To assess the performance of speech enhancement systems, a widely used objective speech quality measure, segmental SNR (SEGSNR), was used since it was shown to be more correlated to subjective evaluation results [8]. SEGSNR is defined as the ratio of the power of clean speech to that of noise signal embedded in a noisy or an enhanced speech signal by tested algorithms. The SEGSNR improvements for all tested noise reduction algorithms in two noise conditions are shown in Fig. 4. Moreover, Fig. 5 plots the enhanced signals as well as clean and noise-corrupted signals for an utterance corresponding to "ka no jyo wa te no kon da go chi so o tu ku ri ma shi ta" when both car noise and passenger's interference are present.

Fig. 4 shows the proposed algorithm results in better speech enhancement performance than the comparative algorithms in all tested noise conditions at various SNR levels consistently, especially in low SNRs. These improvements are attributed to its success in reducing both localized and non-localized noises simultaneously compared to other algorithms. This fact can also be clearly observed from the signals shown in Fig. 5.

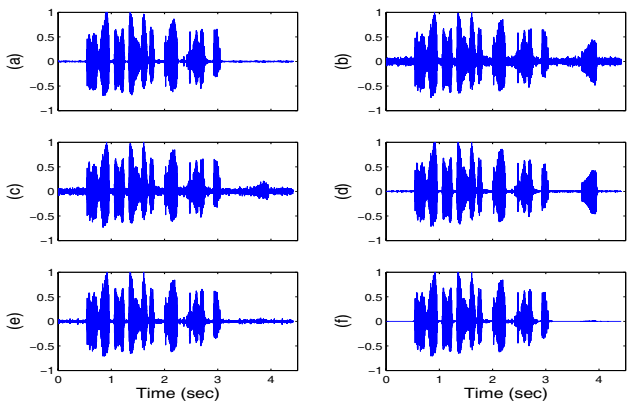


Fig. 5. Waveforms of clean signal (a); noise-corrupted signal (SNR = 10 dB) (b); Delay-And-Sum beamformer with Wiener post-filter output (c); Single-channel OM-LSA estimator output (d); Spectral subtraction output (e); Proposed algorithm output (f).

Technique	Clean	20 dB	15 dB	10 dB	5 dB	0 dB
Noisy	86.33	60.72	48.14	41.17	33.71	29.28
DS+Filter	86.17	58.13	52.94	47.70	39.17	31.87
OM-LSA Est.	85.63	71.37	63.53	51.81	33.39	21.88
Spec. Sub.	86.06	58.18	51.92	43.33	32.52	21.83
Proposed	85.31	74.99	68.29	57.48	47.38	38.12

Table 1. Speech recognition accuracy (%) for data set A.

3.3. Speech Recognition Experiments

To do recognition experiments, a phoneme-based recognition system was constructed. In the experiments, 300 clean sentences were used for training. And the residual 50 sentences, just same as used in the speech enhancement experiments, were used for testing. Standard 16 MFCCs, together with the delta and acceleration features, were used as feature vector of 48 dimensions. In this system, 29 HMM models were generated for 28 mono-phones and a "silence" which was used to be inserted at the start and the end of each sentence. 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. Output distribution probabilities were modelled by means of mixtures of 12 Gaussian components. Standard Viterbi decoding technique was used for recognition.

The speech recognition results for data sets A and B are given in Tables 1 and 2 in terms of the recognition accuracy, defined as: $accuracy = \frac{H-I}{K} \times 100\%$, where H, I and K are the number of correct phonemes, the number of insertions and the total number of phonemes. The recognition results illustrate that the proposed algorithm produces higher speech recognition accuracy compared to other algorithms. And the performance degrades as noises increase. These improvements are attributed to the fact that both statistic characteristic of signals and spatial characteristic of noise field are taken into account in the proposed system.

Technique	Clean	20 dB	15 dB	10 dB	5 dB	0 dB
Noisy	86.33	51.75	43.30	35.71	29.61	26.74
DS+Filter	86.17	50.46	44.79	38.79	35.76	31.39
OM-LSA Est.	85.63	68.67	63.21	45.00	23.55	21.12
Spec. Sub.	86.06	57.21	51.49	39.71	27.88	12.91
Proposed	85.31	73.69	67.21	53.70	38.14	35.88

Table 2. Speech recognition accuracy (%) for data set B.

4. CONCLUSIONS

A novel noise reduction system in arbitrary noise environments has been proposed. The system consists of localized noise suppression, which is based on a hybrid noise estimation technique and spectral subtraction, and non-localized noise suppression which is implemented by a post-filter based on the OM-LSA estimator. The performance of this post-filter is further improved by incorporating an estimator for the *a priori* SAP under the assumption of diffuse noise field. Experimental results indicate that the proposed system results in significant improvement over the comparative algorithms in terms of SEGSNR and speech recognition accuracy.

5. REFERENCES

[1] 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.

[2] Akagi, M. and Kago, T., "Noise reduction using a small-scale microphone array in multi noise source environment", In Proc. ICASSP'02, Orlando, pp. 909-912, 2002.

[3] J. Li and M. Akagi, "Noise Reduction Using Hybrid Noise Estimation Technique and Post-Filtering", To appear in ICSLP2004.

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

[5] 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.

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

[7] 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.

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