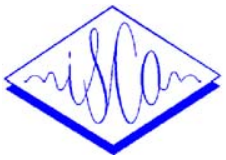
ISCA Archive
http://www.isca-speech.org/archive



INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

NOISE REDUCTION USING HYBRID NOISE ESTIMATION TECHNIQUE AND POST-FILTERING

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

10.21437/Interspeech.2004-718

Abstract

In this paper, a novel noise reduction method using hybrid noise estimation technique and post-filtering is proposed to suppress both localized and non-localized noise components which can not be dealt with by the traditional methods [2][3][4]. To do this, a hybrid noise estimation approach is proposed by combining our previously constructed multi-channel noise estimation approach and a single-channel estimation approach to improve estimation accuracy for localized noise components. The non-localized noise components are suppressed by a single-channel post-filter based on an optimally modified log spectral amplitude (OM-LSA) estimator. To verify the superiorities of the proposed hybrid noise estimation approach and noise reduction system, they are compared to the multi-channel and single-channel scheme based systems under various noise conditions.

1. Introduction

Nowadays, noise reduction systems are in great demand for the increasing number of speech applications, such as automatic speech recognition (ASR) systems and cellular telephony. However, in the recognition or transmission process, speech signals are corrupted by various noises in the enclosures in which they operate, resulting in lower recognition accuracy and decreased intelligibility. A solution to this problem 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 reported in the literature [1]-[5]. A generalized sidelobe canceller (GSC) beamformer, first proposed by Griffiths and Jim, has been widely researched. In the GSC beamformer, adaptive signal processing is normally used to avoid cancellation of the desired speech signal [1]. However, the adaptive signal processing technique decreases the stability of the noise reduction system in real-world environments. A small-scale subtractive beamformer based noise reduction system has recently been proposed by Akagi et.al. [2][3]. Its superiorities lie in its high noise suppression capability, especially for sudden noise, and an analytical noise estimation scheme without adaptive signal processing. However, the basic concept of this system is that undesired noises consist only of localized noise components. Moreover, McCowan has developed a general expression of the post-filter based on the assumption of a diffuse noise field, which is more accurate in practical environments [4].

In this paper, we first introduce a more generalized signal model, containing both localized and non-localized (e.g., diffuse) noise components, which is much more accurate than previous signal models in real-world environments. Then, we present a hybrid noise estimation technique which combines the multi-channel and single-channel techniques in a parallel structure to improve estimation accuracy for the localized noise components. The estimated localized noise components are suppressed by non-linear spectral subtraction. The non-localized noise components are further suppressed by a single-channel post-filter based on the OM-LSA estimator.

2. Overview of the Proposed Noise Reduction System

Assuming a microphone array with three linearly and equidistantly distributed (inter-element spacing is 10cm) omni-directional microphones in a noisy environment, the observed signals consist of: the desired speech signal, arriving from a direction such that the difference in arrival time between the two main microphones is 2ξ; localized noises from directions such that time differences are $2\delta_i$ $(i = 1, 2, \ldots, I)$; and non-localized noises from all directions. Thus, the observed noisy signals imposing on three microphones (left, center, right) can be given by;

$$l(t) = s(t+\xi) + \sum_{i=1}^{I} n_i^c(t+\delta_i) + n_l^{uc}(t) \quad (1)$$

$$c(t) = s(t) + \sum_{i=1}^{I} n_i^c(t) + n_c^{uc}(t) \quad (2)$$

$$r(t) = s(t-\xi) + \sum_{i=1}^{I} n_i^c(t-\delta_i) + n_r^{uc}(t) \quad (3)$$

where n_i^c and $n_\cdot^{uc}$ denote the i-th localized noise signal and non-localized noise signal (e.g. diffuse noise), respectively.

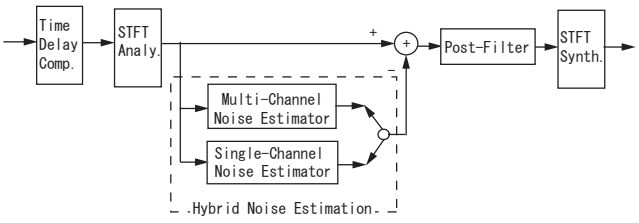


Figure 1: Block diagram of the proposed noise reduction system

Based on this generalized signal model, our purpose is to suppress both the localized noise components and the non-localized noise components while keeping the desired signal intact. To do this, we construct a noise reduction system, shown in Fig. 1, consisting of the following parts: (1) Time delay compensation: it focuses on compensating for the propagation between speech source and microphones on the desired speech signal [2][3]. (2) Localized noise suppression: the localized noise components are first estimated by a hybrid noise estimation approach and then subtracted from the observed signals by spectral subtraction. The hybrid noise estimation approach combines multi-channel and single-channel estimation techniques into a parallel structure to improve estimation accuracy for the localized noise components. (3) Non-localized noise suppression: the non-localized noise components are suppressed by a single-channel post-filter based on an OM-LSA estimator. (4) Spectral analysis and spectral synthesis: they transform the noisy speech signal from the time domain to the frequency domain and inversely transform the enhanced signal from the frequency domain to the time domain, respectively.

Each part shown in Fig. 1 will be explained in the following sections in detail.

3. Suppress Localized Noises

In this section, we will focus on suppressing the localized noise components. To do this, we first propose a hybrid noise estimation method to calculate the localized noise components, which will then be suppressed by spectral subtraction.

3.1. A Hybrid Noise Estimation Method

The proposed hybrid noise estimation method combines the multi-channel estimation approach we proposed previously [2][3] and a single-channel estimation approach into a parallel structure , as shown in Fig. 1.

3.1.1. Multi-Channel Noise Estimation Approach

Based on the generalized signal model, shown in Eqs. (1)-(3), our previous multi-channel estimation approach is reformulated. Two subtractive beamformers $g_{lr}(t)$ and $g_{cr}(t)$ are constructed in the time domain, given by:

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

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

In order to simplify implementation, we make an assumption that non-localized (diffuse) noise components are of the same values at different microphones in the diffuse noise field. Thus, the localized noise spectrum can be estimated from the outputs of the beamformers which blocked the desired speech signal successfully, given by ($\tau = \delta$) [2]:

$$\hat{N}_m^c(\lambda,\omega) = \begin{cases} G_{lr}(\lambda,\omega)/\sin^2\omega\delta, & \text{if} \quad \sin^2\omega\delta > \varepsilon_1 \\ G_{cr}(\lambda,\omega)/\sin^2\omega\frac{\delta}{2}, & \\ \qquad \text{if} \quad \sin^2\omega\delta \le \varepsilon_1 \text{ and } \sin^2\omega\frac{\delta}{2} > \varepsilon_2 & \\ G_{lr}(\lambda,\omega)/\varepsilon_1, & otherwise \end{cases} \quad (6)$$

where (i) $G_{lr}(\lambda,\omega)$ and $G_{cr}(\lambda,\omega)$ are the STFTs of the two subtractive beamformers $g_{lr}(t)$ and $g_{cr}(t)$, respectively; (ii) λ and ω represent frame index and frequency bin index, respectively; (iii) the subscript m, in $\hat{N}_m^c$, indicating the multi-channel technique is used when estimating noise spectrum.

Obviously, this subtractive beamformer-based noise estimation approach has a great capability for estimating localized noise components. However, its estimation accuracy will degrade when the condition $\omega\delta = 2k\pi$ holds since the beamformer does not output any signals. Under this condition, noise spectrum can not be estimated accurately by the multi-channel technique. For this case, a single-channel estimation approach is employed, constructing a hybrid noise estimation approach. In other words, the values of $\sin^2\omega\delta$ and $\sin^2\omega\frac{\delta}{2}$, shown in Eq. (6), control either the output of the single-channel approach or that of multi-channel estimation approach should be selected as the output of this hybrid estimation approach at any time. Thus, the estimated spectra of the localized noise components can be given by:

$$\hat{N}^c(\lambda,\omega) = \begin{cases} \hat{N}_m^c(\lambda,\omega), & \text{if} \quad \max(\sin^2\omega\delta, \sin^2\omega\frac{\delta}{2}) > \varepsilon \\ \hat{N}_s^c(\lambda,\omega), & otherwise \end{cases} \quad (7)$$

where $\hat{N}_m^c(\omega)$ and $\hat{N}_s^c(\omega)$ represent the estimated localized noise spectrum by the multi-channel technique, shown in Eq. (6), and by the single-channel technique which will be given in the following subsection.

3.1.2. Single-Channel Noise Estimation Approach

In the single-channel noise estimation approach, the noise spectrum estimate is updated adaptively in a recursive way as follows:

$$\begin{aligned} \eta_n(\lambda,\omega) &= \alpha_n\eta_n(\lambda-1,\omega) \\ &\quad +(1-\alpha_n)E\left[|N(\lambda,\omega)|^2|C(\lambda,\omega)\right]. \end{aligned} \quad (8)$$

where (i) $\eta_n(\lambda,\omega) = E\left[|N(\lambda,\omega)|^2\right]$ is the variance of noise signal in ω-th frequency bin of λ-th frame; (ii) α_n ($0 < \alpha_n < 1$) is a forgetting factor controlling the update capability in noise estimation. Under speech presence uncertainty, the second term in the right side of Eq. (8) can be represented as:

$$E\left[|N(\lambda,\omega)|^2|C(\lambda,\omega)\right] = q(\lambda,\omega)|C(\lambda,\omega)|^2 + (1-q(\lambda,\omega))\eta_n(\lambda-1,\omega). \quad (9)$$

where $q(\lambda,\omega)$ denotes the speech absence probability in ω-th frequency bin of λ-th frame.

It is of interest to note that the success or failure of the single-channel noise estimation approach is significantly dependent on the conditional speech absence probability $q(\lambda,\omega)$. To combine this single-channel estimation approach with the multi-channel estimation approach mentioned above, the accuracy and robustness of the estimate of speech absence probability must be improved further. For accuracy, based on the estimated noise by the multi-channel technique in most cases and by the single-channel technique only when the multi-channel technique fails, the accuracy of the speech absence probability estimates can be improved significantly. For robustness, the strong correlations of speech presence uncertainty between adjacent frequency components and consecutive frames are taken into account and the *a priori* SNR $\xi(\lambda,\omega)$ and *a posteriori* SNR $\gamma(\lambda,\omega)$ are first smoothed in the time-frequency domain. Based on the complex Gaussian assumption and the smoothed *a priori* SNR $\bar{\xi}(\lambda,\omega)$ and *a posteriori* SNR $\bar{\gamma}(\lambda,\omega)$, the robust speech absence probability $q(\lambda,\omega)$ can be obtained as:

$$q(\lambda,\omega)=\left(1+\frac{1-q^p}{q^p}\frac{1}{1+\bar{\xi}(\lambda,\omega)}\exp\left(\frac{\bar{\xi}(\lambda,\omega)\bar{\gamma}(\lambda,\omega)}{1+\bar{\gamma}(\lambda,\omega)}\right)\right)^{-1} \quad (10)$$

where q^p denotes *a priori* speech absence probability.

3.2. Suppress Estimated Localized Noises

The estimated spectra of localized noises are then subtracted from those of the observed noisy signals by employing the non-linear spectral subtraction to roughly estimate the speech spectra since the non-localized noise components can not be suppressed by this stage.

4. Further Suppress Non-Localized Noises With Post-Filtering

The rough enhanced speech signal obtained in the last section, containing the desired speech signal and non-localized noises, can be considered as a single-channel signal. This observation motivates us to use a state-of-art single-channel noise suppression approach as a post-filter to suppress the non-localized noises. In this section, a single-channel post-filter based on the OM-LSA estimator is adopted due to its independency of the correlation between the noises and its superiority in reducing "musical tones" [5].

5. Experiments and Discussions

In this section, we describe two experiments. One was devoted to evaluating the pure contribution of the hybrid noise estimation method. The second was to evaluate the performance of the proposed noise reduction method as a complete system.

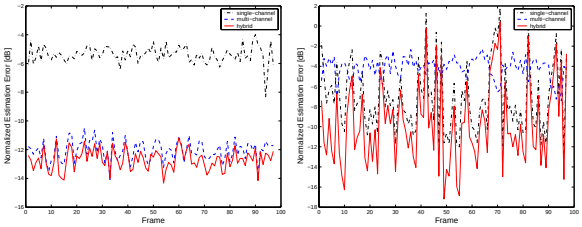


Figure 2: Normalized Noise Estimation Error (dB) for signals processed by single-channel technique (dashdot, average NEEs are -5.28 dB and -6.79 dB for white noise and car noise), multi-channel technique (dashed, average NEEs are -12.89 dB and -4.52 dB for white noise and car noise) and hybrid technique (solid, average NEEs are -13.34 dB and -9.70 dB for white noise and car noise) in white noise condition (left) and car noise condition (right).

5.1. Evaluation of the Hybrid Noise Estimation Approach

The aim of this experiment was to evaluate the performance of the suggested hybrid noise estimation approach. A speech sentence, selected from an ATR database, was corrupted by localized white Gaussian noise and localized car noise with DOA of 40 degrees to the right, respectively. All the sound data were re-sampled to 12kHz and linearly quantized 16 bits before mixing. The experiment was conducted under the following conditions: frame length was 256 samples, frame shift was 128 samples, $\varepsilon = 0.1$, $q^p = 0.5$ and $\alpha_n = 0.9$.

The estimation accuracy of the hybrid noise estimation approach was evaluated in terms of an objective measure: Normalized Estimation Error (NEE), given by:

$$\text{NEE} = \frac{1}{L}\sum_{\lambda=0}^{L-1} 20\log_{10}\frac{\sum_{\omega=0}^{K-1}\left(|\hat{N}(\lambda,\omega)-N(\lambda,\omega)|\right)}{\sum_{\omega=0}^{K-1}|N(\lambda,\omega)|} \quad (11)$$

where $\hat{N}(\lambda,\omega)$ and $N(\lambda,\omega)$ are estimated noise spectrum and "ideal" noise spectrum; L and K represent the number of frames in the signal and the number of samples per frame.

The comparisons of the tested noise estimation approaches are shown in Fig. 2. It is worth noting that the lowest estimation error was achieved by the hybrid noise estimation approach. This observation can be explained by the fact that the noise spectrum was properly estimated by the multi-channel or single-channel techniques in different frequency bins.

5.2. Evaluation of the Noise Reduction System

In this section, two sets of speech sounds were used to evaluate the proposed noise reduction system. One sound data set was generated by mixing 18 speech sentences, selected from an ATR database and uttered by 3 male and 3 female speakers, with localized and non-localized white Gaussian noise and car noise with DOA of 40 degrees to the right for the localized noise at various SNRs [-5,20] dB. A second set of sound data were recorded in the real-world car environment at the speed of 100km/h. The experimental conditions were

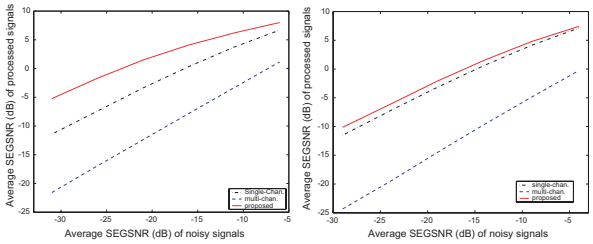


Figure 3: Average Segmental SNR (dB) for signals processed by single-channel technique (dashdot), multi-channel technique (dashed) and hybrid technique (solid) in white noise condition (left) and car noise condition (right).

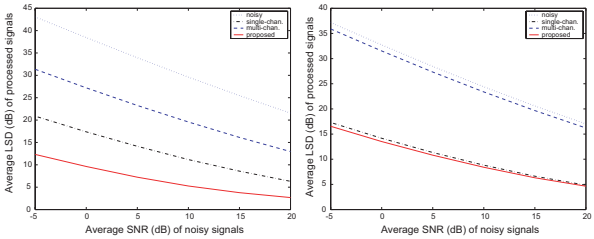


Figure 4: Average Log Spectral Distance (dB) for noisy signal (dotted) and for signals processed by single-channel technique (dashdot), multi-channel technique (dashed) and hybrid technique (solid) in white noise condition (left) and car noise condition (right).

same as those in the first experiment.

The performance of the proposed noise reduction system was evaluated in terms of two objective speech quality measures: segmental SNR (SEGSNR) and log spectral distance (LSD), given by:

$$\text{SEGSNR} = \frac{1}{L}\sum_{\lambda=0}^{L-1} 10\log \frac{\sum_{j=0}^{K-1}\left[s(\lambda K+j)\right]^2}{\sum_{j=0}^{K-1}\left[\hat{s}(\lambda K+j)-s(\lambda K+j)\right]^2} \quad (12)$$

$$\text{LSD} = \frac{1}{L}\sum_{\lambda=0}^{L-1}\left(\frac{1}{K}\sum_{\omega=0}^{K-1}\Big[20\log|S(\lambda,\omega)| - 20\log|\hat{S}(\lambda,\omega)|\Big]^2\right)^{\frac{1}{2}} \quad (13)$$

where $s(.)$ and $\hat{s}(.)$ are the reference speech signal and noisy signal or enhanced signal, and $\hat{S}(\lambda,\omega)$ and $S(\lambda,\omega)$ are the estimated speech spectrum and reference speech spectrum.

Figs. 3 and 4 show the performance improvements of the proposed noise reduction system in the SEGSNR and LSD senses compared to the multi-channel and single-channel based systems for the first sound data set. It is easy to see that the proposed noise reduction system achieved the highest SEGSNR improvement and lowest LSD in the localized and non-localized noise conditions.

Fig. 5 shows the waveforms and spectrograms of a typical noisy speech data /Asahi/ and its enhanced signal by the proposed noise reduction system in the real-world car environment. Obviously, the desired speech is corrupted by the car noise before processing. After processing, the car noise has been significantly reduced with less speech distortion, as see from the waveforms and spectrograms in Fig. 5.

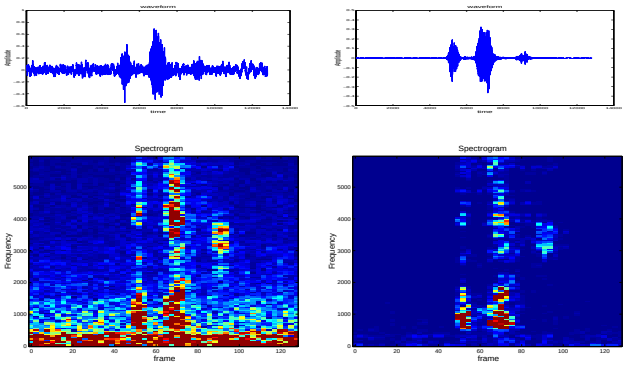


Figure 5: Waveforms (top row) and spectrograms (bottom row) of noisy signals (left column) and enhanced signals (right column) by the proposed noise reduction system. (Speech Data: /Asahi/).

Moreover, compared to the traditional methods (e.g. Delay-And-Sum beamformer), the superiority of the proposed method can be easily verified by the above observations and the fact that the multi-channel technique based system wins the DAS beamformer based system [2].

6. Conclusion

In this paper, a novel noise reduction system using hybrid noise estimation technique and post-filtering was proposed which was able to suppress both localized and non-localized noise components. The superiorities of the proposed hybrid noise estimation approach and noise reduction system were verified in various noise conditions.

7. Acknowledgment

This research is conducted as a program for the "Fostering Talent in Emergent Research Fields" in Special Coordination Funds for Promoting Science and Technology by Ministry of Education, Culture, Sports, Science and Technology.

8. References

[1] L.J. Griffiths and C.W. Jim, "An alternative approach to linearly constrained adaptive beamforming", IEEE Transactions on Antennas and Propagation, vol. Ap-30, pp. 27-34, 1982.

[2] M. Akagi and M. Mizumachi, "Noise Reduction By Paired Microphones", in EUROSPEECH97, pp. 335-338, 1997.

[3] M. Akagi and T. Kago, "Noise Reduction Using a small-Scale Microphone Array in Multi Noise Source Environment", in ICASSP'2002, pp. 909-912, 2002.

[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, Nov. 2003.

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