A SOFT DECISION BASED NOISE CROSS POWER SPECTRAL DENSITY ESTIMATION FOR TWO-MICROPHONE SPEECH ENHANCEMENT SYSTEMS

Xuefeng Zhang, Ying Jia

Intel China Research Center, Beijing, China, 100083
{xuefeng.zhang, ying.jia}@intel.com

ABSTRACT

The coherence function has been successfully applied to two-microphone speech enhancement systems. Recently, cross power spectral density (cross-PSD) estimation was proposed to further improve the coherence based two-microphone speech enhancement systems, especially for the case that the cross correlation between the two channel noises can not be ignored. In this paper, we propose a more robust noise cross-PSD estimation method based on an effective soft decision approach using minimum statistics, which can work well even in highly adverse conditions. Experimental results using real signals recorded in a running car demonstrate the proposed approach can obtain more less speech distortion with the same SNR improvement.

1. INTRODUCTION

The voice communication devices are usually disturbed by diffused noise. especially in mobile communications where hands-free devices are frequently used in noisy environments such as cars. In highly adverse conditions, the ambient noise may be even more powerful than speech, and thus has to be reduced.

Several methods have been proposed to reduce the ambient noise: The one-channel noise reduction algorithm, such as the spectral subtraction [1],[2] with its low computational load, has been investigated with success. Nevertheless, they lead to a compromise between residual noise and speech distortion, especially in the presence of high level noise. In recent years, some two-microphone array systems [3],[4] have been proposed to increase performance by considering spatial characteristics and therefore partially get rid of some hypotheses like noise stationary.

The two-microphone array system has the advantages of convenient placement and low cost. It appears to be a very promising method for speech enhancement. We assume that two observations are available; each one is composed of signal and noise. Whatever the distance between microphones is, the signals are strongly correlated, while the correlation between noises becomes rather weak for a sufficiently great distance. The coherence function is then a pertinent criterion to know whether a speech signal exists or not. R. Le Bouquin etc. [3] used the coherence function for speech enhancement in a noisy environment, which lead to effective noise reduction when the noises are totally uncorrelated. While, in practice, for a real noise field the cross power spectrum between the two channel noises is never exactly zero, especially when the noise is non-stationary. So the cross power spectral density between two noises has to be estimated and subtracted from the cross-PSD of two observed signals. Guerin etc.[5] utilized a fuzzy law to estimate the noise cross-PSD and obtain satisfying results; however its estimation quality is limited in highly adverse environment, such as the lower SNR and non-stationary noise existing.

In this paper we propose an effective soft-decision approach using minimum statistics to estimate the noise cross power spectral density. It can achieve a good estimation even in highly adverse conditions. Simulation experiment based on real signals recorded in a running car shows the proposed algorithm can obtain high noise reduction with less speech distortion.

2. COHERENCE BASED TWO-MICROPHONE SPEECH ENHANCEMENT ALGORITHM

The block diagram of the two-channel microphone speech enhancement algorithm is showed in figure 1. Here $H(f,k)$ represents the speech enhancement filter in frequency domain. Let $s_i(t)$ and $n_i(t)$ denote speech and uncorrelated additive noise signals in the i^{th} channel. The observed signals $x_i(t)$ are given by

$$x_i(t) = s_i(t) + n_i(t) \quad (i=1,2). \tag{1}$$

Their short-time discrete Fourier transform (STFT) are $X_1(f,k)$ and $X_2(f,k)$ respectively, here f is the frequency bin, and k represents the frame index.

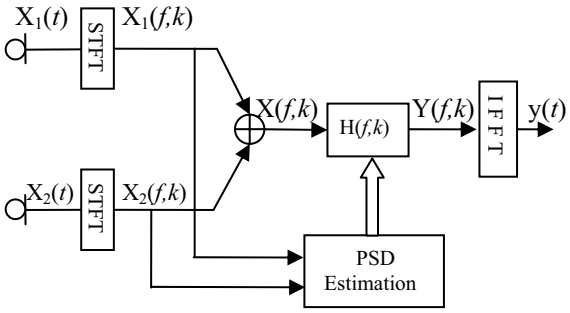


Fig.1. Block diagram of the noise reduction technique

The power spectral densities (PSD) of a signal $u(t)$ and the cross-PSD of two signals $u(t)$ and $v(t)$ are defined as:

$$\gamma_{uu}(f,k) = E[U(f,k)U^*(f,k)] \quad and$$
$$\gamma_{uv}(f,k) = E[U(f,k)V^*(f,k)]. \qquad (2)$$

Let $P_{n_i}(f,k)$, $P_{s_i}(f,k)$ and $P_{x_i}(f,k)$ denote the PSD of the noise, speech and noisy signals on the i^{th} channel respectively. $P_{x_1x_2}(f,k)$, $P_{n_1n_2}(f,k)$ denote the cross-PSD of two observed signals and noises. Then, the coherence between the two signals $x_1(t)$ and $x_2(t)$ is be given by:

$$\rho(f,k) = \frac{P_{x_1x_2}(f,k)}{\sqrt{P_{x_1}(f,k)P_{x_2}(f,k)}}. \qquad (3)$$

If the speech and noises are independent and meanwhile the noises of two channels are spatially uncorrelated, we can get:

$$\hat{P}_{s_1s_2}(f,k) = \hat{P}_{x_1x_2}(f,k). \qquad (4)$$

In this case, the speech signal can be estimated perfectly by filtering the signal $X(f,k)$ shown in Fig.1 using magnitude squared coherence function $|\rho(f,k)|^2$.

While, in practice, for a real noise field, the cross power spectrum between the noises of two channels is never exactly zero. In order to obtain a accurate estimation for the signal cross-PSD $\hat{P}_{s_1s_2}(f,k)$, the noise cross-PSD $P_{n_1n_2}(f)$ has to be estimated and subtracted from the cross-PSD of two observed signal $P_{x_1x_2}(f,k)$. In mathematical description:

$$\hat{P}_{s_1s_2}(f,k) = \left|P_{x_1x_2}(f,k)\right| - \left|P_{n_1n_2}(f,k)\right|. \qquad (5)$$

Therefore, we modify the coherence based filter as:

$$H(f,k) = \frac{\left|P_{x_1x_2}(f,k)\right| - \left|P_{n_1n_2}(f,k)\right|}{\sqrt{P_{x_1}(f,k)P_{x_2}(f,k)}}. \qquad (6)$$

The key task for us is to estimate the different PSD and cross-PSD. However, it is difficult to estimate the noise cross-PSD by regular method without precise priori knowledge about speech and noise. In section 3, we will introduce an effective soft-decision based noise cross-PSD estimation method, which avoid needing too much priori knowledge and vocal activity detector (VAD).

3. POWER SPECTRAL DENSITY ESTIMATION

3.1. PSD estimation

The noisy signal PSD $P_{x_i}(f,k)$ of each channel is estimated by a first order recursive system with a time and frequency dependent forgetting parameter $\lambda(f,k)$:

$$P_{x_i}(f,k) = \lambda(f,k)P_{x_i}(f,k-1) \qquad (7)$$
$$+(1-\lambda(f,k))X_i(f,k)X_i^*(f,k) \quad i=1,2$$

The value of the forgetting parameter plays an important role in the spectral estimation. On the one hand, the estimation has to respect the short-term speech stationary, and consequently λ should take low values; On the other hand, λ has to favor long-term estimation to reduce the estimator variance. Thus, the forgetting factor λ has to take small values during speech presence, and high values during noise only periods. Geurin [5] proposed the following law:

$$\lambda(f,k) = 0.98 - 0.3 \cdot \frac{SNR(f,k)}{1+SNR(f,k)}$$
$$\approx 0.98 - 0.3 \cdot H_{css}(f,k-1), \qquad (8)$$

This proposed law allows the residual noise to be controlled during noise only periods and the variation of speech to be tracked quickly during speech activity, which can effectively reduce the music noise and meanwhile keeps a small estimator variance.

The noise PSD $P_{n_i}(f,k)$ of each channel is estimated by using the minimum statistics method [1],[6]. In contrast to other estimation methods, the minimum statistics method does not use a voice activity detector. Instead it tracks spectral minima in each frequency band without distinction between speech activity and speech pause. Compared to other traditional approaches, the minimum statistics noise estimator has a superior ability to preserve weak speech sounds and therefore delivers a superior intelligibility.

3.2. Cross-PSD estimation

The noisy signal cross-PSD $P_{x_1x_2}(f,k)$ is estimated by a first order recursive system with a time and frequency dependent smoothing parameter expressed in (8):

$$P_{x_1x_2}(f,k) = \lambda(f,k)P_{x_1x_2}(f,k-1)$$
$$+(1-\lambda(f,k))X_1(f,k)X_2^*(f,k) \qquad (9)$$

As mentioned before, an accurate estimation of the noise cross-PSD $P_{n_1n_2}(f,k)$ is critical for the quality of the cross spectral subtraction algorithm. Guerin proposed to utilize a fuzzy law based on energetic considerations to implement a continuous noise estimate. In this method, noise is supposed to be a long-term stationary signal

unlike speech. Therefore, a large energy increase between two adjacent frames may be viewed as the presence of speech, whereas small variations only or a decrease in energy correspond more likely to noise. This assumption is very similar to the minimum statistics method, but it is not accurate as the minimum statistics method. It may lead to a bad speech distortion especially when the input SNR is low and the distance between two sensors is short. We propose a special soft-decision approach using minimum statistics to better estimate the noise cross power spectral density, which can obtain less speech distortion with similar noise reduction degree.

In view of ENERGETIC considerations, noise is supposed to be a long-term stationary signal unlike speech. So we can just update the noise cross-PSD estimation $P_{n_1n_2}(f,k)$ using $P_{x_1x_2}(f,k)$ in noise-only periods while preserve it in speech activity periods.

The noise-only periods can be detected by a special soft-decision approach: two-channel double-talk detection in the time-frequency space. First we compute the comparatively accurate estimation of noisy signal PSD estimation $P_{x_i}(f,k)$ and noise PSD estimation $P_{n_i}(f,k)$ using minimum statistics [6]. Then a simple decision is operated as: if the following inequation:

$$\frac{P_{n_i}(f,k)}{P_{x_i}(f,k)} \le threshold$$

holds for $i=1$ and $i=2$, we may think the period in frequency f and time k is a noise-only period, and act as:

$$P_{n_1n_2}(f,k) \leftarrow P_{x_1x_2}(f,k),$$

otherwise, the period is assumed to be a speech activity period, we preserve the estimation of the former frame:

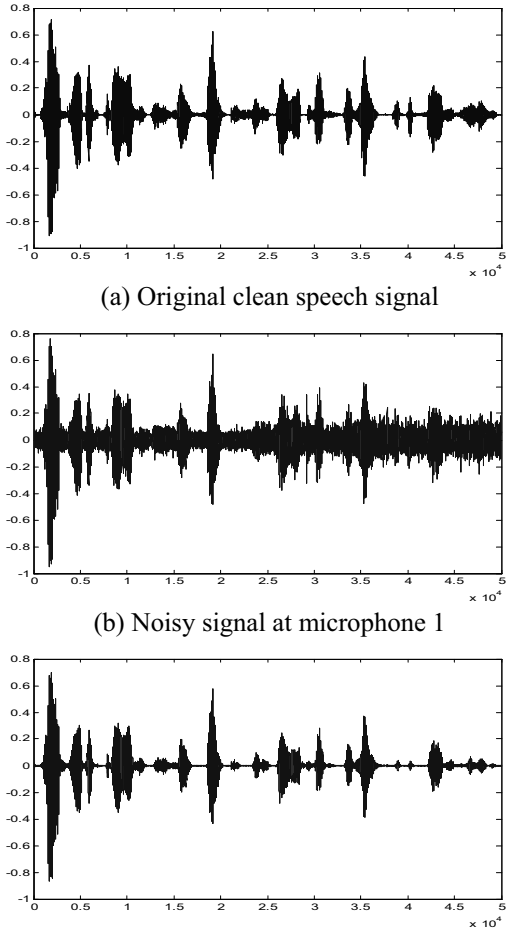
$$P_{n_1n_2}(f,k) \leftarrow P_{n_1n_2}(f,k-1).$$

Here the sign "$\leftarrow$"denotes evaluation operation.

4. EVALUATION

Experiments were conducted on real signals recorded in a running car. The distance between two microphones is 8cm, clean speech signals are recorded at a sampling rate of 12 KHz in absence of background noise (standing car, silent environment). The ambient noise signals, which are mainly composed of the engine, the contact between tires and road and the wind fluctuations noise, are recorded while the car speed is about 60km/h. The input microphone signals are generated by mixing the speech and noise signals with different SNR from -4dB to 14dB. We used 256 points Fast Fourier Transforms of hamming windowed signals, with 128 points overlapping. Figure 2 shows the time waveforms of speech enhancement experiment with input SNR (total) is 6dB.

From the noisy signal waveform we can see the noise is quasi-stationary. The observed signal is greatly corrupted by the noise in the back part. The enhancement result shown in Fig.2.(c) represents our proposed method work well with high noise reduction. Nevertheless, we can also find there are slight speech distortions in the weak speech frames, while this distortion is hardly audible.



(a) Original clean speech signal

(b) Noisy signal at microphone 1

(c) Enhanced result by the proposed method

Fig.2. Waveforms of enhancement experiment.
(a) Original clean speech signal.
(b) Noisy signal at microphone 1.
(c) Enhanced result by the proposed method

In order to quantificational compare the performance of the proposed approach and Guerin's method; two different measures have been evaluated on the results: the segmental SNR and the log spectral distance (LSD). The segmental SNR measure takes both noise suppression and speech distortion into account, which is defined by [7]:

$$SegSNR = \frac{1}{L}\sum_{l=0}^{L-1} 10 \cdot \log \frac{\sum_{n=0}^{N-1} s^2(n+\frac{lN}{2})}{\sum_{n=0}^{N-1}[s(n+\frac{lN}{2}) - \hat{s}(n+\frac{lN}{2})]^2}, \quad (10)$$

where L represents the number of frames in the signal, and N is the number of samples per frame. The SNR at every frame is limited to perceptually meaningful range between

35dB and -10dB. The LSD shows speech distortion degree. The L_2 norm of the log spectral distance id defined by [3]:

$$d_{L_2}(s,\hat{s}) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left|\log S(\omega) - \log \hat{S}(\omega)\right|^2 d\omega\,, \qquad (11)$$

Results of the segmental SNR and LSD measures are presented in figures 3 and 4, respectively. The experimental setup and parameters are the same to the above description.

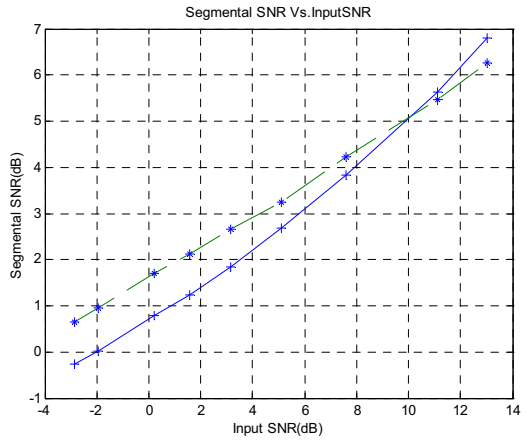


Fig.3. Segmental SNR vs. Input SNR.
The continuous lines correspond to the Guerin's method (+) and the dashed lines correspond to the improved method (*).

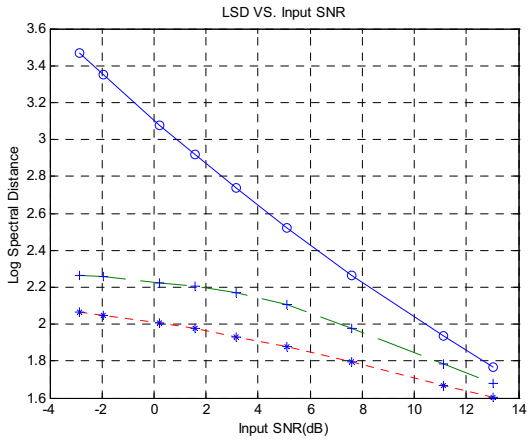


Fig.4. LSD vs. Input SNR.
The continuous lines correspond to the un-enhanced noisy speech (o), the dashed lines correspond to Guerin's method (+) and the dot lines correspond to the improved method (*).

From the two figures we can find the proposed algorithm can obtain a better estimation quality than Guerin's method when the input SNR < 10 dB. When the input SNR > 10 dB, the segmental SNR of Guerin's method have a slight improvement, while its LSD is still larger than the proposed one. This indicates Guerin's method can get higher output SNR under input SNR > 10 dB. However, the higher output SNR is at the cost of the speech distortion.

5. CONCLUSION

In order to further improve the enhancement performance of coherence based two-channel microphone array system, we proposed a more robust noise cross-PSD estimation approach. It can well estimate the noise cross-spectral densities between the two channels of microphone array even in lower input SNR and non-stationary noise existing. Simulation experiment based on real signals recorded in a running car showed the proposed algorithm ensuring high noise reduction with less speech distortion.

6. ACKNOWLEDGEMENT

The authors thank Dr. Wolfgang Herbordt for his helpful discussion.

7. REFERENCE

[1] R. Martin, "Spectral Subtraction Based on Minimum Statisics", *EUSIP-94*, September 1994, pp.1182-1185.

[2] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction", *IEEE Trans. on ASSP*, vol. 27, no. 2, pp. 113-120, 1979.

[3] R. Le Bouquin, G. Faucon, "Using the Coherence Function for Noise Reduction", *IEE Proc.* 139(3), June 1992, pp.276-280.

[4] R. Le Bouquin, A. A. Azirani, and G. Faucon, "Enhancement of Speech Degraded by Coherent and Incoherent Noise Using a Cross-Spectral Estimator," *IEEE Trans.on Speech and Audio Processing.,* 5(5), pp. 484-487, September 1997.

[5] A. Guerin, R. Le Bouquin, and G. Faucon, "A Two-Sensor Noise Reduction System: Applications for Hands-free Car Kit," in *EURASIP JASP 2003*:11 (2003), pp.1125-1134.

[6] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," in *IEEE Trans. on Speech and Audio Processing,* 9(5):504-512, July 2001.

[7] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, Discrete-Time Processing of Speech Signals, IEEE Press, New York, NY, USA, 2nd edition, 2000.