Journal of Zhejiang University SCIENCE A
ISSN 1673-565X (Print); ISSN 1862-1775 (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

A noise cross PSD estimator for dual-microphone speech enhancement based on minimum statistics*

Mohsen RAHMANI†, Ahmad AKBARI, Beghdad AYAD, Nima DERAKHSHAN

(*Research Center for Information Technology, Computer Department, Iran University of Science and Technology, Tehran, Iran*)

†E-mail: m-rahmani@araku.ac.ir

Received May 22, 2008; Revision accepted Oct. 10, 2008; Crosschecked Apr. 27, 2009

Abstract: Some two-microphone noise reduction techniques that work in the frequency domain exploit coherence function between two noisy signals. They have shown good results when noise signals on two sensors are uncorrelated, but their performance decreases with correlated noises. Coherence based methods can be improved when the cross power spectral density (CPSD) of correlated noise signals is available. In this paper, we propose a new method for estimation of the CPSD of the noise, which is based on the minimum tracking technique. Despite the fact that the proposed estimator does not need to implement a voice activity detector (VAD), its performance is comparable to a CPSD estimator that uses an ideal VAD.

Key words: Two-channel noise reduction, Noise estimation, Minima tracking
doi:10.1631/jzus.A0820390 **Document code:** A **CLC number:** TN912

INTRODUCTION

When a speech communication system is used in a noisy environment, the noise picked up by the microphone corrupts the speech signals. To obtain a speech signal that is more intelligible and more pleasant to listen to, a noise reduction system is needed. Coherence based methods are known as a subclass of dual-microphone methods that has shown good results in uncorrelated noise environments. The drawback of these methods is that their performance decreases if captured noises are correlated. To cope with this problem, La Bouquin-Jeannes *et al*.(1997) proposed subtracting the cross power spectral density (CPSD) of noises from the CPSD of the noisy signals. The main difficulty of this improvement is to estimate the CPSD of received noise signals. Guerin *et al*. (2003) proposed estimating the noise CPSD as a function of a posteriori signal-to-noise ratio (SNR) and the CPSD of the noisy signal, based on the premise that the noise CPSD can be estimated in all frames. Zhang and Jia (2005) proposed a soft decision based technique for noise CPSD estimation during speech pauses. They employed the minimum statistics on each channel to estimate the noise PSD in that channel. They exploited these estimated noise PSDs as a criterion to distinguish between speech and pause frames.

In this paper, we present a minimum tracking technique (Martin, 1994; 2001) for noise CPSD estimation. This method is an extension of the single channel minimum tracking method (Martin, 1994) to two-microphone approaches.

CROSS POWER SPECTRAL SUBTRACTION

In two-microphone noise reduction systems, the received signal on microphone i can be written as

$$X_i(f,n)=S_i(f,n)+N_i(f,n), \quad i=1, 2, \tag{1}$$

where $X(f,n)$, $S(f,n)$, and $N(f,n)$ show the noisy speech, the clean speech, and the noise in STFT domain,

* Project supported by the Iran Telecommunications Research Center (ITRC)

respectively. Furthermore, f and n are the frequency and the frame indexes, respectively. We use both channels to compute the spectral modification filter to be applied to the first channel. This filter can be computed based on the following coherence function:

$$\Gamma_{X_1X_2}(f,n)=\frac{\left|P_{X_1X_2}(f,n)\right|}{\sqrt{P_{X_1X_1}(f,n)P_{X_2X_2}(f,n)}}, \quad (2)$$

where $P_{X_1X_1}(f,n)$, $P_{X_2X_2}(f,n)$, and $P_{X_1X_2}(f,n)$ are PSD of $X_1(f,n)$, PSD of $X_2(f,n)$, and CPSD of $X_1(f,n)$ and $X_2(f,n)$, respectively. The CPSD and the PSDs can be estimated as in (La Bouquin-Jeannes *et al.*, 1997):

$$P_{X_iX_j}(f,n)=\lambda P_{X_iX_j}(f,n-1)+(1-\lambda)X_i(f,n)X_j^*(f,n), \quad i,j=1,2, \quad (3)$$

where λ is a smoothing factor in the range [0, 1] and is set to 0.7.

In the coherence based methods, it is assumed that the received noise signals are uncorrelated, which is often violated in the real world. La Bouquin-Jeannes *et al.*(1997) adapted the coherence methods to reduce the effect of coherent noises. In this approach, starting from the coherence function defined in Eq.(2), the estimated CPSD of noise signals is subtracted from the CPSD of noisy signals in the numerator to obtain a noise reduction filter $H_{\mathrm{CPSS}}(f,n)$:

$$H_{\mathrm{CPSS}}(f,n)=\frac{\left|P_{X_1X_2}(f,n)\right|-\left|P_{N_1N_2}(f,n)\right|}{\sqrt{P_{X_1X_1}(f,n)P_{X_2X_2}(f,n)}}. \quad (4)$$

A precise estimation of the noise CPSD, $\left|P_{N_1N_2}(f,n)\right|$, is crucial to obtain an accurate estimation of the speech signal.

CPSD NOISE ESTIMATION USING MINIMA TRACKING

In this section, we describe our approach for CPSD minima tracking. The proposed noise CPSD estimator operates as follows.

The smoothed noisy CPSD, $R_{X_{12}}(f,n)$, is computed as

$$R_{X_{12}}(f,n)=\gamma(f,n)R_{X_{12}}(f,n-1)+(1-\gamma(f,n))X_1(f,n)X_2^*(f,n). \quad (5)$$

The difference between Eqs.(3) and (5) is that in Eq.(5) the smoothing factor, $\gamma(f,n)$, is estimated for each frequency bin and each frame. The amplitude of the smoothing parameter must have two features: (1) $\left|R_{X_{12}}(f,n)\right|$ has to be able to track the non-stationary behavior of the CPSD; meanwhile, it decreases the variance in noise-only frames. (2) The smoothed CPSD should return to the noise CPSD level in the gaps between speech components, so that noise power can be estimated in these regions (Derakhshan *et al.*, 2007). In (Martin, 1994; Derakhshan *et al.*, 2007) the smoothing parameter is controlled by time and frequency dependent SNR to perform more smoothing on the noise-only regions. In our approach, smoothing is done such that the lower the coherence, the more smoothing is performed. The empirical value for the smoothing factor is determined as

$$\gamma(f,n)=\begin{cases}\sqrt{1-t_1}, & \Gamma_{X_1X_2}(f,n)\le t_1,\\ \sqrt{1-\Gamma_{X_1X_2}(f,n)}, & t_1<\Gamma_{X_1X_2}(f,n)<t_2,\\ \sqrt{1-t_2}, & \Gamma_{X_1X_2}(f,n)\ge t_2,\end{cases} \quad (6)$$

where t_1 and t_2 are two thresholds that are empirically set to 0.2 and 0.8, respectively.

The short-time noise CPSD is estimated by finding the local minima on the CPSD of the noisy signals, $R_{X_{12}}(f,n)$. These values are found on the short intervals of L frames:

$$R_{X_{12}\min}(f,n)=\min\left\{\left|R_{X_{12}}(f,m)\right|, n-L+1<m<n\right\}, \quad (7)$$

where L is the length of the search window and is set to the maximum lifetime of the speech components. Since the short-term minimum of amplitude CPSD is always smaller than the mean of amplitude CPSD, the local minima of $R_{X_{12}}(f,n)$ is a biased estimate of the amplitude CPSD that decreases the noise CPSD estimation accuracy. For an accurate noise CPSD estimate this bias must be compensated. For this reason, it is multiplied by a value named the 'bias compensation factor', $B_{\min}$:

$$\left|\hat{P}_{N_1N_2}(f,n)\right| = R_{X_{12}\min}(f,n)B_{\min}. \quad (8)$$

In the case of single-channel noise PSD estimation, La Bouquin-Jeannes *et al*.(1997) used a constant bias compensation factor to correct the error between the estimated and the true noise PSDs. Martin (2006) showed that the bias compensation factor is a function of the search window length, L, and the noise PSD estimate variances. Here, $B_{\min}$ is set to be a constant value for a fixed minima search window.

The block diagram of the noise reduction method is shown in Fig.1.

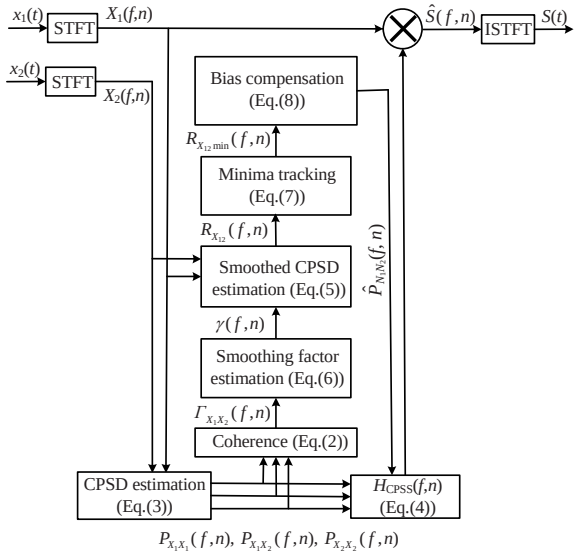

Fig.1 Block diagram of the noise reduction method

In this block diagram, the CPSD and coherence values are estimated using Eqs.(3) and (2) respectively, and then the smoothing parameter, $\gamma(f,n)$, is estimated using the coherence between noisy signals as in Eq.(6). The smoothed CPSD is then estimated using Eq.(5). In the next steps, the noise CPSD is estimated by minima tracking using Eqs.(7) and (8). The estimated noise CPSD is exploited to calculate the noise reduction filter as in Eq.(4). Finally, this filter is applied to the noisy signal from channel 1. The enhanced speech signal in the time domain is computed using an inverse FFT.

A known drawback of the minimum tracking approach is the inability of the PSD estimation to track fast increases in the noise power (Derakhshan *et al*., 2007). This problem arises because a spectrum smoothing is not performed before minima tracking. To cope with this problem, in (Martin, 2001; Cohen, 2002; Derakhshan *et al*., 2007) the smoothing parameter is controlled by either the past estimations of noise or another stage of minima tracking. We instead use a coherence-dependent smoothing factor. The advantage is that the smoothing factor is independent of the noise estimation in previous frames, and thus previous incorrect noise estimations do not affect the smoothing procedure.

Fig.2 shows the noise CPSD estimated using the minimum tracking technique where the level of noise on both channels is increased and then decreased synchronously. The increase in noise power is 40 dB, which results in a 40 dB increase in noise CPSD. Noise CPSD is estimated in the presence of speech signal. Parameter $\gamma(f,n)$ in Eq.(6) is set to a fixed value (dashed line) and to the variable value by Eq.(6) (solid line). True noise CPSD is depicted as a dotted line. In transition regions where the true noise CPSD increases or decreases suddenly, $\gamma(f,n)$ obtained by Eq.(6) introduces less error and hence is more acceptable than a fixed smoothing factor. When the noise CPSD is low, the coherence value is high and so the smoothing value is low; as a result the smoothed noisy CPSD tracks changes fast. This is why the estimation with an adaptive smoothing parameter tracks the true CPSD faster when the noise level increases. It is also seen in some cases, for example when time t=3~5 s, that the estimation with a fixed parameter is closer to the true noise CPSD compared with variable $\gamma(f,n)$ obtained by Eq.(6). But the latter is the method of choice, especially at the transition point.

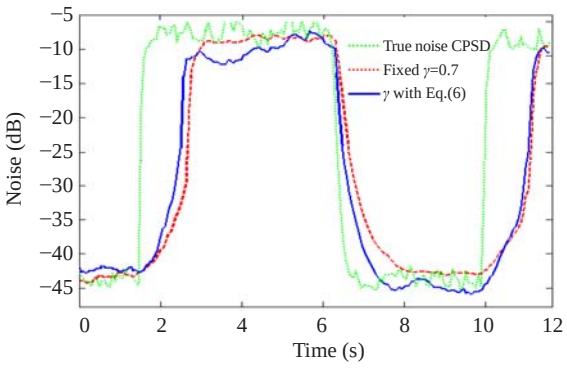

Fig.2 Performance of the noise CPSD estimators obtained by taking the average over all frequencies

EVALUATION

We used a speech database recorded using three omni-directional microphones installed on a headset on a dummy head to evaluate implemented methods. The clean speech was played from a speaker installed on the mouth of the model's head. In each experiment, we used two microphones simultaneously. Fig.3 shows the position of the microphones, where the distance between microphones 1 and 2 is 20 mm and the distance between microphones 1 and 3 is 66 mm.

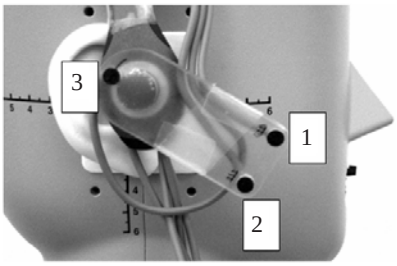

Fig.3 Position of microphones

To generate noisy signals, two noise types were added to speech signals: car noise and babble noise. Car noise was recorded in a Peugeot 405, driving at the speed of about 80 km/h. The babble noise was recorded in a noisy cafeteria. The noise signals were added to their corresponding speech signals (noise from microphone *i* was added to speech from microphone *i*).

Four methods were compared (Table 1). The first method was a single-channel Wiener filter, where the noise PSD is estimated using a minima statistics tracking method proposed in (Derakhshan *et al*., 2007). The second was CPSS-VAD, which uses the noise reduction filter in Eq.(4), where $\left|P_{N_1N_2}(f,n)\right|$ is estimated in noise-only frames. In this technique, noise-only frames are labeled manually. The third was the CPSS-FXD method, which uses the same noise reduction filter as in Eq.(4), but the noise CPSD is estimated using a minimum tracking approach where the smoothing factor value is fixed at 0.7. Finally, the fourth method, CPSS-MTR, uses the proposed method for CPSD noise estimation. In all methods, signals are sampled at 8 kHz and split into 50 percent overlapping frames of 256 samples. Each frame is multiplied by a hamming-window function. After modification, the frames are reassembled to create a continuous output signal. In the reconstruction step, overlapping of frames is considered. The parameters of minimum tracking algorithms are empirically set; the length of the search window is 1 s and the bias compensation factor is set to 2.5.

The evaluation results are shown in Table 1. PESQ (ITU-T P.862, 2001) is used for evaluation, which is a psychoacoustics-based objective measure originally proposed to assess the performance of the codecs. Each score is calculated by averaging over four PESQ scores of four utterances with the same SNR. To have a sense of PESQ scoring on our signals, we also report the PESQ scores for noisy signals in Table 2.

The PESQ scores for two-microphone methods are higher than those of the Wiener method. This was an expected result, since two-microphone methods use more information compared to single-channel methods. The PESQ values for the proposed method are higher than those for the minima tracking with a fixed smoothing factor. This is also observed even with a manually defined VAD, where the PESQ scores of CPSS-MTR are comparable with those of CPSS-VAD. This can be explained by the fact that CPSS-VAD does not take the behavior of the noise during the speech frames into account, while the minimum tracking estimates the noise CPSD in sub-bands during speech and non-speech frames.

Table 1 *PESQ* results for enhanced signals for input *SNR*=0, 5, and 10 dB

Method	*PESQ*					
	Microphones (1, 2)*			Microphones (1, 3)**		
	SNR=0	5	10	*SNR*=0	5	10
Wiener filter	2.244 (1.327)	2.432 (1.666)	2.581 (2.029)	2.282 (1.367)	2.493 (1.690)	2.638 (2.032)
CPSS-VAD	2.585 (1.382)	2.681 (1.870)	2.810 (2.361)	2.679 (1.510)	2.826 (2.033)	2.946 (2.452)
CPSS-FXD	2.564 (1.342)	2.591 (1.886)	2.761 (2.381)	2.573 (1.479)	2.757 (1.892)	2.889 (2.243)
CPSS-MTR	2.571 (1.355)	2.721 (1.865)	2.789 (2.343)	2.608 (1.596)	2.780 (1.982)	2.908 (2.391)

Car noise (babble noise). * With a distance of 20 mm; ** With a distance of 66 mm

Table 2 *PESQ* scores for noisy signals (with the input *SNR*=0, 5, 10, 15, and 20 dB)

Noisy signal type	*PESQ*				
	SNR=0	5	10	15	20
Car noise	2.103	2.338	2.491	2.685	2.921
Babble noise	1.297	1.560	1.884	2.278	2.611

To validate the objective performance evaluation, an informal listening test was conducted. We asked five listeners to judge the quality of 18 sets of sentences (6 conditions×3 sentences). Each listener listened to clean speech, noisy speech, and three enhanced signals. The enhanced speech signals were produced by CPSS-VAD, CPSS-FXD, and CPSS-MTR. Each listener gave a score between one (poor) and five (excellent) to each enhanced signal. This scale corresponds to the mean opinion score (MOS) scale presented in (Deller *et al*., 2000). It represents the listener's general appreciation of both the speech distortion and the residual noise. We also wanted listeners to be careful about musical noise. Listeners used headphones during experiments. The listening test results are presented in Fig.4.

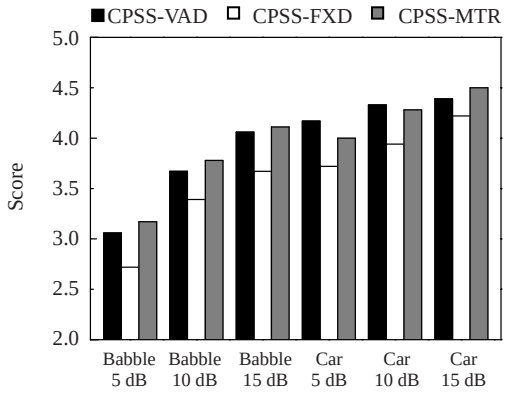

Fig.4 Listener test scores for enhanced signals

The results are reported for microphone pair (3, 4), averaged over 15 scores (5 listeners×3 sentences). Noise type: babble noise and car noise; input *SNR*: 5, 10, and 15 dB

In most cases, listeners preferred CPSS-MTR over the other methods (Fig.4). In some cases the listeners' scores for CPSS-MTR were comparable with those for an ideal VAD. This confirms the good performance and effectiveness of the proposed method.

CONCLUSION

In this paper, we proposed a minimum tracking approach to estimate the noise CPSD in double-microphone enhancement methods. A dynamic smoothing factor based on the coherence function was employed for estimating noise CPSD. Our proposed method is able to estimate the noise CPSD well without using VAD. When we exploited this estimator in a double-microphone noise reduction system, we obtained enhanced speech signals comparable to those obtained using a VAD based technique.

References

Cohen, I., 2002. Noise estimation by minima controlled recursive averaging for robust speech enhancement. *IEEE Signal Processing Lett.*, **9**(1):12-15. [doi:10.1109/97.988717]

Deller, J.R., Hansen, J.H.L., Proakis, J.G., 2000. Discrete-time Processing of Speech Signals (2nd Ed.). IEEE Press, New York, USA.

Derakhshan, N., Ayatollahi, A., Akbari, A., Rahmani, M., 2007. Noise Power Spectrum Estimation Using Time-variant Spectral Smoothing and Low-delay Minima Tracking. SPECOM, p.542-548.

Guerin, A., La Bouquin-Jeannes, R., Faucon, G., 2003. A two-sensor noise reduction system: applications for hands-free car kit. *EURASIP J. Appl. Signal Processing*, **2003**(11):1125-1134. [doi:10.1155/S1110865703305098]

ITU-T P.862, 2001. Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-end Speech Quality Assessment of Narrow-band Telephone Networks and Speech Codecs. International Telecommunication Union, Geneva.

La Bouquin-Jeannes, R., Azirani, A.A., Faucon, G., 1997. Enhancement of speech degraded by coherent and incoherent noise using a cross-spectral estimator. *IEEE Trans. Speech Audio Processing*, **5**(5):484-487. [doi:10.1109/89.622576]

Martin, R., 1994. Spectral Subtraction Based on Minimum Statistics. 7th European Signal Processing Conf., p.1182-1185.

Martin, R., 2001. Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Trans. Speech Audio Processing*, **9**(5):504-512. [doi:10.1109/89.928915]

Martin, R., 2006. Bias compensation methods for minimum statistics noise power spectral density estimation. *Signal Processing*, **86**(6):1215-1229. [doi:10.1016/j.sigpro.2005.07.037]

Zhang, X., Jia, Y., 2005. A Soft Decision Based Noise Cross Power Spectral Density Estimation for Two-microphone Speech Enhancement Systems. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, p.813-816. [doi:10.1109/ICASSP.2005.1415238]