Residual noise Control for Coherence Based Dual Microphone Speech Enhancement

Behzad Zamani
Islamic Azad University – Sharekord Branch
bzamani@iust.ac.ir

Mohsen Rahmani
Iran University of Science and Technology,
m_rahmani@iust.ac.ir

Ahmad Akbari
Iran University of Science and Technology,
akbari@iust.ac.ir

Abstract

Among various enhancement methods, dual microphone methods are utilized for their low cost implementation and use of spatial noise reduction. In this paper, we present a double microphone noise reduction method for nearby microphones. Our method is based on a modified coherence based filter which reduces correlated noises using cross spectral subtraction. In our proposed method the amount of noise reduction is controlled in order to achieve an acceptable compromise between residual noise and speech distortion. The performance of the enhancement technique is carried out using PESQ measure. Obtained results confirm the ability of the proposed approach to reduce distortion of signal.

1. Introduction

Over the last decades, by developing new communication systems, speech enhancement has become more important than before. For instance, in new mobile headsets, because of the distance between the mouth and the microphone, the received signal is noisy. A noise reduction system helps to increase the quality of noisy speech.

Speech enhancement systems need to have acceptable residual noise and acceptable speech distortion. Some of frequency based noise reduction systems use a spectral modification filter to modify the noisy speech spectrum and remove the background noise. In these methods, distortion of speech signal and residual noise are two well-known annoying effects which can not be simultaneously minimized [1]. In [1] some compromises between noise reduction and speech distortion are presented. Some other methods are also presented later in the literature [2],[3].

General speech enhancement approaches are divided into two main categories: single microphone and multi microphone methods. Some known single microphone methods are spectral subtraction [4], Wiener filter [5] and minimum mean square error [6]. Single microphone methods have some limitations in real environments. These methods induce musical noise and speech distortion [7]. The advantage of multi-microphone approaches is their ability to exploit spatial noise reduction, which is reduction of noise based on the knowledge of the position of speech sources. Some multi-microphone methods are independent component analysis [8], beamforming techniques [9][10] and generalized sidelobe canceling [11][12]. The performance of multiple channel noise reduction algorithms is improved by increasing number of microphones [9][10]. But larger number of microphones implies higher costs and increasing demands in computational load. We have chosen dual microphone approaches as the lowest cost multi channel methods.

Coherence based methods [13],[14] are known as a subclass of the dual microphone methods, which have shown good results in uncorrelated noise environments. But their performance decreases if captured noises are correlated. To cope with this problem, it is proposed in [13][15] to subtract the Cross Power Spectral Density (CPSD) of noises from the CPSD of the noisy signals. Hereafter, we refer this method as Cross Power Spectral Subtraction (CPSS).

In this paper, we modify the CPSS filter to reduce the distortion of enhanced signal, keeping an accepted residual noise. Results show that by accepting some residual noise and reducing distortion, recognition rate is increased.

2. Basic Two Microphone Speech Enhancement

In our system, it is assumed that each microphone received both noise and speech. For microphone *i* in STFT domain, the received signal can be written as:

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

978-0-7695-3504-3/08 $25.00 © 2008 IEEE
DOI 10.1109/ICCEE.2008.56

Where *X(f,n)*, *S(f,n)* and *N(f,n)* show the noisy speech, the clean speech and the noise in STFT domain, respectively. Furthermore, *f* and *n* are frequency and frame indexes, respectively. We can use both of X_1 and X_2 in equation (1) in order to compute the spectral modification filter for enhancing speech and reducing noise. This procedure is shown in Fig 1. As it can be seen in the figure, we apply the computed filter, *H(f,n)* on the first channel.

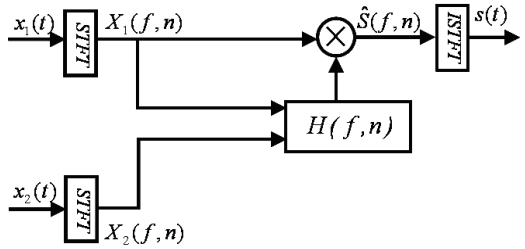

Fig. 1: Block diagram of the two-microphone speech enhancement

Spectral modification filter *H(f,n)* can be computed using the coherence function between two channels. The magnitude coherence function of two signals X_1 and X_2 is defined as:

$$\Gamma_{X1X2}(f,n)=\frac{\left|P_{X_1X_2}(f,n)\right|}{\sqrt{P_{X_1X_1}(f,n)P_{X_2X_2}(f,n)}} \quad (2)$$

In this equation, $P_{X1X1}(f,n)$, $P_{X2X2}(f,n)$ and $P_{X1X2}(f,n)$ are power spectrum density of $X_1(f,n)$, power spectrum density of $X_2(f,n)$ and cross PSD of $X_1(f,n)$ and $X_2(f,n)$, respectively. The cross power spectral density and the power spectrum densities can be estimated as:

$$P_{XiXj}(f,n)=\lambda_x P_{XiXj}(f,n-1)+(1-\lambda_x)X_i(f,n)X_j(f,n)^* \qquad i,j\in\{1,2\} \quad (3)$$

Where λ_x is a smoothing factor in the range [0, 1], with a recommended value of 0.7 [15].

In the coherence based methods, the frequency components of the filter vary according to the amount of coherence between channels. It is assumed that the received speech signals are correlated and the received noise signals are uncorrelated. Thus, higher values of coherence function correspond to increased level of desired speech in the signal. On the contrary, lower values of the coherence function correspond to increased level of noise in the signal. These properties lead us to use coherence function in speech enhancement systems.

This assumption that the received noise signals are uncorrelated is often not a valid assumption. In [15, authors proposed a way to improve the coherence based method. They adapted coherence based methods for correlated noises. In this approach, an estimation of noise CPSD signals is subtracted from CPSD of noisy signals:

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

The subtraction in the numerator of Equation (4) is done in order to achieve the CPSD of clean signals **Error! Reference source not found.**:

$$H_{CPSS}(f,n)=\frac{\left|P_{S_1S_2}(f,n)\right|}{\sqrt{P_{X_1X_1}(f,n)P_{X_2X_2}(f,n)}} \quad (5)$$

2.1. Using a priori SNR estimation

Like spectral subtraction method, this subtraction may induce musical noise and high distortion in speech signal. To reduce the amount of musical noise and distortion, we propose to employ a decision directed approach **Error! Reference source not found.**. In our approach, the numerator of the filter is changed as (7)

$$\begin{aligned}\left|P_{S_1S_2}(f,n)\right| &= \left|P_{X_1X_2}(f,n)\right|\frac{\left|P_{S_1S_2}(f,n)\right|}{\left|P_{S_1S_2}(f,n)\right|+\left|P_{N_1N_2}(f,n)\right|}\\ &= \left|P_{X_1X_2}(f,n)\right|\frac{SNRcs(f,n)}{SNRcs(f,n)+1}\end{aligned} \quad (6)$$

SNRcs(f,n) is the ratio between the clean speech CPSD and the noise CPSD, we call it cross SNR. To estimate the cross SNR, a technique similar to a-priori SNR estimation is employed. The advantages of a priori signal to noise ratio are shown in many references such as [21]. Using a priori SNR, we can reduce the amount of musical noise and distortion. A priori cross SNR, similar to single channel a priori SNR [6], is estimated as follow:

$$\begin{aligned}Rpo(f,n) &= max[\frac{\left|X_1(f,n)X_2(f,n)^*\right|}{\left|P_{N1N2}(f,n)\right|}-1]\\ Rpr(f,n) &= \lambda_{DD}H(f,n-1)^2\frac{\left|X_1(f,n-1)X_2(f,n-1)^*\right|}{\left|P_{N1N2}(f,n-1)\right|}+(1-\lambda_{DD})Rpo(f,n)\end{aligned} \quad (7)$$

λ_{DD} is smoothing in the rang of [0-1], which is set to values close to one. For example, it is set to 0.97 in [6].

Using equations (5) and (7) and considering cross SNR estimation in equation (8), $H_{CPSS}(f,n)$ is modified as follow:

$$H_{CPSS,Rpr}(f,n)=\frac{\left|P_{X_1X_2}(f,n)\right|}{\sqrt{P_{X_1X_1}(f,n)P_{X_2X_2}(f,n)}}\frac{Rpr(f,n)}{Rpr(f,n)+1} \quad (8)$$

2.1. Second-order headings

A precise estimation of the noise CPSD is crucial in order to obtain an accurate estimation of the speech

signal. A technique for estimation of the noise CPSD is to update the noise estimation in non speech regions and freeze it in speech activity. This can be done using a VAD. A VAD based technique, estimates CPSD as follow:

$$\left|\hat{P}_{N1N2}(f,n)\right| = \begin{cases} \lambda_n \left|\hat{P}_{N1N2}(f,n-1)\right| + (1-\lambda_n)\left|X_1(f,n)X_2(f,n)^*\right| & pause \\ \left|\hat{P}_{N1N2}(f,n-1)\right| & speech\ frames \end{cases} \quad (9)$$

Where, λ_n is a smoothing parameter between 0 and 1. Single or dual microphone VADs can be used to distinguish the speech/pause regions. We use a coherence based VAD [16], in that, the speech/pause regions are determined by tacking a threshold on the coherence function. This is based on this fact that, the coherence values for speech frames are more than the coherence values for noise only frames.

3. Residual Noise Control

Spectral modification filters are usually applied with the purpose of attenuating all noise. Eliminating noise in the spectral domain generates musical noise and speech distortion. By controlling the amount of noise reduction, we can reduce these effects. To control the level of noise reduction using a single channel Wiener filter, we can use the following filter [2]:

$$H_{Winer,\alpha}(f,n) = \frac{P_{S_1S_1}(f,n) + \alpha P_{N_1N_1}(f,n)}{P_{S_1S_1}(f,n) + P_{N_1N_1}(f,n)} \quad (10)$$

In which, α is a parameter for noise reduction control and its value varies between 0 and 1. In ideal noise estimation, the level of background noise reduction can be written as a function of α, as shown in the following equation [2].

$$NR_{dB} = 10log(\alpha^2 P_N) - 10log(P_N) = 20log(\alpha) \quad (11)$$

Where, PN is the power of background noise in the noisy signal. For example, for adjusting filter gain to reduce 10db of background noise, α must be equal to 0.32. Using α=0 denotes the conventional Wiener filter, where the estimated noise is supposed to be removed completely. In the case α=1, the filter is equal to unity and there is no noise reduction.

As we mentioned, the CPSS is a successful method in noise correlated environments, but it makes speech distortion compared to primitive method of coherence. Our goal is to extend the single channel residual noise control to CPSS method. To control the level of noise reduction in the CPSS method, we propose the following filter. Using this filter, we can have a compromise between residual noise and speech distortion.

$$H_{Rpr,\alpha}(f,n) = \frac{|P_{S_2S_1}(f,n)| + \alpha |P_{N_2N_1}(f,n)|}{\sqrt{P_{X_1X_1}(f,n)P_{X_2X_2}(f,n)}} \quad (12)$$

It must be noted, here in the case α=1, the modification filter is not unity and the filter reduces noise. This is because of the spatial noise reduction of the coherence function. But, using parameter α, we can tune the filter to values between the CPSS and the primitive coherence function. Thus, we can control the level of noise reduction. Figure 1 shows the black diagram of the proposed method.

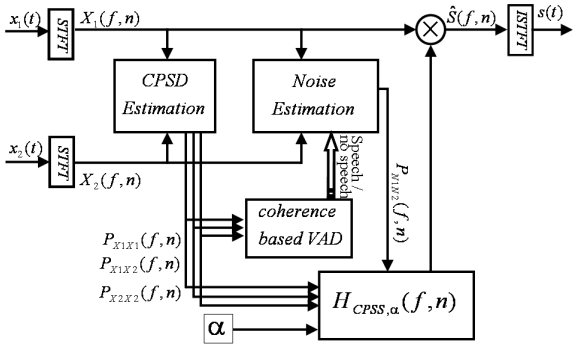

Fig. 2: Block diagram of the proposed method.

In order to exploit a priori benefits, Equation (12) can be rewritten as:

$$H_{Rpr,\alpha}(f,n) = \frac{|P_{X_2X_1}(f,n)|}{\sqrt{P_{X_1X_1}(f,n)P_{X_2X_2}(f,n)}} \frac{Rpr_{cc} + \alpha}{Rpr_{cc} + 1} \quad (13)$$

4. Evaluation

We used a realistic recording to evaluate the proposed method. We have used a speech database recorded using 4 microphones installed on a headset on a dummy head. The clean speech is played from speaker installed on the mouth of the head. In each experiment, we have used 2 microphones simultaneously. The distances between employed microphones are as follow:

- Microphones 1 and 2: It's related to the size of the head. For our model, it is 180mm.
- Microphones 1 and 3: It's fixed to 66mm.
- Microphones 3 and 4: It's fixed to 20mm.

The microphone position is shown in Fig 3.

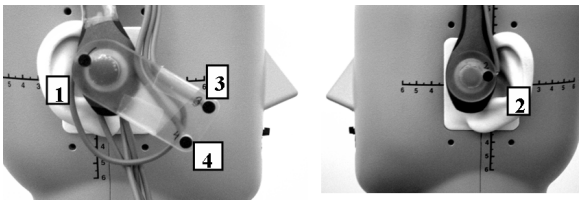

Fig 3: position of microphones on the head.

To generate noisy signals, two noise types have been added to speech signals: car noise and babble noise. Car noise has been recorded in a Samand (an Iranian car similar to Peugeot 405) with the speed of about 80 km/h. Babble noise has been recorded in a noisy cafeteria. The noise signals have been added to their corresponding speech signals (noise from microphone *i* is added to speech from microphone *i*).

The performance assessment is carried out using PESQ (P.862) measure [22] which is a psychoacoustics-based objective measure originally proposed to assess the performance of the codecs. Table 1 shows the PESQ score for noisy and enhanced signals. The results are reported for three microphone pairs and 3 input SNRs.

Table 1: PESQ scores for noisy and enhanced signals.

a) microphone pair (1,2)

	babble			***car***		
Input SNR (dB)	0	5	10	0	5	10
noisy signal (microphone 1)	1.495	1.671	2.308	1.246	1.423	2.068
Eq. (14), α=0	2.065	2.295	2.677	1.818	2.049	2.427
Eq. (14), α=0.1	2.224	2.403	2.777	1.98	2.157	2.527
Eq. (14), α=0.2	2.11	2.482	2.916	1.86	2.23	2.664
Eq. (14), α=0.3	2.071	2.432	2.94	1.826	2.177	2.692
Eq. (14), α=0.4	1.966	2.378	2.857	1.714	2.122	2.602
Eq. (14), α=1	1.646	1.837	2.564	1.398	1.582	2.306

b) microphone pair (1,3)

	babble			***car***		
Input SNR (dB)	***0***	***5***	***10***	***0***	***5***	***10***
noisy signal (microphone 1)	1.494	1.671	2.309	1.246	1.423	2.068
Eq. (14), α=0	2.083	2.261	2.563	1.84	2.008	2.314
Eq. (14), α=0.1	2.214	2.393	2.682	1.97	2.142	2.427
Eq. (14), α=0.2	2.149	2.468	2.767	1.894	2.216	2.52
Eq. (14), α=0.3	2.079	2.407	3.042	1.825	2.153	2.8
Eq. (14), α=0.4	2.051	2.365	2.909	1.8	2.113	2.662
Eq. (14), α=1	1.604	1.771	2.459	1.356	1.523	2.212

c) microphone pair (3,4)

	babble			***car***		
Input SNR (dB)	***0***	***5***	***10***	***0***	***5***	***10***
noisy signal (microphone 3)	1.499	1.674	2.31	1.254	1.428	2.062
Eq. (14), α=0	2.085	2.26	2.566	1.839	2.007	2.31
Eq. (14), α=0.1	2.218	2.399	2.682	1.965	2.148	2.426
Eq. (14), α=0.2	2.145	2.464	2.769	1.898	2.215	2.522
Eq. (14), α=0.3	2.079	2.402	3.045	1.822	2.155	2.798
Eq. (14), α=0.4	2.049	2.363	2.911	1.808	2.116	2.659
Eq. (14), α=1	1.606	1.77	2.461	1.358	1.516	2.209

To have a better view, The results for car noise on microphone pair 1,2 is depicted in Figure 4.

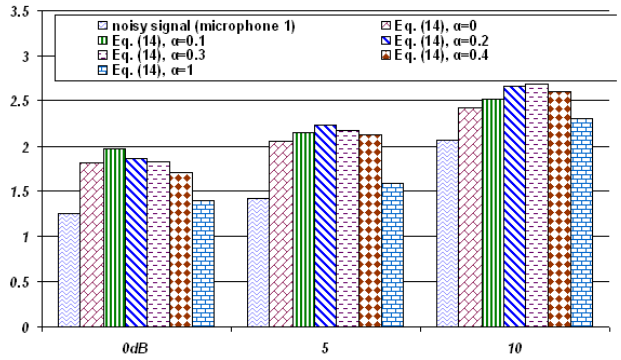

Fig 4: results for car noise, microphone pair 1,2.

The PESQ score for noisy signal on microphone 3 is higher than the score for microphone 1. This is because of the more distance between speaker and microphone 1 in comparison with the distance between speaker and microphone 3. It is known that, the performance of the coherence based methods increases when the distance between microphones increases. The PESQ scores in the table for microphone pair 3,4 (comparing with 1,2) do not confirm it. This may results from that the initial scores of noisy signals for microphone pair 1,2 is lower than 3,4.

As it is seen, in all the cases the best results are not for the original CPSS method. If one compares the results for the coherence based method and the original CPSS method, it is seen that for very noisy signals (0dB and 5dB input SNR) the results obtained by CPSS method is better than the results of coherence based method, while for high SNR signals (10dB and 20dB input SNR) the results obtained by the coherence is better than the CPSS results. It is also seen for all methods: as the input SNR increases, the methods with lower distortion and higher residual noise are more appropriate. A motivation for this effect is that for high input SNR signals high noise reduction is not much essential, but the method with lower distortion is more important.

5. Conclusion

In this paper, we have proposed a techniques to control the level of residual noise in the coherence based method. For that reason, we have adjusted the cross power spectral subtraction modification filter to control the amount of noise reduction. The enhanced signals obtained on realistic database confirm the advantage of residual noise control. It was shown that, rigid noise reduction especially when the background noise is low, decreases the recognition rate and speech quality and by controlling the level of noise reduction

less enhanced speech with low distortion can be achieved.

5. References

[1] Y. Ephraim, H. L. Van Trees, A Signal Subspace Approach for Speech Enhancement IEEE Trans. on Speech and Audio Processing, 3(4) (1995) 251-266.

[2] J. C. Benesty, J. Y Huang, S Doclo, New insights into the noise reduction Wiener filter, IEEE Trans. on Audio, Speech, And Language Processing, 14(4) (2006) 1218-1234.

[3] M. Rahmani, R. Abdipour, A. Akbari, B. Ayad, Background Noise Control for Speech Enhancement Applications, The 11th annual computer society of Iran computer conference, Tehran, Iran, 24-26 Jan 2006. pp 236-239.

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

[5] X. Huang, A. Acero, and H. W. Hon, Spoken Language Processing: A Guide to Theory, Algorithm, and System Development, Prentice Hall PTR, Upper Saddle River, N.J., 2001.

[6] Y. Ephraim, D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator", IEEE Trans. on Acoustics, Speech, Signal Processing, vol. 32 (6), Dec. 1984

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

[8] hyvarinen, E. Oja, "Independent Component Analysis: Algorithms and Applications", Neural Networks, no. 4-5, pp. 411-430, 2000.

[9] D. Van,K M. Buckley "Beamforming: A Versatile Approach to Spatial Filtering" . IEEE ASSP Magazine, pages 4--24, Apr. 1988.

[10] H. Krim and M. Viberg, "Two decades of array signal processing research", IEEE Signal Processing Magazine, vol. 13, Jul 1996.

[11] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," IEEE Trans. Antennas and Propagation, vol. 30, no. 1, pp. 27-34, 1982.

[12] D. R. Campbell, and P. W. Shields, "Speech enhancement using subband adaptive Griffiths-Jim signal processing," Speech Communication 39, pp. 97-110, 2003.

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

[14] J.B. Allen, D.A. Berkley, J. Blauert, 'Multi microphone signal processing technique to remove room reverberation from speech signals', J. Acoust. Soc. of Amer., vol. 62, no. 4, pp. 912–915, 1977.

[15] R. Le Bouquin, A. A. Azirani, G. Faucon, "Enhancement of Speech Degraded by Coherent and Incoherent Noise Using a Cross-Spectral Estimator," IEEE Trans. on Speech and Audio Processing, Sep. 1997.

[16] R. Le Bouquin, G. Faucon, "Voice Activity Detector Based on The Averaged Magnitude Squared Coherence", International Conference on Signal Processing Applications and Technology, Oct 1995.

[17] X. Zhang, Y. Jia, "A Soft Decision Based Noise Cross Power Spectral Density Estimation for Two-Microphone Speech Enhancement Systems," ICASSP 2005, Philadelphia, March 2005.

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

[19] M. Rahmani, A. Akbari, B. Ayad, "A modified Coherence based method for dual microphone speech enhancement", IEEE int. conf. on signal processing and communication, Dubai, November 2007.

[20] P. Sovka, P. Pollak, J. Kybic, "Extended Spectral Subtraction", Proceeding of European Signal Processing Conference, EUSIPCO-96, Trieste, Italia, Sep. 1996.

[21] P. Scalart, J.V. Filho, "Speech enhancement based on a priori signal to noise estimation", ICASSP1996, vol. 2, pp 629 – 632, May 1996.

[22] Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs, ITU-T Recommendation P.862, Feb 2001.