POWER LEVEL DIFFERENCE AS A CRITERION FOR SPEECH ENHANCEMENT

Nima Yousefian, Mohsen Rahmani, Ahmad Akbari
ni_yousefiyan @comp.iust.ac.ir, {m_rahmani, akbari}@iust.ac.ir
Audio and Speech Processing Laboratory, Computer Engineering Department, Iran University of Science and Technology

ABSTRACT

This paper deals with the problem of speech enhancement in near field condition, when two microphones are available. Proposed technique relies on the difference in power of received signals at the two microphones. This difference is employed to estimate the clean speech signal power. The method has the capability of dealing with non-stationary noise, a drawback of many noise reduction techniques. Superiority of the presented method over some of prominent methods in this field is demonstrated by conducting both subjective and objective quality tests.

Index Terms**—** noise suppression, non-stationary noise, power of signal

1. INTRODUCTION

Single microphone speech enhancement algorithms are favored in many applications because they are relatively easy to apply. Nevertheless, their performance is limited especially when non-stationary noise is present. The performance of noise reduction algorithms can be expected to improve, when more than one microphone is available. In this case, the spatial characteristic of the sound field can be exploited.

Beamforming is one of the simplest and most robust means of spatial filtering with several modified versions. The fundamental assumption of basic beamforming techniques is that noise components at different microphones are mutually uncorrelated. In many cases, however, the obtained noise reduction level is not sufficient and post-filtering techniques are used to further enhance the output of the beamformer. However, in a diffuse noise field, where the low-frequency noise components are coherent, the noise reduction performance degrades remarkably. A major shortcoming of many multi-channel post-filtering techniques is that highly non-stationary noise components are not dealt with. Unfortunately, transient interferences are often much too brief and abrupt for the above post-filtering methods [1].

In this paper, a novel approach for dual-channel speech enhancement systems is presented. This approach applies the Power Level Difference (PLD) between the two channels as a criterion for speech enhancement. Although this cue has been proposed and tested for sound source localization with microphone array [5], none of the previous work has investigated its applicability to the speech enhancement problem. Meanwhile, the proposed method possesses several advantages such as independence from the time delays estimation between input signals and appropriate outputs in case of presence of non-stationary noise.

2. PRELIMINARIES AND PRIOR WORK

Let us assume that two Omni-directional microphones are placed in a noisy environment to receive the desired signal. Let the signals received at the microphones after delay compensation be:

$$x_1(m) = h_1(m) * s(m) + n_1(m) \tag{1}$$

$$x_2(m) = h_2(m) * s(m) + n_2(m) \tag{2}$$

where $x_1(m)$ and $x_2(m)$ denote the signals obtained by the microphones, $h_1(m)$ and $h_2(m)$ are the impulse responses associated with the speech source for the first and second microphones, respectively; $s(m)$ is the main source signal and finally $n_1(m)$ and $n_2(m)$ are noise signals received at each microphone. By converting the above equations to frequency domain we obtain:

$$X_1(n,k) = H_1(n,k)S(n,k) + N_1(n,k) \tag{3}$$

$$X_2(n,k) = H_2(n,k)S(n,k) + N_2(n,k) \tag{4}$$

where $X_i(n,k)$ denotes the Fourier transform of the $x_i(m)$ for the frame n and the k-th frequency bin; $S(n,k)$ denotes the Fourier transform of $s(m)$ also $N_i(n,k)$ and $H_i(n,k)$ denote the Fourier transform of $n_i(m)$ and $h_i(m)$ respectively. We are now interested in describing two well known techniques for speech enchantment briefly.

One of the basic techniques in this field, known as coherence technique, employs the correlation between speech signals in the two channels for speech enhancement [1][2]. The idea behind this technique is that the speech signals in the two channels are correlated while the noise signals are uncorrelated. The coherence function used to filter the observations is defined by the following equation:

$$G_{Coh}(n,k) = \frac{P_{X1X2}(n,k)}{\sqrt{P_{X1}(n,k)P_{X2}(n,k)}} \tag{5}$$

Where P_{x1} and P_{x2} denote the Power Spectral Density (PSD) of signals x_1 and x_2 respectively and P_{x1x2} is the Cross Power Spectral Density (CPSD) between them. By estimating noise characteristics during silent intervals (noise alone), and considering its PSD in calculating coherence function, the work in [3] achieved much higher efficiency than the initial coherence algorithm in [2]. This modified coherence- based method is referred as improved coherence in the rest of the paper. Although this method is powerful in noise reduction especially when speech is degraded by stationary noise, introducing musical noise in enhanced signal is the major disadvantage of the method.

Beside the above mentioned techniques, Time Delay of Arrival (TDOA) of input signals at two microphones is another criterion for the speech enhancement problem. One of the most well-known methods which utilizes TDOA for the enhancement is phase based method [4]. The method states that in ideal situation (assuming no noise and no reverberation) the difference in the phases of the

978-1-4244-2354-5/09/$25.00 ©2009 IEEE

input signals in the two channels after delay compensation should be zero. In practical applications, the phase error will often not be zero due to noise or reverberation. In [4] authors have defined an error term as:

$$\theta(n,k) = \angle X_1(n,k) - \angle X_2(n,k) - \omega\tau \qquad (6)$$

where τ is the time delay estimation between signals received at two microphones. After defining the error term the empirical proposed filter is defined by:

$$G_{Phb}(n,k) = \frac{1}{1+\gamma\theta^2(n,k)} \qquad (7)$$

where γ is a constant coefficient. Although the phase based method has satisfactory performance in presence of non-stationary noise, it requires the accurate time delay estimation between the two signals which is often a problematical task in adverse situations. Furthermore, poor performance of phase based method when distance between two microphones is negligible is a major drawback of the method.

3. POWER LEVEL DIFFERENCE (PLD)

This section presents the proposed dual-microphone speech enhancement approach. First, PLD concepts are introduced and it is shown how PLD can be utilized as an estimator of the clean speech signal. Then, we present the proposed PLD-based algorithm for speech enhancement. In fact, the basic principle behind PLD is that in near field sensor arrays, where the distances between the source and two microphones are distinct, signals emitted from the source of interest have different power levels in different microphones, while the levels associated with noise signals are almost identical. To explain this fact, let us consider *$h_{12}(m)$* as the room impulse responses between the first and second microphone. Now, we can rewrite (3) and (4) as:

$$X_1(n,k) = S_1(n,k) + N_1(n,k) \qquad (8)$$

$$X_2(n,k) = H_{12}(n,k)S_1(n,k) + N_2(n,k) \qquad (9)$$

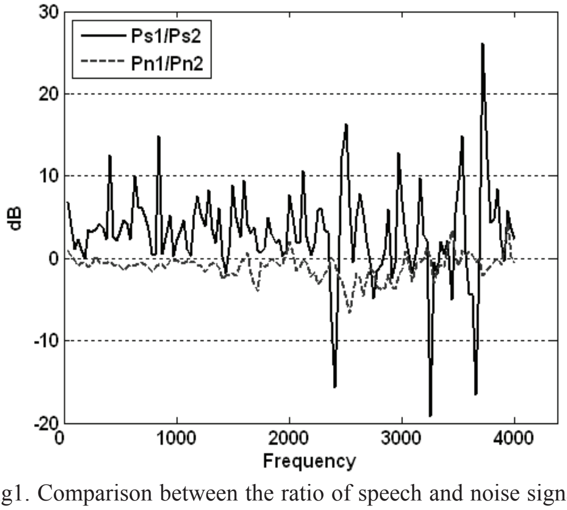

Fig1. Comparison between the ratio of speech and noise signals in the two channels in dB for one frame of speech. The distance between microphones is 62mm (both microphones installed on a headset on a dummy head. The clean speech was played from a loudspeaker installed on the mouth of the head. First microphone is closer to the source).

Where *$H_{12}(n,k)$* denotes the Fourier transform of *$h_{12}(m)$* and *$S_1(n,k)$* is equal to *$H_1(n,k)S(n,k)$* in (3). If we assume that speech and noise are independent, the following equations are derived

$$P_{X1}(n,k) = P_{S1}(n,k) + P_{N1}(n,k) \qquad (10)$$

$$P_{X2}(n,k) = |H_{12}(n,k)|^2 \; P_{S1}(n,k) + P_{N2}(n,k) \qquad (11)$$

Where prefix *P* stands for power of signal. Now, if we consider the last two equations and subtract the latter from the former, we have:

$$P_{X1}(n,k) - P_{X2}(n,k) = P_{S1}(n,k)(1-|H_{12}(n,k)|^2) + \Delta P_N(n,k) \qquad (12)$$

Where *$\Delta P_N(n,k)=P_{n1}(n,k) - P_{n2}(n,k)$*. In diffuse noise field, the difference in the power of noise signals in the two channels is negligible and we can omit term *$\Delta Pn(n,k)$,* from the last equation. This assumption is also valid for noise signals coming from far sources. Figure 1 illustrates the ratio between powers of noise signals in the two channels and compares it with that of the clean speech signals. According to the above discussion by taking absolute values we can rewrite (12) as follows:

$$|\Delta P_X(n,k)| = |(1-|H_{12}(n,k)|^2)| \; P_{S1}(n,k) \qquad (13)$$

Where *$\Delta P_X(n,k)= P_{X1}(n,k) - P_{X2}(n,k)$*. It is expected that the microphone closer to the source always receives more powerful signal than the farther one. In practical situations, the achieved value for *$\Delta P_X(n,k)$* may violate this assumption. Some effects such as reverberation in environment are the main reasons for this violation. In our experiments we found that this phenomenon usually occurs in speech pause intervals. By taking absolute values we ensure that *$\Delta P_X(n,k)$* always takes a nonnegative value. However, the last equation reveals that the difference between the powers of the input signals in the two channels is proportional to the clean speech signal power. We are interested to depict this fact by figures 2 and 3. It can be concluded from the figures that the difference between two channels follows the power of speech signal in speech intervals and lower frequencies more accurately.

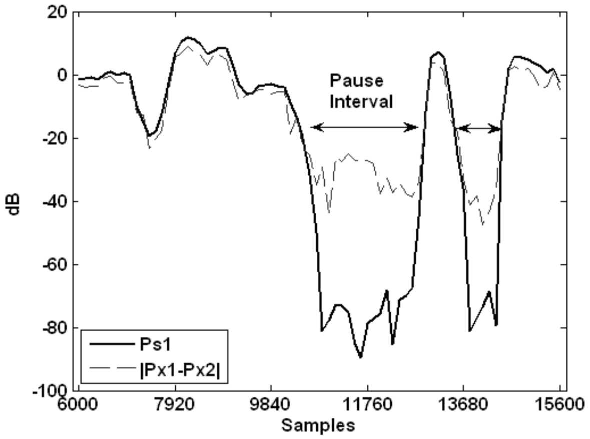

Fig2. Comparison between the power of the difference of noisy signals in the two channels and clean speech signals in dB for 9600 samples (1.2 sec) of noisy signals, frequency=300 Hz. SNR=10dB (babble noise).

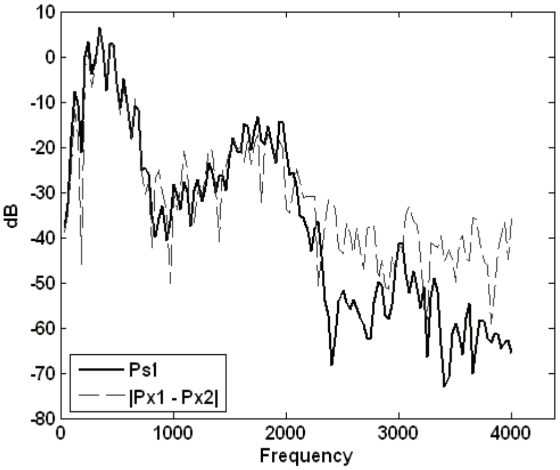

Fig3. Comparison between the power of the difference of noisy signals in the two channels and clean speech signals in dB for one frame of speech. SNR=10dB (babble noise).

In this work, the Power Spectral Density (PSD) of noisy signal is computed by the following equation:

$$P_X(n,k) = \lambda_X P_X(n-1,k) + (1-\lambda_X)\,|\,X(n,k)\,|^2 \qquad (14)$$

where λ_X is a forgetting factor.

Now, we propose a PLD-based algorithm for speech enhancement. To obtain the desired filter, we start from the Wiener filter, defined by the following equation:

$$H_W(n,k) = \frac{P_S(n,k)}{P_S(n,k) + P_N(n,k)} \qquad (15)$$

By multiplying both numerator and denominator of the last equation by $|(1-|H_{12}(n,k)|^2)|$, using (13) and naming proposed filter $G_{\Delta P}$ we have:

$$G_{\Delta P}(n,k) = \frac{|\,\Delta P_X(n,k)\,|}{|\,\Delta P_X(n,k)\,| + |\,(1-|\,H_{12}(n,k)\,|^2)\,|\,P_N(n,k)} \qquad (16)$$

As it is evident from (16), the proposed filter, requires estimating the PSD of the noise and the ratio between impulse responses associated with the speech source for the microphone pair, in each frame. First, In order to calculate the PSD of noise, we use speech pauses intervals to learn the noise characteristics. In this work, P_n is learned simply by a recursive equation from the first frames (silent interval) of the noisy signal.

$$P_N(n,k) = \lambda_N P_N(n-1,k) + (1-\lambda_N)\,|\,X(n,k)\,|^2 \qquad if\; n<T \qquad (17)$$

where λ_N is a forgetting factor and T is a simple threshold on the number of frames. Second, calculation of $|H_{12}(n,k)|$ can easily be made by utilizing (11), and assuming the independence of noise and speech signals again. We can illustrate the proposed PLD-based algorithm by a block diagram shown in figure 4. During experimental evaluation, we found that time alignment of the input signals has no major impact on the performance of the proposed algorithm. Thus, the delay compensator module is not included in the block diagram. Since the precise time delay estimation is not a straightforward operation, this issue can be considered as a remarkable advantage of the proposed method

To substantiate the proposed method, we compare the clean speech signal power estimated by the proposed algorithm with the real speech signal power in figure 5. The figure also depicts speech signal power estimated with single channel spectral subtraction method [6]. Both the proposed and power spectral subtraction methods estimate the noise power by the formula in (17).

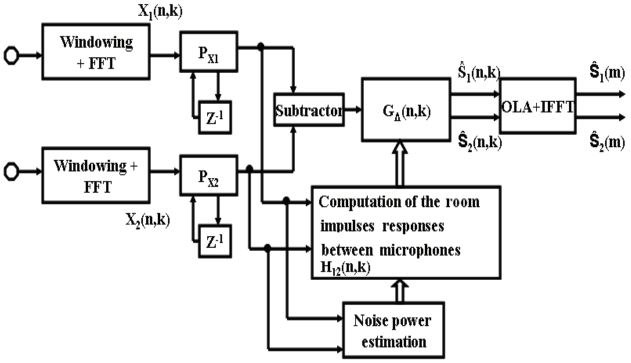

Fig4. Block diagram of proposed PLD-based method

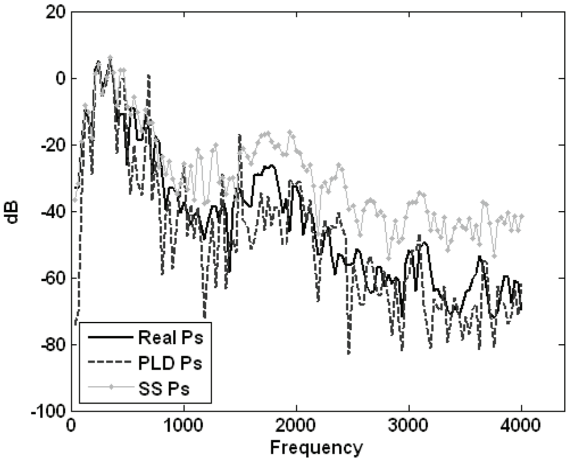

Fig5. Comparison of the clean speech signal power and power of the signal estimated by the PLD based method and spectral subtraction method in dB for one frame of speech. SNR=10dB (babble noise). The distance between microphones is 62mm and the estimates are related to the microphone closer to the source.

4. EVALUATION

To assess the performance of the method, speech files which have been recorded by a multi-microphone headset are employed. The headset has four microphones in positions that should be realistically achievable in a real headset design. The dataset used in the experiment consists of 60 different utterances, comprising six distinct speakers (three male, three female). Speech signals are recorded at a sampling rate of 8 kHz. Each frame contains 256 samples with 50% overlap between adjacent frames, is spectrally decomposed by a fast FT (FFT) with a Hanning window. We compare the performance of the proposed PLD technique with that of the two introduced algorithms in section 2: improved coherence [3] and phase based [4] methods.

For estimating the PSD of noisy signals λ_x in(14) is set to 0.7. In addition, both PLD and improved coherence methods use the equation (17) for estimating the PSD of noise with λ_n= 0.9. The T in (17) is set to 20, which means the noise power is learned from about the first 300ms of the input signal. The γ of the phase based method in (7) is set to 5. We should notice that all results reported here, are associated with the microphone closer to the source.

In order to evaluate the performance, an informal listening test was conducted. We asked 7 listeners to judge the quality of enhanced audio signals. Each listener listened to clean speech, noisy speech and enhanced signals of the three methods. Each listener gave a score between one (poor) and five (excellent) to each output. This scale corresponds to the MOS scale presented in [7]. It represents listener's general appreciation. The listening test results are presented in Table1. It can be seen from the table that the proposed technique outperforms both improved coherence and phased based techniques in most cases. Assuming that speech is degraded by non-stationary noise (babble), the results of the phase based method is higher than that of PLD at 5dB. This may be because of the noise power estimation technique, which we have applied in this work. We have estimated the noise power only from the 20 first frames of the input signal, not strictly admissible, when speech is degraded by non-stationary noise. However, this paper is not devoted to study of advanced noise estimation algorithms and in this evaluation we use the simple method stated in (17) for the estimation.

Table1. Listeners test average scores for enhanced signals

Input SNR and Noise Type	Improved Coherence	Phase Based	PLD (Proposed)
5dB (Babble)	2.5	3.58	3.29
5dB (car)	3.38	3	3.71
15dB (babble)	3.46	3.88	4
15 dB (car)	3.66	3.65	4.32

We also assess the performance of different methods by an objective measure: Perceptual Evaluation of Quality (PESQ) [8]. PESQ prediction maps mean opinion score (MOS) estimates to a range between -0.5 and 4.5, where 1.0 corresponds to non-acceptable speech signal and 4.5 corresponds to a distortion-less one. Figure 6 add 7 show the PESQ scores for the noisy signal and signals enhanced by the mentioned techniques, where speech is degraded by car and babble noise respectively.

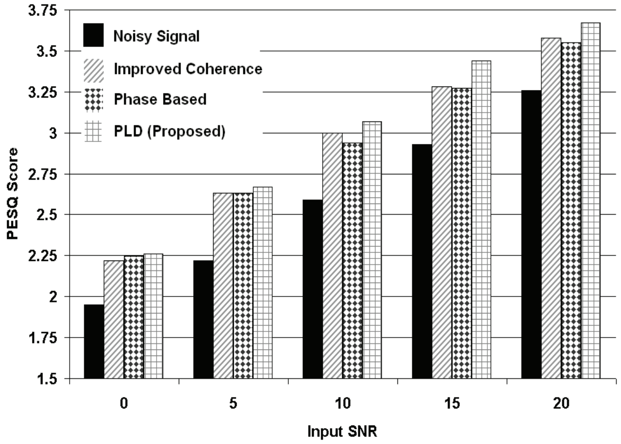

Fig6. PESQ scores for the noisy signal and signals enhanced by the different methods. Speech is degraded with car noise

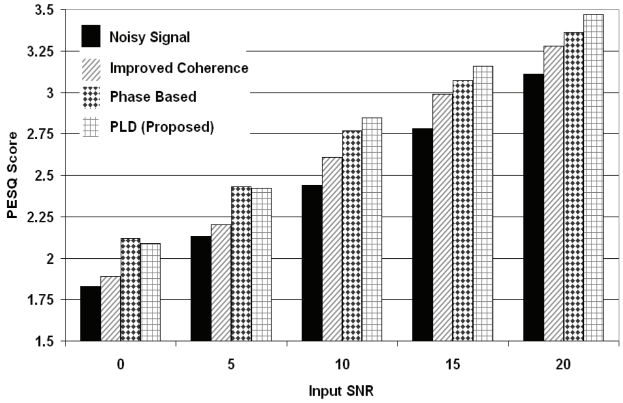

Fig7. PESQ scores for the noisy signal and signals enhanced by the different methods. Speech is degraded with babble noise

5. CONCLUSION

This study introduced a new approach for dual-channel speech enhancement in near field condition, based on power level difference (PLD). We first showed how the PLD approach can be utilized as a criterion for the speech enhancement problem, and then proposed a Wiener based filter applying PLD. Employing various quality measures indicates the superiority of the method over some of the eminent techniques in this field.

REFERENCES

[1] I. Cohen, "Multi-Channel Post-Filtering in Non-Stationary Noise Environments" IEEE Trans. Acoustics, Speech, and Signal Processing, Vol. 52, PP. 1149-1160, May 2004.

[2] R. Le Bouquin, G. Faucon, "Using the coherence function for noise reduction," in Proc. IEE on Communications, Speech and Vision. Vol. 139, pp.276-280, June 1992.

[3] R.L. Bouquin-Jeanes, A.A. Azirani, G. Faucon. "Enhancement of speech degraded by coherent and incoherent noise using a cross-spectral estimator," IEEE Trans. Speech and Audio Processing, vol. 5, pp. 484–487, September 1997.

[4] P. Aarabi, G. Shi, "Phase-based dual-microphone robust speech enhancement", IEEE Trans. Systems, Man and Cybernetics, vol. 34, pp. 1763-1773, August 2004.

[5] S.T. Birchfield, R. Gangishetty, "Acoustic localization by interaural level difference", in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing., ICASSP-05, vol. 4, pp. 1109-1112, 2005.

[6] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction" IEEE Trans. Acoustics, Speech, and Signal Processing, vol. 27, no. 2, pp. 113-120, 1979.

[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] ITU-T Recommendation, Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech coders, ITU-T Recommendation P.862, February 2001.