DUAL-MICROPHONE NOISE REDUCTION FOR MOBILE PHONE APPLICATION

Zhong-Hua Fu

School of Computer Science
Northwestern Polytechnical University
Xi'an, China

Fan Fan, Jia-Dian Huang

ZTE Corporation
China

ABSTRACT

This paper addresses the dual-microphone noise reduction problem in mobile phone application. We propose to use the inter-microphone Posteriori SNR Difference (PSNRD) for Speech Presence Probability (SPP) estimation, which is more robust than the Power Level Difference (PLD) that is often used previously. Additionally, we use the recently reported multichannel minimum variance distortionless response (MVDR) filter for noise reduction. We divide the noises into quasi-stationary and transient components, and propose an SPP-based noise correlation matrix estimator. Analysis and experiments on data recorded in real environments verify the robustness of the PSNRD. The SPP-based estimator is more appropriate for the MVDR filter in tracking transient noise, and the proposed MVDR filter can lead to high noise reduction and small speech distortion.

Index Terms— Dual-microphone, noise reduction, MVDR filter, speech presence probability

1. INTRODUCTION

The fast developments of mobile techniques provide us many kinds of mobile devices. Mobile phone is nowadays used almost everywhere for communication around the world. Since the acoustic environments in which people are talking are quite different, it is necessary to use speech enhancement technology to provide high quality speech signals.

It is known that single channel speech enhancement is not capable of dealing with highly non-stationary noise in outdoor environments. To get better speech quality, many mobile phone manufacturers begin to deploy two or three microphones in their products. In a typical dual-microphone configuration, the front bottom microphone is a primary one, while the rear top microphone is a secondary one. Note that the dual-microphone noise reduction for mobile phone is different from general dual-channel speech enhancement, e.g. in hearing-aids [1] or speech enhancement [2] . It generally assumes that the primary microphone receives stronger speech than the secondary one, and due to the shadow of user's head, hand and the phone body, the secondary microphone receives approximate or stronger noise than the primary one. So the inter-microphone Power Level Difference (PLD) is often exploited for voice activity detection (VAD) and noise estimation, such as in [3]. However, the difference between the sensitivities of the dual microphones influences the PLD greatly, which makes it difficult to decide the corresponding thresholds.

Another problem is that the secondary signal is only used in VAD and noise estimation but not in the enhancement procedure. Recently, a series of researches on the multichannel MVDR filter based on a new decomposition theory is examined. They verified that the new MVDR filter can provide very promising results compared to Wiener filter [4, 5, 6]. However, they all focus on the theoretical works and assume the noise is estimated perfectly. This, in fact, is a very challenge problem and directly influences the performance of the optimal filter.

In this paper we firstly propose to use the Posteriori SNR Difference (PSNRD) of inter-microphone for Speech Presence Probability (SPP) estimation. We will show that it is more reliable to assume that the posteriori SNR of the primary microphone is higher than that of the secondary one. Secondly, we introduce an SPP-based noise correlation matrix estimator and involve the multichannel MVDR filter. With data recorded in different environments using two dual-microphone deployments, the experimental results show that the proposed MVDR filter can provide high noise reduction and small speech distortion.

2. PSNRD AND SPP ESTIMATION

The noises in mobile applications are quite complicated. We divide them into quasi-stationary and transient components. The former has a spectrum that changes slowly, such as engine noise and sensor noise. It can be estimated using many reported single channel estimator, such as in [7, 8]. The latter consists of interferences that burst suddenly and unexpectedly, which is very difficult to suppress in single microphone situation. By transforming into time-frequency domain, the dual-microphone signal model can be expressed as

$$y_m(k,n) = x_m(k,n) + v_m(k,n) + u_m(k,n) = x_m(k,n) + o_m(k,n), (m = 1,2), \quad (1)$$

where $y_m(k,n)$, $x_m(k,n)$, and $o_m(k,n)$ are the spectral coefficients of the received signals, the desired signals, the undesired environment sound signals, respectively. $v_n(k,m)$, and $u_m(k,n)$ are the quasi-stationary noises and the transient interferences. k is subband index and n is time-frame index. m is the microphone index. $m = 1$ refers to the primary microphone and $m = 2$ is the secondary microphone. We assume that $x_m(k,n)$, $v_m(k,n)$, and $u_m(k,n)$ are uncorrelated to each other.

The normalized PLD of inter-microphone is simply defined as

$$\Delta\phi(k,n) = \frac{\phi_{y_1}(k,n) - \phi_{y_2}(k,n)}{\phi_{y_1}(k,n) + \phi_{y_2}(k,n)}, \quad (2)$$

where $\phi_{y_m}(k,n) = E\left[|y_m(k,n)|^2\right]$ is the variance of $y_m(k,n)$. The PLD is often used for VAD in mobile phone applications. The reason is that when a user holds his/her dual-microphone phone and talks, his/her mouth is close to the primary microphone, and his/her head, hand, and the phone body will disturb the sound propagation to the secondary microphone. Therefore the speech picked up by the primary microphone is louder than the secondary microphone. The additive noises are different. For example, in diffuse noise field, the noises picked up by both microphones are approximately the same. While in free space, if some other speaker is talking on the opposite side to the target speaker, the secondary microphone will pick up stronger interferences than the primary one.

However, the above observations are based on the assumption that the sensitivities of the two microphones are same. Apparently, in (2), if we change the gain of each microphone, i.e. the sensitivity, the PLD will change as well. Hence it is difficult to determine the correct thresholds for VAD.

We propose another VAD flag, the posteriori SNR difference (PSNRD), which is defined as

$$\Delta\gamma(k,n) \triangleq \frac{\gamma_1(k,n) - \gamma_2(k,n)}{\gamma_1(k,n) + \gamma_2(k,n)}, \quad (3)$$

where $\gamma_m(k,n)$ is the posteriori signal to quasi-stationary noise ratio (PSNR), i.e.

$$\gamma_m(k,n) \triangleq \frac{\phi_{y_m}(k,n)}{\phi_{v_m}(k,n)}. \quad (4)$$

Note that the denominator of (4) is $\phi_{v_m}(k,n)$, i.e. the variance of the quasi-stationary noise. The basic assumption then, is that the SNR of the primary microphone is always higher than that of the secondary microphone. In case that the primary microphone has higher sensitivity, this assumption still holds since all signals in the primary microphone are amplified by same gain. Therefore the PSNRD is independent of the microphones' sensitivity.

Then the SPP estimator is straightforward using a simply linear mapping from the PSNRD, as

$$p(k,n) = \begin{cases} 1 & \Delta\gamma(k,n) > \Delta\gamma_{\max} \\ \frac{\Delta\gamma(k,n) - \Delta\gamma_{\min}}{\Delta\gamma_{\max} - \Delta\gamma_{\min}} & \text{else} \\ 0 & \Delta\gamma(k,n) < \Delta\gamma_{\min} \end{cases} \quad (5)$$

3. DUAL-CHANNEL NOISE REDUCTION

The reported noise reduction approaches with dual-microphone phone, such as in [3], aim at finding correct spectral gains, and applying them on the signals of the primary microphone to enhance the desired speech. The inter-microphone correlation is used but with strict limitation. Recently, a multi-channel MVDR filter considering both the inter-channel and inter-frame correlations is proposed [6], which improves the full-band SNR with small speech distortion. However, in spite of the very promising performance, the correlated works on the MVDR filter [4, 5] all assume the noise correlation matrix is perfectly estimated, which in fact, influences the MVDR performance greatly. In this section, we firstly introduce the multi-channel MVDR filter into our dual-microphone noise reduction problem, then provide an SPP-based estimator for correlation matrix estimation. The detail explanation about the multi-channel optimal noise reduction filter in STFT domain refers to [9].

Firstly, we rewrite the signal model in (1) into array notation:

$$\underline{\mathbf{y}}(k,n) = \underline{\mathbf{x}}(k,n) + \underline{\mathbf{v}}(k,n) + \underline{\mathbf{u}}(k,n) = \underline{\mathbf{x}}(k,n) + \underline{\mathbf{o}}(k,n), \quad (6)$$

where

$$\mathbf{y}(k,n) = [y_1(k,n), y_2(k,n)]^T$$
$$\underline{\mathbf{y}}(k,n) = [\mathbf{y}(k,n)^T, \mathbf{y}(k,n-1)^T, \cdots, \mathbf{y}(k,n-L+1)^T]^T,$$

and L is the number of consecutive time-frames used for each subband. The superscript ^{T} denotes transpose operator. $\underline{\mathbf{x}}$, $\underline{\mathbf{v}}$, $\underline{\mathbf{u}}$, and $\underline{\mathbf{o}}$ are defined in a similar way. Then the dual-microphone noise reduction can be simply expressed as

$$\hat{x}_1(k,n) = \underline{\mathbf{h}}^H(k,n)\,\underline{\mathbf{y}}(k,n) = \underline{\mathbf{h}}^H(k,n)\,\underline{\mathbf{x}}(k,n) + \underline{\mathbf{h}}^H(k,n)\,\underline{\mathbf{o}}(k,n) = x_{1\mathrm{f}}(k,n) + o_{\mathrm{rn}}(k,n), \quad (7)$$

where the superscript ^{H} denotes transpose-conjugate operator. We take the clean speech signal picked by the primary microphone, i.e. $x_1(k,n)$, as the desired signal. By decomposing the vector $\mathbf{x}(k,n)$ into two orthogonal components

depending on the correlation with the desired $x_1(k,n)$, (7) is rewritten as [6]

$$\begin{aligned}\hat{x}_1(k,n) &= x_1(k,n)\,\underline{\mathbf{h}}^H(k,n)\,\underline{\mathbf{d}}^*_{\underline{\mathbf{x}}}(k,n) + \\ &\quad \underline{\mathbf{h}}^H(k,n)\,\underline{\mathbf{x}}_{\mathrm{i}}(k,n) + \underline{\mathbf{h}}^H(k,n)\,\underline{\mathbf{o}}(k,n) \\ &= x_{\mathrm{fd}}(k,n) + x_{\mathrm{ri}}(k,n) + o_{\mathrm{rn}}(k,n), \end{aligned} \quad (8)$$

where

$$E\left[x_1(k,n)\,\underline{\mathbf{x}}^*{}_{\mathrm{i}}(k,n)\right] = \underline{\mathbf{0}}, \quad (9)$$

and

$$\underline{\mathbf{d}}_{\underline{\mathbf{x}}}(k,n) = \frac{E\left[x_1(k,n)\,\underline{\mathbf{x}}^*(k,n)\right]}{E\left[\left|x_1(k,n)\right|^2\right]} = \frac{\Phi_{\underline{\mathbf{x}}}(k,n)\,\mathbf{i_0}}{\phi_{x1}(k,n)}, \quad (10)$$

where $\Phi_{\underline{\mathbf{x}}}(k,n) = E\left[\underline{\mathbf{x}}(k,n)\,\underline{\mathbf{x}}^H(k,n)\right]$ is the correlation matrix of $\underline{\mathbf{x}}(k,n)$. $\mathbf{i_0}$ is the first column of the identity matrix.

Now using MSE criterion with distortionless constraint, one can derive the optimal MVDR filter as

$$\mathbf{h}_{\mathrm{MVDR}}(k,n) = \frac{\Phi^{-1}_{\underline{\mathbf{y}}}(k,n)\,\underline{\mathbf{d}}^*_{\underline{\mathbf{x}}}(k,n)}{\underline{\mathbf{d}}^T_{\underline{\mathbf{x}}}(k,n)\,\Phi^{-1}_{\underline{\mathbf{y}}}(k,n)\,\underline{\mathbf{d}}^*_{\underline{\mathbf{x}}}(k,n)}. \quad (11)$$

$\Phi_{\underline{\mathbf{y}}}(k,n)$ is the correlation matrix of $\underline{\mathbf{y}}(k,n)$. The key merit of the MVDR filter is that no distortion will occur in the desired signal, and the residual noise is more pleasant compared to Wiener filter. Although in theory, the MVDR filter can provide maximum output SNR as Wiener filter; in practice, its noise reduction performance is modest compared to Wiener.

Now, the essential problem is to estimate the correlation matrix, $\Phi_{\underline{\mathbf{y}}}(k,n)$ and $\underline{\mathbf{d}}_{\underline{\mathbf{x}}}(k,n)$. $\Phi_{\underline{\mathbf{y}}}(k,n)$ is simply estimated using regressive smoothing [4],

$$\Phi_{\underline{\mathbf{y}}}(k,n) = a_y\Phi_{\underline{\mathbf{y}}}(k,n-1) + (1-a_y)\left[\underline{\mathbf{y}}(k,n)\underline{\mathbf{y}}^H(k,n)\right], \quad (12)$$

and $\underline{\mathbf{d}}_{\underline{\mathbf{x}}}(k,n)$ is obtained using

$$\underline{\mathbf{d}}_{\underline{\mathbf{x}}}(k,n) = \frac{\Phi_{\underline{\mathbf{y}}}(k,n)\,\mathbf{i_0}}{\phi_{y_1} - \phi_{o_1}} - \frac{\Phi_{\underline{\mathbf{o}}}(k,n)\,\mathbf{i_0}}{\phi_{y_1} - \phi_{o_1}}, \quad (13)$$

where $\Phi_{\underline{\mathbf{o}}}(k,n)$ is the correlation matrix of the environment noise. Since the outdoor noise is very complicated, it is difficult to estimate $\Phi_{\underline{\mathbf{o}}}(k,n)$ directly. Considering that we divide the noise into quasi-stationary and transient components, the correlation of the former, i.e. $\Phi_{\underline{\mathbf{v}}}(k,n)$, can be estimated using similar regressive smoothing as (12).

To estimate the whole noise correlation matrix, we use an SPP-based estimator as

$$\Phi_{\underline{\mathbf{o}}}(k,n) = p(k,n)\Phi_{\underline{\mathbf{v}}}(k,n) + \left[1-p(k,n)\right]\Phi_{\underline{\mathbf{y}}}(k,n). \quad (14)$$

If the SPP is close to 1, the noise correlation matrix is decayed to the stationary noise estimate to reduce speech distortion, and if the SPP is close to 0, the noise estimates are roughly equal to the observed noisy estimates. Therefore, if the SPP is correct, this estimator will track the transient noise quickly during speech absence. The following experimental results will verify its performance.

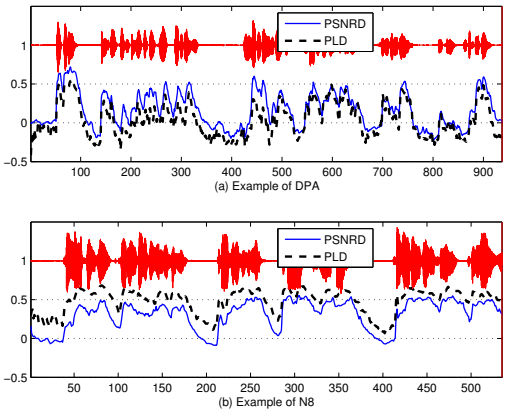

Fig. 1. Comparisons between PLDs and PSNRDs on the 468.74 Hz frequency bin

4. EXPERIMENTS AND EVALUATIONS

This section presents the experiments and performance evaluations. Two kinds of dual-microphone deployments are considered. One is a real mobile phone, i.e. Nokia N8, which is embedded with two microphones. The other is a normal mobile phone sticked with a pair of small DPA4060 microphones. One is on the front-bottom, the other is on the rear-top.

All experiment data are recorded in real environments. The clean speech signals are recorded from two male talkers in two quiet office rooms using DPAs and N8 respectively. The noise signals are recorded in a large crowded cafeteria using DPAs, and in a running bus using N8, respectively. All recordings are sampled at 8 kHz. 0 dB, 5 dB, and 10 dB SNR noisy signals are mixed in computer. We use Hanning window of 256 samples (32 ms) for STFT, with half overlap. The minimum statistics (MS) approach [7] is used to estimate the stationary noise in each channel.

We first examine the robustness of the PSNRD. Fig.1 shows an comparison. Note that the blue solid line refers to the PSNRDs and the green dotted line refers to the PLDs. The upper figure shows a DPA example with 10 dB cafeteria noise, and the lower figure shows a N8 example with 10 dB bus noise. The clean speech waveforms are also plotted to show speech activities. One can find that during noise only periods, the PLDs in the DPA case are almost negative, but in the N8 case, they are about bigger than 0.2. Hence it is difficult to select a correct threshold for PLD-based VAD. On the contrary, the PSNRDs are quite robust. They are near 0 during noise only periods in both cases.

Then we examine the noise reduction performance. All regression smoothing factors for estimating the correlation matrixes are same, 0.85 in our experiments. According to [10], same smoothing factors for both mixed signal and noise signal lead to best performance of the MVDR. The thresholds of the PSNRD for both dual-microphone deployments

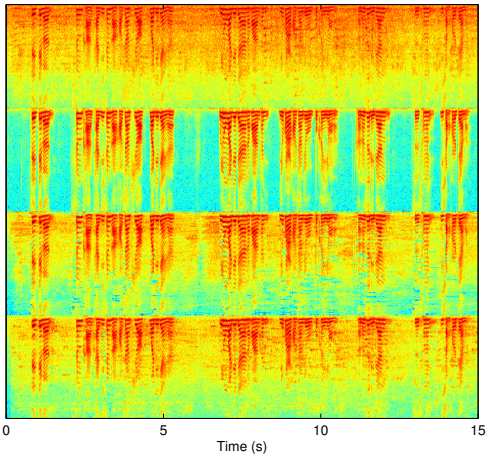

Fig. 2. From above to bottom: noisy signal, clean signal, classical MVDR, and proposed MVDR. The 10dB DPA with the cafeteria noise is used.

Table 1. Speech distortion indices (dB)

	Bus noise			Cafeteria noise		
input SNR	10	5	0	10	5	0
class. MVDR	-14.4	-12.8	-10.4	-11.1	-8.7	-5.7
prop. MVDR	-12.9	-11.3	-8.5	-9.4	-6.6	-2.8

are same, i.e. $\Delta\gamma_{\max} = 0.2$ and $\Delta\gamma_{\min} = 0$. The first 10 frames are used for initializing the correlation matrixes. We set $L = 4$ for consideration of inter-frame correlation. To verify the effect of the proposed noise estimator, we use the classical MVDR with the same parameters, where the noise correlation matrix is estimated using regressive smoothing directly from the noise signals [4].

Fig.2 shows an example, where the 10 dB DPA case with the cafeteria noise is used. We can find the proposed MVDR performs slightly better than the classical MVDR in noise only period. Hence it is not appropriate to use constant smoothing in estimation the noise correlation matrix, especially when the noise suddenly changes.

To evaluate the quantitative performance, we use the full-band speech distortion indices and the full-band array gains to evaluate the speech distortion and the SNR improvement, respectively, see [4] for details. The lower the distortion index, the smaller distortion in the desired speech signals. The higher the array gain, the more noise reduction. The quantitative results of the speech distortion indices and the array gains are list in Table 1 and Table 2 respectively. Note that the speech distortion indices are averaged during the speech active parts.

We can see that the speech distortion of the proposed MVDR are slightly higher than the classical MVDR, while the array gain of the former is better. Additionally, the performances in the bus noise are generally better than that in the cafeteria noise, since the crowded cafeteria noise is more challenge. The lower distortion of the classical MVDR is due to the directly utilization of the noise signals. But the constant smoothing influences its tracking capability of transient noise, and the SPP-based smoothing is more appropriate. Note that the distortion indices are somehow not very low considering the MVDR criteria, even using the noise directly as in the classical MVDR. That might because the MVDR theory is based on stationary signal, and to estimate the correlation matrix of non-stationary signals like speech is always an approximation. However, the listening experiences are quite pleasant and natural, without artificial effects as the Wiener filter does.

Table 2. Array gains (dB)

	Bus noise			Cafeteria noise		
input SNR	10	5	0	10	5	0
class. MVDR	8.1	10.4	11.3	5.8	7.6	8.9
prop. MVDR	16.0	18.2	16.1	7.9	8.3	9.8

5. CONCLUSIONS

This paper addresses the dual-microphone noise reduction problem in mobile phone application. We propose to use the posteriori SNR difference for speech presence probability estimation, which is robust and independent of the microphone's sensitivity. Then we use the dual-channel MVDR filter with the newly reported decomposition theory and consider both the inter-channel and inter-microphone correlations, where an SPP-based correlation matrix estimator is introduced. The experimental results verify the robustness of the PSNRD, and the SPP-based estimator is more appropriate for MVDR filter in tracking transient noise. The proposed MVDR filter can lead to high noise reduction with small speech distortion.

6. RELATION TO PRIOR WORK

The work in this paper has focused on the dual-microphone noise reduction problem in mobile phone application. The work by Jeub, et al. [3] exploits the inter-microphone PLD for noise PSD estimation. We proposed using the posteriori SNR differences, which is more robust. For noise reduction, we involved the MVDR filter with the newly reported decomposition theory [4, 5]. The correlated MVDR researches all suppose that the noise estimates are known. We proposes a practical work and proposed an SPP-based correlation matrix estimator.

7. ACKNOWLEDGEMENT

This work was supported by the National Natural Science Foundation of China (60901077) and 2012 NWPU Fundamental Research Foundation.

8. REFERENCES

[1] D. Marquardt, V. Hohmann, and S. Doclo, "Binaural cue preservation for hearing aids using multi-channel wiener filter with instantaneous ITF preservation," in *IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, 2012, pp. 21–24.

[2] N. Yousefian and P. C. Loizou, "A dual-microphone speech enhancement algorithm based on the coherence function," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 599–609, 2012.

[3] M. Jeub, C. Herglotz, C. Nelke, C. Beaugeant, and P. Vary, "Noise reduction for dual-microphone mobile phones exploiting power level differences," in *IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, 2012, pp. 1693–1696.

[4] J. Benesty and Y. Huang, "A single-channel noise reduction MVDR filter," in *IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, 2011, pp. 273–276.

[5] J. Benesty, J. Chen, Y. Huang, and T. Gaensler, "Time-domain noise reduction based on an orthogonal decomposition for desired signal extraction," *Journal of the Acoustical Society of America*, vol. 132, no. 1, pp. 452–446, 2012.

[6] E. A. P. Habets, J. Benesty, and J. Chen, "Multi-microphone noise reduction using interchannel and interframe correlations," in *IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, 2012, pp. 305–308.

[7] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech Audio Processing*, vol. 9, no. 5, pp. 504–512, 2001.

[8] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise psd tracking with low complexity," in *IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, 2010, pp. 4266–4269.

[9] J. Benesty, J. Chen, and E. A. P. Habets, *Speech Enhancement in the STFT Domain*. Berlin: Springer-Verlag, 2011.

[10] Y. Huang and J. Benesty, "A multi-frame approach to the frequency-domain single-channel noise reduction problem," *IEEE Transactions on Speech Audio Processing*, vol. 20, no. 4, pp. 1256–1269, 2012.