A Novel Speech Enhancement System Based on the Coherence-Based Algorithm and the Differential Beamforming

Lei WANG†, Member and Jie ZHU†a), Nonmember

SUMMARY This letter proposes a novel speech enhancement system based on the ‘L’ shaped triple-microphone. The modified coherence-based algorithm and the first-order differential beamforming are combined to filter the spatial distributed noise. The experimental results reveal that the proposed algorithm achieves significant performance in spatial filtering under different noise scenarios.

key words: speech enhancement, coherence-based algorithm, differential beamforming, spatial filtering

1. Introduction

Multi-channel speech enhancement techniques can obtain better performance than single-channel techniques by utilizing the spatial diversity of speech and noise signals, especially under situation of non-stationary directional noise (e.g. the competing talker) [1]. As a result, multiple microphones are increasingly embedded in the man-machine interactive system like intelligent speakers, speech recognition devices, hearing aid and cochlear implant devices [1], [2].

Spectral postfiltering and beamforming are two popular classes of multi-channel speech enhancement techniques. The coherence-based algorithm, a spectral postfiltering method, was proposed to evaluate the presence of speech and noise. Later, Nima Yousefian et al. ameliorated it to suppress the directional noise in an endfire dual-microphone configuration [3], [4]. On the premise of coherent noise field, the real-value spectral gain of every frequency bin is derived from the real and imaginary parts of the coherence function.

Beamforming realizes noise elimination by means of spatial filtering and interference rejection. To deal with speech signals, many response-invariant broadband beamformers are developed. Harmonically nested subarrays design equally-spaced linear subarrays for certain frequency bands but require a large number of microphones [5]. Differential microphone arrays (DMA), a form of filter-and-sum beamformers, provide a facile implementation for broadband signal processing by narrowband decomposition [6].

In this paper, we find the algorithm proposed in [4] obtains better property when applied on the broadside dual-microphone. However, the broadside configuration introduces a front-back transparency. We construct a two-sensor DMA with a cardioid beampattern (one null at 180°). Hence we propose a triple-microphone speech enhancement system adopting joint algorithms. The coherence-based algorithm and the differential beamforming cooperate to denoise the spatial interference. Three metrics (PESQ, STOI, fwSNR) are calculated and indicate the proposed algorithm is more effective than the coherence-based algorithm, the ordinary differential beamformer and the Generalized Side-lobe Canceller (GSC) which is designed for the directional noise.

2. System Overview

The proposed speech enhancement system, as is shown in Fig. 1, consists of three omnidirectional microphones following the ‘L’ shape, which are numbered M1, M2 and M3. Both M2 and M3 are at a distance of d from M1. Compared to the incident direction of the target speech, M1 and M2 are placed at the broadside direction and M1 and M3 are at the endfire direction. The target speaker is placed at 0° and the interference is placed at θ°. The distance among sound sources and microphones satisfies the far-field hypothesis.

2.1 Coherence-Based Algorithm of Broadside Array

In [4], the author proposed a coherence-based speech enhancement algorithm applied on the endfire dual-microphone. In this paper, we analogize the algorithm to the broadside dual-microphone and prove that its effectiveness on noise reduction is better than the previous method.

Fig. 1 Schematic diagram of the proposed speech enhancement system.
2.1.1 Description of the Coherence-Based Algorithm

Noise fields can be divided into the diffuse noise field and the coherent noise field. In this paper, we focus on the speech enhancement in the coherent noise field generated by directional sound sources, which is the dominate form in a low-reverberation condition. In a coherent noise field, based on the broadside dual-microphone configuration, the coherence function of two input signals is modeled by [7],

$$
\Gamma_u = e^{j \omega \tau \sin \theta}
$$

(1)

where \( \tau = df_s / c, d \) is the microphone spacing, \( f_s \) denotes the sampling frequency, \( c \) is the speed of sound and \( \theta \) is the incident angle. We assume a homogeneous speech and noise field, i.e. \( \Phi_n \) and \( \Phi_s \) are the power spectral density (PSD) of speech and noise signals, respectively. Hence the coherence function of two signals can be expressed as,

$$
\Gamma_u = \frac{\Phi_{n1n2}}{\sqrt{\Phi_{n1n1}\Phi_{n2n2}}} = \frac{\Phi_s}{\Phi_u}
$$

(2)

where \( \Phi_{n1n2} \) denotes the cross-power spectral density (CSD).

In case of small microphone spacing \( d \), the signal-to-noise rate (SNR) at two receivers can be considered as the same value \( \text{SNR} \). From [4], the coherence function of two received signals can be modeled as:

$$
\hat{\Gamma}_s = \frac{\text{SNR}}{1 + \text{SNR}} \Gamma_s + \frac{1}{1 + \text{SNR}} \Gamma_n
$$

(3)

where \( \Gamma_s \) and \( \Gamma_n \) are the coherence functions of clean speech signals and noisy signals. From Eq. (2), \( \Gamma_n = \Phi_n / \Phi_u \). In [4], the authors discuss the algorithm in condition of one-noise source. Now we extend it to multi-noise condition. Assuming there are \( K \) noise sources incident from \( \theta_1, \theta_2, \ldots, \theta_K \) azimuths and speech and noise signals are uncorrelated, the following equations hold: \( \Phi_n = \sum_{i=1}^{K} \Phi_{n_i} \) and \( \Phi_s = \sum_{i=1}^{K} \Phi_{s_i} \), where \( \Phi_n \) and \( \Phi_s \) are PSD and CSD of the \( i \)th noise source.

Therefore, Eq. (3) can be rewritten as,

$$
\hat{\Gamma}_s = \frac{\text{SNR}}{1 + \text{SNR}} \Gamma_s + \frac{1}{1 + \text{SNR}} \sum_{i=1}^{K} r_i \Gamma_{n_i}
$$

(4)

where \( \Gamma_{n_i} = \Phi_{n_i} / \Phi_u \) is the coherent function of the \( i \)th noise source and \( r_i = \Phi_{s_i} / \sum_{i=1}^{K} \Phi_{n_i} \). Obviously, \( \sum_{i=1}^{K} r_i = 1 \). According to Eq. (1), the following equation is obtained,

$$
\sum_{i=1}^{K} r_i \Gamma_{n_i} = \sum_{i=1}^{K} r_i e^{j \omega \tau \sin \theta_i} = e^{j \mu}
$$

(5)

where \( \mu \) is a substitution parameter. Plugging Eqs. (1) and (5) into Eq. (4), the following equation is obtained:

$$
\hat{\Gamma}_s = \frac{\cos(\omega \tau \sin 0) + j \sin(\omega \tau \sin 0)}{1 + \text{SNR}} \frac{\text{SNR}}{1 + \text{SNR}}
$$

$$
+ \frac{\cos(\mu) + j \sin(\mu)}{1 + \text{SNR}} \frac{1}{1 + \text{SNR}}
$$

(6)

The real component of Eq. (6) is,

$$
\Re = \frac{\cos \mu + \text{SNR}}{1 + \text{SNR}}
$$

(7)

From Eq. (7), \( \text{SNR} \) can be expressed as,

$$
\text{SNR} = \frac{\cos \mu - \Re}{\Re - 1}
$$

(8)

Similarly, deriving from the imaginary part \( \Im \) of Eq. (6),

$$
\Im = \frac{\sin \mu - \Im}{\Im}
$$

(9)

AppARENTLY, the right sides of Eqs. (8) and (9) are equivalent. After rearranging the equations, Eq. (10) is obtained:

$$
\Re - \Im \sin \mu = 0
$$

(10)

Since \( \Im \) is known, only need to work out the value of \( \sin \mu \) to obtain the \( \text{SNR} \) in Eq. (9). Known \( \sin^2 \mu + \cos^2 \mu = 1 \), a quadratic equation is obtained as:

$$
[\Im^2 + (1 - \Re)^2] \sin^2 \mu + 2(\Re - 1) \Re \cdot \sin \mu = 0
$$

(11)

Two solutions: \( \sin \mu = 0 \) and \( \sin \mu = 2(1 - \Re) \Im / (\Im^2 + (1 - \Re)^2) \) are available. The former leads to \( \text{SNR} = -1 \), which is physically meaningless. So the latter solution is fed into Eq. (9):

$$
\text{SNR} = \frac{1 - \Re^2 - \Im^2}{\Im^2 + (1 - \Re)^2}
$$

(12)

A square-root-Wiener filter \( G(\omega) = \sqrt{\text{SNR}(\omega)/(\text{SNR}(\omega) + 1)} \) is derived to obtain the enhanced signal.

2.1.2 Analysis on the SNR Estimator

From the above, the proposed SNR estimator is derived from the coherent function between observed signals. Compared to those methods based on the voice activity detection (VAD), the proposed method relies on no noise detection, which makes it effective for the nonstationary noise.

In [4], the authors proposed a SNR estimator as,

$$
\text{SNR} = \frac{\sin \mu - \Im}{\Re - \sin \omega \tau}
$$

(13)

In contrast to Eq. (9), the additional term = \( \sin \omega \tau \) in the denominator of Eq. (13) depends on \( \tau \) (theoretically \( \tau = df_s / c \)). But the real time delay is susceptible to the multipath propagation effect (reverberation) in real environments [8].

By contrast, our derivative procedure and the result are simplified to some extent. Besides, the estimated SNR in Eq. (12) is derived only from the real and imaginary parts of the coherent function, without involving the time delay.

To evaluate the accuracy of the algorithm, we use the two estimators to calculate SNR of an identical speech scenario. Like the configuration in Fig. 1, a target speaker is
placed at 0°. And a ‘factory’ noise is placed at 90°. The SNR estimator in [4] is called the endfire method and ours is called the broadside method. Figure 2 shows the comparison between the true SNR and the estimated SNR computed by two methods. It can be find that the SNR estimated by the broadside method is closer to the true value.

Note that when \( \theta = 180° \), \( \sin \theta = 0 \), \( \Gamma = 1 \), which is identical to the condition of \( \theta = 0° \). Hence the speech located at 0° and the noise located at 180° are indistinguishable.

### 2.2 First-Order Differential Microphone Array (DMA)

The beampattern describes the directional sensitivity of microphone beamformers to a signal impinging on arrays from \( \theta \) azimuth [6]. Uniform linear DMAs provide a convenient way to design the frequency-invariant beampatterns by applying particular constraints such as null location and maximum gain. To overcome the front-back transparency, M1 and M3 form a first-order DMA with a cardioid beampattern which has only one null at 180°.

From [6], the beampattern of a uniform linear DMA with a small sensor spacing can be described as,

\[
B_{M,N}(\omega, \theta) = \sum_{n=0}^{N} \cos^p \theta \left[ \left( \frac{j \omega \tau_0}{n!} \right)^p \sum_{m=1}^{M} (m-1)^p H_m(\omega) \right]
\]  

(14)

where \( M \) is the number of microphones, \( N \) denotes the order of the Maclaurin’s series of \( e^{(m-1) \omega \tau_0 \cos \theta} \), \( \theta \) is the incident angle (0° is the endfire direction), \( \omega \) is the angular frequency, \( \tau_0 = d/c \) is the delay of each successive microphone where \( d \) is the microphone spacing, \( c \) is the sound speed and \( H_m(\omega) \) is the weight function applied to the \( m \)th microphone output.

For the case: \( M = 2 \), the order \( N = 1 \) can approximates \( B_{M,N} \) well due to the small \( \tau_0 \). Eq. (14) is simplified as,

\[
B_{2,1}(\omega, \theta) = H_1(\omega) + H_2(\omega) + j H_2(\omega) \omega \tau_0 \cos \theta
\]  

(15)

To build a frequency-invariant first-order DMA beampattern, Eq. (15) is rewritten as,

\[
B_{2,1}(\theta) = a_0 + a_1 \cos \theta
\]  

(16)

Compare Eqs. (15) and (16), we get,
In all of the systems, the placement of sound sources is the same as shown in Fig. 1 and the microphone spacing is 2 cm. The target source data consists of 20 sentences spoken by different speakers from TIMIT database. The interference database consists of two kinds of noise: stationary noise (SN) and non-stationary noise (competing talker (CT)). The sampling frequency are all 16kHz. The received noisy speech is obtained by convolving the original signal with the room impulse responses (RIR). The RIR is measured by the source-image method [9] in a room with dimensions 4.6m*5m*2.8m. The reverberation time $T_{60} \approx 80$ ms. Several metrics is computed to evaluate the speech enhancement performance: the perceptual evaluation of speech quality (PESQ), the short-time objective intelligibility (STOI) measure and the frequency-weighted segmental SNR (fwSNR) [10], [11]. For all of them, a larger score indicates better performance.

To evaluate the effectiveness of spatial filtering, we test the algorithms in case of one noise located at $\theta$ azimuth. Two kinds of noise (SN (‘babble’ noise) and CT) are tested respectively. $\theta$ is changed from 0° to 180° with a 5° step size. The input SNR is 0dB. The averaged PESQ is displayed in Fig. 4, which can be considered as a spatial response diagram of directional noise. The proposed method achieves much better performance than other algorithms at almost all the azimuths and is robust to both SN and CT. The performance of BScoh deteriorates at 180° and GSC can only work well in condition of stationary noise source.

To assess the the speech enhancement performance in various-SNR and multi-source noise scenarios, we test the algorithms at five SNR levels ($-10$dB, $-5$dB, 0dB, 5dB, 10dB) and two noise location configurations (Scenario 1: ‘babble’ noise at 60°, ‘factory’ noise at 120°; Scenario 2: ‘babble’ noise at 180°, two CTs at 90° and 270°). The PESQ, STOI and fwSNR results are displayed in Fig. 5. For every metric, the proposed method outperforms other comparisons at all the noise configurations. Hence the proposed algorithm is potential to work in complex noise scenarios.

### 4. Conclusion

In this letter, we present a joint speech enhancement algorithm combining the coherence-based algorithm and the differential beamforming, which is implemented by two-step spatial filtering and proven to improve the speech quality and intelligibility more effectively than the comparisons. Due to the small size of the microphone array and low computation cost of the algorithm, we offer a potential realizable method that can be applied in portable or compact equipment.

### Acknowledgments

This work was supported by the National Key Research and Discovery Plan No. 2017YFF0210903 and the National Natural Science Foundation of China (grant 11433002).

### References

[1] J. Benesty, J. Chen, and Y. Huang, Microphone Array Signal Processing, Springer Berlin Heidelberg, 2008.
[2] R. Ma, G. Liu, Q. Hao, and C. Wang, “Smart microphone array design for speech enhancement in financial VR and AR,” 2017 IEEE SENSORS, pp.1–3, 2017.
[3] N. Yousefian and P.C. Loizou, “A dual-microphone speech enhancement algorithm based on the coherence function,” IEEE Trans. Audio, Speech, Language Process., vol.20, no.2, pp.599–609, 2012.
[4] N. Yousefian and P.C. Loizou, “A dual-microphone algorithm that can cope with competing-talker scenarios,” IEEE Trans. Audio, Speech, Language Process., vol.21, no.1, pp.145–155, 2013.
[5] Y.R. Zheng, R.A. Goubran, and M. El-Tanany, “Experimental evaluation of a nested microphone array with adaptive noise cancellers,” IEEE Trans. Instrum. Meas., vol.53, no.3, pp.777–786, June 2004.
[6] L. Zhao, J. Benesty, and J. Chen, “Design of robust differential microphone arrays,” IEEE/ACM Trans. Audio, Speech, Language Process., vol.22, no.10, pp.1455–1466, 2014.
[7] M.S. Brandstein and D.B. Ward, Microphone Arrays: Signal processing techniques and applications, Springer, 2001.
[8] J. Chen, J. Benesty, and Y. Huang, “Time delay estimation in room acoustic environments: An overview,” EURASIP J. Adv. Signal Process., vol.2006, no.1, 026503, 2006.

[9] J.B. Allen and D.A. Berkley, “Image method for efficiently simulating small-room acoustics,” J. Acoust. Soc. Am., vol.65, no.4, pp.943–950, 1979.

[10] Y. Hu and P.C. Loizou, “Evaluation of objective quality measures for speech enhancement,” IEEE Trans. Audio, Speech, Language Process., vol.16, no.1, pp.229–238, 2007.

[11] C.H. Taal, R.C. Hendriks, R. Heusdens, and J. Jensen, “A short-time objective intelligibility measure for time-frequency weighted noisy speech,” IEEE International Conference on Acoustics, Speech and Signal Processing, pp.4214–4217, 2010.