Microphone Mini-array Based Speech Enhancement Using ISDS-MGSC Algorithm

Qingning Zeng [*] Qinghua Liu [*] Shan Ouyang [*] and Waleed Abdulla [†]
[*] Guilin University of Electronic Technology, Guilin
E-mail: hmzengqn@guet.edu.cn Tel: +86-773-2991320
[†] University of Auckland, Auckland, New Zealand
E-mail: w.abdulla@auckland.ac.nz Tel: +64-9-3737599ext.88969

Abstract—**Microphone Mini-array based speech enhancement is a challenging research subject since the aperture of microphone array is greatly limited. An efficient algorithm is presented in this paper for microphone mini-array based speech enhancement systems. The algorithm proposes an Improved Shared Distorted Signal (ISDS) method for Modified Generalized Sidelobe Canceller (MGSC). The idea is to apply ISDS to cancel the speech signal in the blocking subsystem of MGSC, which introduces noticeable speech improvement. In experiment an average improvement over several noise sources settings of ~14dB is achieved as compared to ~3dB improvement by MGSC.**

I. Introduction

Speech enhancement is important in many fields such as speech communication and speech recognition. Microphone array based speech enhancement systems perform much better than their single-channel speech enhancement counterparts [1].

However, in many applications the necessary aperture of the microphone array is too big to suit that application. For example, when it is applied to a mobile phone, hearing aid or personal data assistance (PDA), the employed microphone array should be small enough to be embedded into such small devices[2-5]. In this paper we call the array which can be embedded in these small devices as mini-array. Obviously, mini-array based speech enhancement is more challenging than those using the common arrays since the aperture of the mini-array and the number of the microphones employed are greatly limited.

Many effective algorithms used for the common array based speech enhancement have little or even no effect when applied on systems using the mini-array structure. For example, Modified Generalized Side-lobe Canceling (MGSC) algorithm is one of the most effective algorithms in common array based speech enhancement [6-9]. Yet, it has very limited effect whence used with mini-array structures. Also, for the basic GSC algorithm, it may even degrade the Signal to Noise Ratio (SNR) of the original noisy speech when it is used with the mini-arrays. In this paper, we propose an Improved Shared Distorted Signal (ISDS) method for the blocking subsystem of the Modified Generalized Sidelobe Canceling (MGSC) algorithm. The method takes into consideration the highly correlated signals acquired by the adjacent microphones in the mini-array structures. The proposed ISDS-MGSC algorithm effectively blocks the speech signal in the blocking subsystem, which implicates improved enhancement performance. The experimental results verified the advantages of the proposed algorithm. The paper is structured as follows: Section II describes the proposed algorithm. Section III shows the performance of the proposed algorithm under several settings for noise type and sources location. Finally, section IV derives conclusions about the proposed algorithm.

II. Proposed Algorithm

Suppose speech $s(k)$ and noise (or noises) $n(k)$ are generated by independent sources. They arrive at microphone M_i through multi-paths and are acquired by M_i as $s_i(k)$ and $n_i(k)$ respectively. The actual signal acquired by microphone M_i can be represented by

$$x_i(k) = s_i(k) + n_i(k),\ k = 0,1,2,\cdots, i = 1,2,\cdots,N \quad (1)$$

where N（$N \ge 2$） is the number of microphones employed in the array [10].

A. ISDS-MGSC Algorithm

Fig. 1 is the structure of the proposed ISDS-MGSC algorithm, where FBF is a fixed beamformer, z^{-d} is a delay unit, VAD is a Voice Activity Detector and MANC is a Multi-channel Adaptive Noise Canceller. The adaptive filter MANC is adapted only during Non Voice Period (NVP) to minimize the power of the system output $e(k)$, and its coefficients are kept constant during Voice Period (VP).

The main difference between ISDS-MGSC and MGSC lies in their blocking subsystems. The system in the dotted rectangular as shown in Fig.1 is the subsystem of ISDS-MGSC. It consists of an adaptive filter A and N adaptive filters B_i ($i = 1,2,\cdots,N$). Filter A has multi-channel inputs and each adaptive filter B_i has one channel input. Filter A adapts its coefficients to cancel the residual noise in the primarily enhanced speech $\tilde{y}(k)$ by FBF during pure noise period (i.e. NVP), and keeps its coefficients constant during VP. In this way a speech-related signal $\hat{y}(k) = e_A(k)$ can be established and it is input to adaptive filters B_i to cancel the

This work is supported by the Nature Science Foundation of China (60272038), and the Science Foundation of Guangxi (0639028, 0832007Z)

speech signal in noisy speech $x_i(k)$ to get estimated noise $\tilde{n}_i(k)$ ($i=1,2,\cdots,N$). Thus, N channels of estimated noises can be obtained to put into MANC. Obviously, the subsystem in the dotted rectangular in Fig. 1 plays the function to block speech signal for MGSC system.

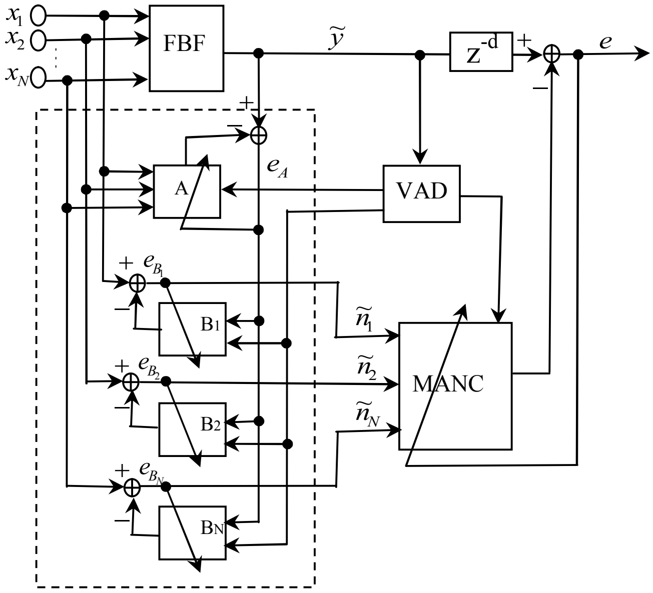

Fig. 1 ISDS-MGSC

B. Signal Blocking Principal

In the blocking subsystem of ISDS-MGSC, for every channel of noisy speech signal $x_i(k)$ we need to block the speech signal $s_i(k)$ to get the estimation $\tilde{n}_i(k)$ of noise $n_i(k)$ ($i=1,2,\cdots,N$).

Not losing generality, we may assume the primarily enhanced speech signal by fixed beamformer FBF to be

$$\tilde{y}(k) = s_1(k) + \tilde{n}(k) \tag{2}$$

$\tilde{y}(k)$ usually contain less noise than any noisy speech $x_i(k)$. If a Delay And Sum (DAS) algorithm is employed for FBF, we need only to select $x_1(k)$ to be the reference signal for time aligning.

In Non Voice Period (NVP), consider

$$\hat{y}(k) = e_A(k) = \tilde{n}(k) - \mathbf{wn}(k) \tag{3}$$

where $\mathbf{wn}(k)$ is the output of the adaptive FIR filter A, $\mathbf{w}$ is the $1 \times N(L+1)$ -dimension coefficient vector of filter A, and L is the sample delay number for every channel of the signal input to filter A.

$$\mathbf{w} = (\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_N) \tag{4}$$

$$\mathbf{w}_i = (w_{i0}, w_{i1}, \cdots, w_{iL})$$

$\mathbf{n}(k)$ is a $N(L+1)\times 1$-dimension vector for noise.

$$\mathbf{n}(k) = [\mathbf{n}_1(k), \mathbf{n}_2(k), \cdots, \mathbf{n}_N(k)]^T \tag{5}$$

$$\mathbf{n}_i(k) = [n_i(k), n_i(k-1), \cdots, n_i(k-L)]$$

Adjust the coefficient vector $\mathbf{w}$ of filter A to minimize the power of $e_A(k)$ and suppose the optimal coefficient vector to be

$$\begin{aligned} \mathbf{w}^* &= (\mathbf{w}_1^*, \mathbf{w}_2^*, \cdots, \mathbf{w}_N^*) \\ &= (w_{10}^*, w_{11}^*, \cdots, w_{1L}^*, w_{20}^*, w_{21}^*, \cdots, w_{2L}^*, \\ &\qquad \cdots\cdots, w_{N0}^*, w_{N1}^*, \cdots, w_{NL}^*) \end{aligned} \tag{6}$$

The corresponding minimum error is noted as $e_A^*(k)$. Since the microphones are closed placed in a mini-array, the noise signals between different channels are, generally speaking, highly correlated. As a results, residual noise $e_A^*(k)$ will generally be much smaller than $\tilde{n}(k)$.

Then, in the Voice Period (VP) which immediately follows the previous NVP, we assume the environment for noise transmitting remains unchanged or changes quite slowly. Under this assumption, we may keep the coefficient vector of filter A unchanged, which is optimized during the previous NVP. Thus the output of filter A during this VP would be

$$\begin{aligned} \mathbf{w}^*\mathbf{x}(k) &= \mathbf{w}^*[\mathbf{s}(k) + \mathbf{n}(k)] \\ &= \mathbf{w}^*\mathbf{s}(k) + [\tilde{n}(k) - e_A^*(k)] \end{aligned} \tag{7}$$

where $\mathbf{x}(k)$ and $\mathbf{s}(k)$ represent the noisy speech vector and pure speech vector with the forms like $\mathbf{n}(k)$ in (5). Thus, according to（2）and（7）

$$\begin{aligned} e_A(k) &= \tilde{y}(k) - \mathbf{w}^*\mathbf{x}(k) \\ &= [s_1(k) + \tilde{n}(k)] - [\mathbf{w}^*\mathbf{s}(k) + \tilde{n}(k) - e_A^*(k)] \\ &= p(k) + e_A^*(k) \end{aligned} \tag{8}$$

$$\begin{aligned} p(k) &= s_1(k) - \mathbf{w}^*\mathbf{s}(k) = s_1(k) - \sum_{i=1}^{N}\sum_{j=0}^{L} w_{ij}^* s_i(k-j) \\ &= s_1(k) - \sum_{i=1}^{N}\sum_{j=0}^{L} w_{ij}^* h_{s_1 s_i}(k-j) * s_1(k-j) \end{aligned} \tag{9}$$

where $h_{s_1 s_i}(k)$ is impulse response of the intermediate media from speech signal $s_1(k)$ to $s_i(k)$.

From（8）, it can be found that the residual noise $e_A^*(k)$ in $\hat{y}(k)$ is generally much less than the noise $\tilde{n}(k)$ in $\tilde{y}(k)$. From (9), $p(k)$ can be regarded as a distorted version of the pure speech signal $s_1(k)$ and thus is related with $s_i(k)$ ($i=1,2,\cdots,N$). As a result, $\hat{y}(k) = e_A(k)$ is more suitable than $\widetilde{y}(k)$, which is used by Hoshuyama [9], to be input to adaptive filter B_i to cancel the speech $s_i(k)$ in $x_i(k)$ to get the noise estimation $\widetilde{n}_i(k)$.

Like ANC algorithm, we may suppose the noise and the speech are uncorrelated. In order to cancel the speech signal in $x_i(k)$, we need only to adjust the coefficients of the filter B_i to minimize the power of $e_{B_i}(k)$.

$$e_{B_i}(k) = x_i(k) - y_{B_i}(k) \tag{10}$$

where $y_{B_i}(k)$ is the output of filter B_i . For simplicity, filter B_i may also take FIR type.

However, $\hat{y}(k) = e_A(k)$ may still contain residual noise due to the incomplete correlation between different channel noises. This fact will cause partial noise cancellation in the estimated noise during VP and results in the residual noise in the final enhanced signal $y(k)$ during VP being stronger than the residual noise during NVP. To make a steady residual noise in the final output, we can adjust the coefficients of filter B_i not only during VP but also during NVP. That is, to adjust the coefficients of every filter B_i all the time.

To sum up, for the blocking process, we only need to adjust the coefficients of filter A in Fig.1 during NVP to minimize the power of $e_A(k)$ and to adjust the coefficients of every filter B_i all the time to minimize the power of $e_{B_i}(k)$. Then the output $e_{B_i}(k)$ would be the estimation $\widetilde{n}_i(k)$ of the pure noise $n_i(k)$.

III. Experimental Results

In the experiment, four small microphones were used to construct a planar array with an aperture of less than 5cm. The speech and the noises were generated concurrently by loudspeakers from different locations. The speech data was from a section of recorded speech in the computer and the noise data was from the NoiseX92 database. The sampling rate used to digitize the acquired signals was 8 kHz.

The experiment was made in a common study room of dimensions 5x4x2.8m. The array was put on a desk. The center of the array was 1.4m from the front wall, 1.8m from the left wall and 1.23m from the floor. There were two sofas, a cabinet and another two desks in the room. The room had two glass windows and a wooden door.

One of the experiment cases listed as Case 9 is shown in Fig.2. For simplicity, the figure is a planar one since the loudspeakers emitting speech and noises have almost the same height from the floor as the array in the experiment. In this case, the speech loudspeaker was placed 30cm in front of the microphone array at (0,30). Noise loudspeakers were concurrently activated to emit Volvo, Leopard, Factory2 and White noises. They were positioned at (-100,100), (50,50), (200,250) and (0,100)cm respectively. Other 8 cases tested are: Case 1 Speech at (0,30) and Leopard noise at (0,100). Case 2 Speech at (0,30) and Leopard noise at (200,250). Case 3. Speech at (0,30) and Volvo noise at (-100,100). Case 4 Speech at (200,250) and Volvo noise at (-100,100). Case 5 Speech at (0,30), Volvo noise at (-100,100) and Leopard noise at (50,50). Case 6 Speech at (0,30), Volvo noise at (-100,100) and Factory2 noise at (200,250). Case 7 Speech at (0,30), Leopard noise at (50,50) and Factory2 noise at (200,250). Case 8 Speech at (0,30), Volvo noise at (-100,100), Leopard noise at (50,50) and Factory2 noise at (200,250).

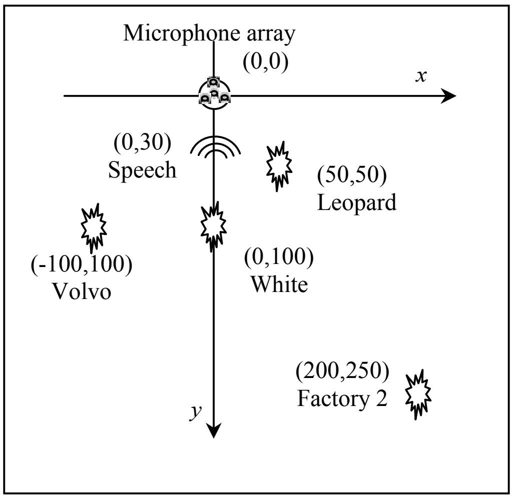

Fig.2 Case 9 of the experiment environments

Table 1 SNRs (dB) of noisy speech and enhanced speech through GSC, MGSC and ISDS-MGSC

Algorithm Case	Noisy Speech	GSC	MGSC	ISDS-MGSC
1	2.64	2.29	6.86	19.90
2	13.29	8.98	16.50	23.55
3	11.58	5.06	13.88	23.32
4	7.38	9.30	12.12	20.01
5	2.62	2.38	6.90	19.60
6	13.00	7.46	13.45	23.36
7	2.56	2.20	5.78	19.35
8	2.54	2.16	5.80	19.01
9	2.52	1.98	5.22	18.30
Average	**6.46**	**4.65**	**9.61**	**20.71**
Improved		**-1.81**	**3.15**	**14.25**

Table 1 shows the SNRs and SNR improvements of the original and enhanced speeches by use of different algorithms including the GSC, MGSC and ISDS-MGSC. The last two rows are the average SNRs and average SNR improvements.

In table 1, the original noisy speech signal is $x_1(k)$ acquired from microphone M_1. Here the SNR is calculated by

$$SNR = 10\log_{10}\left[\left(\alpha\sum_{k\in T_s}x^2(k)-\sum_{k\in T_n}x^2(k)\right)\left(\sum_{k\in T_n}x^2(k)\right)^{-1}\right] \qquad (11)$$

where $x(k)$ is noisy speech signal, T_s is the set of the signal samples containing in Voice Period (VP) while T_n is the set of the signal samples containing in Non Voice Period (NVP), $\alpha = m(T_n)/m(T_s)$, here $m(T_n)$ and $m(T_s)$ represent the numbers of the samples in T_n and T_s respectively.

In the ISDS-MGSC algorithm, use microphone M_1 as the standard calibrating microphone, a correlation method to calculate the time delays and the DAS algorithm for fixed beamformer FBF. VAD employs an energy and zero-crossing rate method. Whenever VAD is failed, use the artificially decided results about VP and NVP. The adaptive FIR filter MANC has a length of 120 and a LMS adaptation algorithm with learning rate $\mu = 0.01$. In ISDS-MGSC processing, the length of filter A is 100 and all filters B_i ($i = 1,2,3,4$) have the same length of 40. All filters employ the LMS adaptation algorithm with learning rate $\mu = 0.01$.

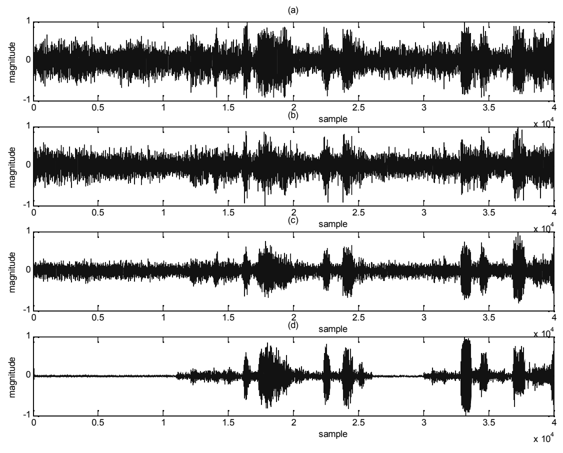

Fig. 3 Speech enhancement results
(a) Noisy speech (b) Enhanced speech by GSC
(c) Enhanced speech by MGSC
(d) Enhanced speech by ISDS-MGSC

Fig. 3 shows the signals concerned in case 9. Fig.3 (a) is the time domain waveform of noisy speech signal $x_1(k)$ from microphone M_1. Its SNR=2.25 dB. Fig.3 (b) is the enhanced speech by GSC with SNR=1.98dB. Fig.3 (c) is the enhanced speech by MGSC with SNR=5.22 dB. Fig.3 (d) is the enhanced speech by ISDS-MGSC with SNR=18.30dB.

From Table 1 and Fig.3, we can find that the proposed ISDS-MGSC algorithm achieves much more SNR improvement than conventional GSC and MGSC. Listening to the original noisy speech and the enhanced speech by above algorithms, we can also find that the enhanced speech by the proposed ISDS-MGSC algorithm has the highest quality. Other quality evaluation methods for enhanced speech also show the advantage of the proposed algorithm.

IV. CONCLUSIONS

ISDS-MGSC algorithm is proposed for microphone mini-array based speech enhancement systems. By introducing Improved Shared Distorted Signal (ISDS) method, the signal blocking subsystem of the Modified Generalized Sidelobe Canceller (MGSC) efficiently blocks the speech signal and offers better noise estimations, which can lead to better speech enhancement result for MGSC. Experimental results verified the advantages of the proposed algorithm. As shown in Table 1, an average improvement of ~14dB is achieved by the proposed algorithm as compared to ~3dB improvement by MGSC and ~-1.81dB improvement by GSC.

REFERENCES

[1] J. Benesty, S. Makino, J. Chen. *Speech Enhancement*. Berlin: Springer, 2005

[2] Greenberg J E, Zurek P M. "Microphone-array Hearing Aids", in M. Brandstein and D. Ward eds Microphone Arrays: *Signal Processing Techniques and Application*. Springer, Berlin, pp.229-253, 2001

[3] A. Spriet, G. Rombouts, M. Moonen, J. Wouters. "Combined Feedback and Noise Suppression in Hearing Aids". *IEEE Transactions on Audio, Speech and Language Processing,* vol.15, no.6, pp.1777-1790, 2007

[4] M. Ogasawara, T. Nishino, K. Takeda. "A Small Dodecahedral Microphone Array for Blind Source Separation". *IEEE proceedings of 2010 ICASSP,* pp. 229-232, 2010

[5] H. Puder. Acoustic noise control: "An Overview of Several Methods Based on Applications in Hearing Aids". *IEEE Pacific Rim Conference on Communications, Computers and Signal Processing,* 2009, pp. 871-876, 2009

[6] G. L. Fudge, D. A. Linebarger. "A Calibrated Generalized Sidelobe Canceller for Wideband Beamforming". *IEEE Transaction on Signal Processing,* 1994, vol.42, no.10, pp. 2871-2875, 1994

[7] S. Gannot, D. Bueshitein, E. Weinstein. "Analysis of the Power Spectral Deviation of the General Transfer Function GSC". *IEEE Transaction on Signal Processing,* vol.52, no.4, pp. 1115-1121, 2004

[8] E. Warsitz, A. Krueger, R. Haeb-Umbach. "Speech Enhancement with a New Generalized Eigenvector Blocking Matrix for Application in a Generalized Sidelobe Canceller". *IEEE International Conference on Acoustics, Speech and Signal Processing , Las Vagas,* pp.73-76, 2008

[9] O. Hoshuyama, A. Sugiyama, A. Hirano. "A Robust Adaptive Beamformer for Microphone Arrays with a Blocking Matrix Using Constrained Adaptive Filters". *IEEE Transactions on Signal Processing,* vol.47, no.1, pp.2677-2684, 1999

[10] Q. Zeng, W. Abdulla. "Speech Enhancement by Multichannel Crosstalk Resistant ANC and Improved Spectrum Subtraction". *EURASIP Journal on Applied Signal Processing*, vol.2006, Article ID 61214, 10 pages, 2006