BLIND MULTICHANNEL IDENTIFICATION FOR SPEECH DEREVERBERATION AND ENHANCEMENT

Zhu Liang Yu

Center for Signal Processing,
Nanyang Technological University,
Singapore, 639798
Email: ezlyu@ntu.edu.sg

Meng Hwa Er

School of EEE,
Nanyang Technological University,
Singapore, 639798
Email: emher@ntu.edu.sg

ABSTRACT

A multichannel wideband signal dereverberation and enhancement method is proposed in this paper. It uses the blindly estimated impulse responses (IRs) relating the signal source and each sensor to form a multiple input inverse filter (MINT) for speech dereverberation and ehancement with extended generalized sidelobe canceller (GSC). With the replacement of MINT for fixed beamformer and modification of blocking matrix in conventional GSC, the resulting extended GSC not only dereverberate the distorted target signal, but also suppress the interference/noise. Computer simulation results show the effectiveness of the proposed method.

1. INTRODUCTION

Multichannel speech dereverberation and enhancement is an important research topic in speech applications. It is known that an acoustic enclosure usually reduces the intelligibility of speech transmitted through it because the transmission path is not ideal. The microphone received signal consists of the direct path signal from the source, multipath reflections and reverberation. Moreover, signal also contains background interference and noise, which should be suppressed by further processing.

To combat the reverberation/multipath effects and enhance the target signal, the matched filter (MF) array [1] was proposed. The study showed that the MF array significantly improves the quality of signal captured in reverberant environment. However, this method has two drawbacks: Firstly, a large amount of sensors are required to achieve high dereverberation and interference suppression performance. The other one is that the channel IRs must be known a priori. It's a great challenge to obtain better dereverberation performance with a small number of sensors.

Miyoshi etc. proposed a new method [2], we call it as MINT, to find the exact inverse of a point in a room by using multiple sensors. In this method, the channel IR relating the signal source and each sensor is modelled as FIR filter. It was proved in [2] that the exact inverse filter can be obtained provided that the transfer functions of multiple IRs are coprime. The minimum number of sensors is two. This method indicates that the perfect dereverberation performance can be achieved with small number of sensors. In practical applications, if more sensors are available, the chances of the IRs being non-coprime diminishes. However, this method still has the drawback that the IRs are assumed to be known a priori. Moreover, the MINT obtained does not have the capability of noise suppression, which is different from MF array.

Either MF array or MINT used, the identification of the IRs is necessary. The IRs identification based on training signal [3] is not practical when the signal source moves or the environment changes. A potentially better solution is the blind channel identification (BCI) technique [4, 5]. Although there are many BCI methods proposed in literature, few of them can be applied in acoustic applications due the large length of multi-channel IR, nonwhite inputs, intensive computational load, etc. The normalized multichannel frequency-domain LMS (NMCFLMS) developed in [6] indicated the possibility of using adaptive LMS-type method to estimate the multichannel IRs of large length. However, this method requires very high input SNR, which is too strict for practical applications. In this paper, we use the normalized blind frequency-domain least mean square (NBFLMS) method [7] to estimate the multi-channel IRs. NBFLMS is insensitive to input noise and has better estimation performance than NMCFLMS [7].

With the blindly estimated channel IRs, we can use MINT to dereverberate the distorted speech with small number of sensors. To calculate the MINT, the row action projection (RAP) method is adopt due to its simple algorithm and high performance [8]. Since the MINT does not have the capability of noise suppression, we propose to combine the MINT with GSC. The fixed beamforer of GSC is replaced by the MINT. Moreover, to combat the signal cancellation problem of GSC in reverberant environment, its blocking matrix should also be modified. In [9], we proposed an extended GSC as well as its implementation in time and frequency domain for such purpose.

This paper is organized as follows. The system model is reviewed in Section 2. In Section 3, the MINT for speech dereverberation as well as RAP method are discussed. The NBFLMS is then presented in Section 4. The extended GSC is discussed in Section 5. The resulting GSC with MINT has perfect signal dereverberation and noise suppression capability. Some numerical results are shown in Section 6. In Section 7, a brief conclusion is given.

2. SYSTEM MODEL

Notations used in this paper are defined before we formulate the problem and develop the algorithm. $E\{\cdot\}$, $(\cdot)^*$, $(\cdot)^T$, $(\cdot)^H$, $\star$ and $||\cdot||$ stand for mathematical expectation, complex conjugate, vector/matrix transpose, vector/matrix Hermitian transpose, linear convolution, and Euclidean norm, respectively. The identity matrix is $\mathbf{I}$.

There are M sensors used. Each sensor picks up the target sig-

nal as well as the environment noise. The signal $s(k)$ propagates through the ith channel with IR $h_{i,k}, i = 1, 2, \cdots, M$, and is corrupted by additive environment noise $n_i(k)$. The received signal $x_i(k)$ of ith channel is expressed as

$$x_i(k) = h_{i,k} \star s(k) + n_i(k), \quad i = 1, 2, \cdots, M. \tag{1}$$

In most applications, the transfer function $h_{i,k}$ can be approximated as FIR filters with length L and coefficient vector

$$\mathbf{h}_i = [h_{i,0}\ h_{i,1}\ \cdots\ h_{i,L-1}]^T. \tag{2}$$

Here, we assume that all the IRs are fixed or changing very slowly although the time variation can be tracked by the proposed method.

3. SPEECH DEREVERBERATION USING MINT

Assume that the channel IRs are known. Filters $g_{i,k},\ i = 1, \cdots, M$ can be used to form MINT It was proved in [2] that the exact inverse of room transfer function exists uniquely if these two conditions hold:

1. All the transfer functions $\{Z(h_{i,k})\}$ are co-prime, where $Z(\cdot)$ denotes Z transform.
2. The order of $g_{i,k}$ is less than $h_{i,k}$.

In this paper, we assume that $h_{i,k}$ and $g_{i,k}$ have same order. The total response $d(k)$ of length $2L-1$ relating source and MINT output is

$$d(k) = \sum_{i=1}^{M} h_{i,k} \star g_{i,k} \tag{3}$$

In matrix form, we can express

$$\mathbf{d} = [\mathbf{H}_1^T\ \cdots\ \mathbf{H}_M^T]\mathbf{g} = \mathbf{H}\mathbf{g} \tag{4}$$

where

$$\begin{aligned}
&\mathbf{d} = [d(0)\ d(1)\ \cdots\ d(2L-1)]^T \\
&\mathbf{H}_i = \begin{bmatrix} h_i(0) & \cdots & h_i(L-1) & 0 & \cdots & 0 \\ 0 & h_i(0) & \cdots & h_i(L-1) & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & h_i(0) & \cdots & h_i(L-1) \end{bmatrix} \\
&\mathbf{g} = [\underbrace{g_1(0)\ \cdots\ g_1(L-1)}_{\mathbf{g}_1^T}\ \cdots\ \underbrace{g_M(0)\ \cdots\ g_M(L-1)}_{\mathbf{g}_M^T}]^T
\end{aligned} \tag{5}$$

In the speech dereverberation applications, the ideal total response vector $\mathbf{d}$ is

$$\mathbf{d} = [1\ 0\ \cdots\ 0]^T. \tag{6}$$

If $\mathbf{h}_i$ is known, $\mathbf{g}$ can be calculated by solving (4). In this paper, we adopt the RAP method [8], which is an efficient way to solve (4). Moreover, the RAP provides stable and fast solution. With RAP method, the vector $\mathbf{g}$ can be updated using

$$\begin{aligned}
e(k) &= d(p) - \mathbf{h}_p\mathbf{g}(k) \\
\mathbf{g}(k+1) &= \mathbf{g}(k) + \lambda\frac{e(k)}{||\mathbf{h}_p||^2}\mathbf{h}_p^T.
\end{aligned} \tag{7}$$

where $\mathbf{h}_p$ is the pth row of $\mathbf{H}$ at kth iteration and $p = k\ mod\ (M+1)$. The relaxation parameter λ should be select in $[0, 1]$. If the measurement is noisy, usually a small value is chosen for λ. Otherwise, a large value is chosen to speed up convergence.

4. BLIND MULTICHANNEL IDENTIFICATION

The derivation of NBFLMS is based on cross relation (CR) criteria [10, 11] in frequency domain using overlap-save method [12]. Refer to [7] for the detailed derivation of the NBFLMS.

Theorem 1 *The constrained NBFLMS algorithm is*

$$\begin{aligned}
\bar{\boldsymbol{h}}_k(m) = \tilde{\boldsymbol{h}}_k(m-1) &- \rho\boldsymbol{\mathcal{W}}_{N'\times N'}^{10}\boldsymbol{\mathcal{P}}^{-1}(m) \\
&\sum_{i=1}^{M}\left[\boldsymbol{\mathcal{R}}_{x_ix_i}(m)\tilde{\boldsymbol{h}}_k(m) - \boldsymbol{\mathcal{R}}_{x_ix_k}(m)\tilde{\boldsymbol{h}}_i(m)\right] \\
&k = 1, 2, \cdots, M
\end{aligned}$$

The unconstrained NBFLMS algorithm is

$$\begin{aligned}
\bar{\boldsymbol{h}}_k(m) = \tilde{\boldsymbol{h}}_k(m-1) &- \rho\boldsymbol{\mathcal{P}}^{-1}(m) \\
&\sum_{i=1}^{M}\left[\boldsymbol{\mathcal{R}}_{x_ix_i}(m)\tilde{\boldsymbol{h}}_k(m) - \boldsymbol{\mathcal{R}}_{x_ix_k}(m)\tilde{\boldsymbol{h}}_i(m)\right] \\
&k = 1, 2, \cdots, M
\end{aligned}$$

where

$$\begin{aligned}
\boldsymbol{\mathcal{R}}_{x_ix_i}(m) &\approx \beta\boldsymbol{\mathcal{R}}_{x_ix_i}(m-1) + \boldsymbol{\mathcal{D}}_{x_i}^H(m)\boldsymbol{\mathcal{D}}_{x_i}(m) \\
\boldsymbol{\mathcal{R}}_{x_ix_k}(m) &\approx \beta\boldsymbol{\mathcal{R}}_{x_ix_k}(m-1) + \boldsymbol{\mathcal{D}}_{x_i}^H(m)\boldsymbol{\mathcal{D}}_{x_k}(m) \\
\boldsymbol{\mathcal{P}}(m) &= \sum_{i=1, i\neq k}^{M}\hat{\boldsymbol{\mathcal{R}}}_{x_ix_i}(m) \\
\hat{\boldsymbol{\mathcal{R}}}_{x_ix_i}(0) &= \boldsymbol{\mathcal{R}}_{x_ix_i}(0) \\
\hat{\boldsymbol{\mathcal{R}}}_{x_ix_i}(m) &= \lambda\hat{\boldsymbol{\mathcal{R}}}_{x_ix_i}(m-1) + \boldsymbol{\mathcal{R}}_{x_ix_i}(m) \\
\boldsymbol{\mathcal{W}}_{N'\times N'}^{10} &= \mathbf{F}_{N'}\begin{bmatrix} \mathbf{I}_{L\times L} & \mathbf{0}_{L\times(N-1)} \\ \mathbf{0}_{(N-1)\times L} & \mathbf{0}_{(N-1)\times(N-1)} \end{bmatrix}\mathbf{F}_{N'}^{-1}
\end{aligned}$$

where $\mathbf{F}_{N'}$ is the $N \times N$ DFT matrix and $\boldsymbol{\mathcal{D}}_{x_i}(m)$ is a diagonal matrix whose diagonal elements are the FFT transform of vector $\tilde{\mathbf{x}}_i(m)$ given by

$$\tilde{\mathbf{x}}_i(m) = [x_i(mN - L + 1)\ \cdots\ x_i(mN + N - 1)]^T.$$

β and λ are forgetting factors.

To avoid trivial solution, the updated filter coefficient vectors are normalized to vector with unit norm.

$$\tilde{\boldsymbol{h}}_k(m) = \frac{\bar{\boldsymbol{h}}_k(m)}{||\boldsymbol{h}(m)||}, \quad \boldsymbol{h}(m) = [\bar{\boldsymbol{h}}_1(m)\ \cdots\ \bar{\boldsymbol{h}}_M(m)]^T$$

The computational load of NBFLMS in Theorem 1 is low since the matrices $\boldsymbol{\mathcal{R}}_{x_ix_k}(m)$ and $\boldsymbol{\mathcal{P}}(m)$ are both diagonal.

5. EXTENDED GENERALIZED SIDELOBE CANCELLER

One of the advantages of frequency domain filtering is that fast Fourier transform (FFT) can be used to reduce the computational complexity. From the property of discrete Fourier transform (DFT), we know that the circular convolution of two finite length sequences can be obtained by transforming both sequences to their respective frequency domain (using FFT), performing an element-by-element multiplication on the transformed samples, and transforming the

result back to time domain (using inverse FFT). Another advantage of frequency domain filtering is that FFT performs an orthogonal transform on the input data. With the power normalization method, it is possible to derive fast convergence algorithm for adaptive updating. In this paper, the algorithm is derived based on the overlap-save technique [12] for frequency domain filtering.

With the blindly estimated channel IRs, the GSC-ATF is formed in frequency domain with MINT to replace the fixed beamformer in [9]. The output $\mathbf{y}(m)$ of the fixed beamformer is

$$\begin{aligned}\mathbf{y}(m) &= \sum_{i=1}^{M} \mathbf{y}_i(m) = \mathbf{W}_{N\times N'}^{01} \sum_{i=1}^{M} (\breve{\mathbf{X}}_i(m)\tilde{\mathbf{g}}_i) \\ &= \mathbf{W}_{N\times N'}^{01} \mathbf{F}_{N'}^{-1} \sum_{i=1}^{M} \boldsymbol{\mathcal{D}}_{\breve{\mathbf{X}}_i}(m)\tilde{\boldsymbol{g}}_i\end{aligned} \tag{8}$$

The output signal vector of the ith blocking filter $\mathbf{z}_i(m)$ is expressed as

$$\mathbf{z}_i(m) = \mathbf{W}_{N\times N'}^{01} \mathbf{F}_{N'}^{-1} \sum_{j=1}^{M} \boldsymbol{\mathcal{D}}_{\breve{\mathbf{X}}_j}(m)\tilde{\boldsymbol{b}}_{ij}, \tag{9}$$

where

$$\begin{aligned}\tilde{\boldsymbol{b}}_{ij} &= \mathbf{F}_{N'} \mathbf{W}_{N'\times L}^{10} \mathbf{b}_{ij} \\ \mathbf{b}_{ij} &= [b_{ij}(0) \ \cdots \ b_{ij}(L-1)]^T\end{aligned} \tag{10}$$

The output signal vector $\mathbf{v}(m)$ of the multichannel adaptive filter is

$$\mathbf{v}(m) = \mathbf{W}_{N\times N''}^{01} \mathbf{F}_{N''}^{-1} \sum_{i=1}^{M} \boldsymbol{\mathcal{D}}_{\breve{\mathbf{Z}}_i}(m)\tilde{\boldsymbol{w}}_i, \tag{11}$$

where $N'' = L_w + N - 1$ and $\breve{\mathbf{Z}}_i(m)$ is the circulant matrix whose first column vector is the extended vector $\tilde{\mathbf{z}}_i(m)$ expressed as

$$\tilde{\mathbf{z}}_i(m) = [z_i(mN - L_w + 1) \ \cdots \ z_i(mN) \ \cdots \ z_i(mN + N + 1)]^T. \tag{12}$$

The vector $\tilde{\boldsymbol{w}}_i$ is expressed as

$$\begin{aligned}\tilde{\boldsymbol{w}}_i &= \mathbf{F}_{N''} \mathbf{W}_{N''\times L_w}^{10} \mathbf{w}_i = \boldsymbol{\mathcal{W}}_{N''\times L_w}^{10} \boldsymbol{w}_i \\ \mathbf{w}_i &= [w_i(0) \ \cdots \ w_i(L_w - 1)]^T \\ \boldsymbol{w}_i &= \mathbf{F}_{L_w} \mathbf{w}_i\end{aligned} \quad . \tag{13}$$

Define the error vector $\boldsymbol{e}(m)$ in the frequency domain as

$$\boldsymbol{e}(m) = \mathbf{F}_N(\mathbf{y}(m, D) - \mathbf{v}(m)), \tag{14}$$

where $\mathbf{y}(m, D)$ is the delayed signal vector of $\mathbf{y}(m)$,

$$\mathbf{y}(m, D) = [y(mN - L - D + 1) \ \cdots \ y(mN + N - D + 1)]^T. \tag{15}$$

Theorem 2 *The constrained Newton-LMS type algorithm for GSC-ATF is*

$$\begin{aligned}\hat{\boldsymbol{w}}_k(m) &= \hat{\boldsymbol{w}}_k(m-1) - \rho' E\{\bigtriangledown^2 C(m)\}^{-1} \bigtriangledown C(m) \\ &= \hat{\boldsymbol{w}}_k(m-1) + \rho \boldsymbol{\mathcal{W}}_{N''\times N''}^{10} \boldsymbol{\mathcal{R}}_{z_k}(m)^{-1} \\ &\quad \cdot \boldsymbol{\mathcal{D}}_{\breve{Z}_k}^H(m) \boldsymbol{\mathcal{W}}_{N''\times N}^{01} \boldsymbol{e}(m)\end{aligned} \tag{16}$$

and the unconstrained Newton-LMS algorithm is

$$\hat{\boldsymbol{w}}_k(m) = \hat{\boldsymbol{w}}_k(m-1) + \rho \boldsymbol{\mathcal{R}}_{z_k}(m)^{-1} \boldsymbol{\mathcal{D}}_{\breve{Z}_k}^H(m) \boldsymbol{\mathcal{W}}_{N''\times N}^{01} \boldsymbol{e}(m) \tag{17}$$

where ρ is the stepsize.

	1	2	3	4	5
x	0.80	0.80	1.00	1.20	1.20
y	2.0	2.0	2.0	2.0	2.0
z	1.6	1.2	1.6	1.6	1.2

Table 1. Position (x,y,z) of five microphones (in meter)

The computational load of the algorithms in Theorem 2 is low since the matrices $\boldsymbol{\mathcal{D}}_{\breve{Z}_k}^H(m)$ and $\boldsymbol{\mathcal{R}}_{z_k}(m)$ are both diagonal matrix. The matrix inverse and multiplication are simplified to element inverse and multiplication. The implementations in time and frequency domain have identical theoretical performance. However, the latter one has lower computational load and faster convergence speed.

6. NUMERICAL STUDY

In this section, we asses the performance of the proposed method for wideband signal dereverberation and enhancement. A microphone array is used. The acoustic enclosure is a small office room with dimension $(x \times y \times z) = (2.8m \times 3.2m \times 2.2m)$, wall reflection coefficients 0.8 and floor/celling reflection coefficients 0.4. This microphone array is placed in front of the PC monitor for sound capture. The position of the microphones are given in Table I. A source signal is placed at the position $(1.0m, 1.5m, 1.4m)$. The IR relating speech source and each microphone is calculated using image method [13] with sampling rate $8kHz$. In this simulation, the IR length is set as $L = 256$, so that most of the reverberation is taken into account. A white background noise is used in the simulation.

We firstly study the derverberation performance of the MINT. The total IR relating the soruce signal and beamformer output is shown in Fig. 1. It is obvious that the total IR is a very close to a delay impulse, which mean the reverberation is reduced. The comparison of frequency responses (FRs) of total IR and one channel IR is shown in Fig. 2. The FR of total IR is almost flat while the FR of one channel has large spectral fluctuation.

In the following simulation, we show the speech enhancement capability of the proposed method. The channel IRs are estimated when the input SNR is $5dB$. After that, a strong point interference is placed at (2.5m, 2.8m, 1.3m). The average signal-to-interference ratio (SIR) is $0dB$. In Fig. 3(a), the waveform of the clean speech signal is shown. In Fig. 3(b), the received signal of the first microphone is shown. This signal contains strong interference and background noise. The output signal of the proposed method is shown in Fig. 3(c). The noise is suppressed efficiently.

7. CONCLUSION

A method is proposed to use blind channel identification in multichannel wideband signal dereverbeation and enhancement. The proposed method exploits the IRs estimated by NBFLMS to form a MINT, which is used to replace the fixed beamformer in extended GSC. The resulting GSC beamformer not only has perfect dereverberation but also high noise suppression performance with a small number of sensors. Simulation results show the effectiveness of the proposed method.

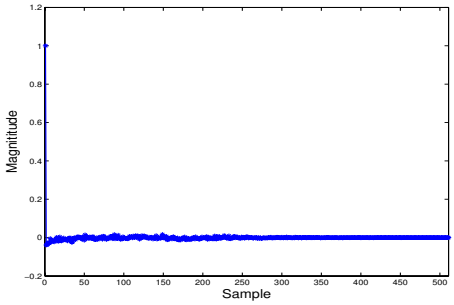


Fig. 1. The total IR relating source signal to MINT output using estimated channel IRs

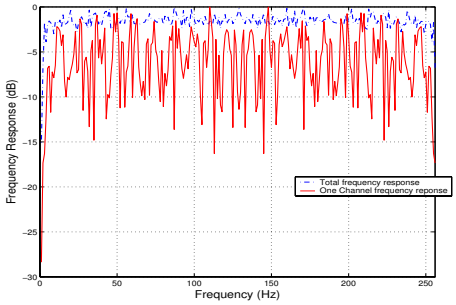


Fig. 2. Comparison of the frequency responses of total IR and one channel IR

8. REFERENCES

[1] J. L. Flanagan, A. C. Surendran, and E. E. Jan, "Spatially selectivity sound capture for speech and audio processing," *Speech Communication*, vol. 13, pp. 207–222, 1993.

[2] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoust. Speech Processing*, vol. 36, no. 2, pp. 145–152, Feb. 1988.

[3] E. E. Jan, P. Svaizer, and J. L. Flanagan, "Matched-filter processing of microphone array for spatial volume selectivity," in *Proc. ISCAS*, May 1995, pp. 1460–1463.

[4] G. B. Giannakis, Y. Hua, P. Stoica, and L. Tong, Eds., *Signal Processing Advances in Wireless and Mobile Communication: Trends in Channel Estimation and Equalization*. Upper Saddle River, NJ: Prentice Hall PTR, 2001, vol. I.

[5] Z. Ding and Y. Li, *Blind Equalization and Identification*. New York: Marcel Dekker, Inc., 2001.

[6] Y. Huang and J. Benesty, "A class of frequency-domain adaptive approaches to blind multichannel identification," *IEEE Trans. Signal Processing*, vol. 51, pp. 11–24, Jan. 2003.

[7] Z. L. Yu and M. H. Er, "A robust frequency domain adaptive blind multichannel identification algorithm for acoustic applications," accepted to present in ICASSP 2004.

[8] R. J. Mammone, *Computational Methods of Signal Recognition and Recovery*. New York: John Wiley & Sons, 1992.

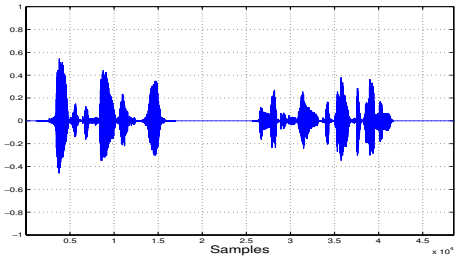
[9] Z. L. Yu and M. H. Er, "An extended GSC with arbitrary transfer functions in time and frequency domain," accepted to present in ISCAS 2004.

[10] H. Liu, G. Xu, and L. Tong, "A deterministic approach to blind equalization," in *IEEE Conference Record of The Twenty-Seventh Asilomar Conference on Signals, Systems and Computers*, vol. 1, Nov. 1993, pp. 751 –755.

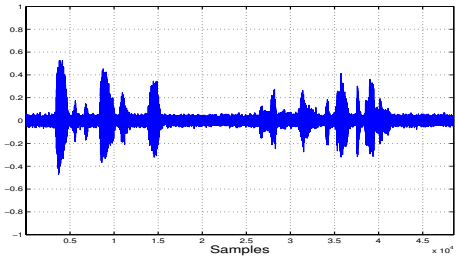
[11] G. Xu, H. Liu, L. Tong, and T. Kailath, "A least-square approach to blind channel identification," *IEEE Trans. Signal Processing*, vol. 43, pp. 2982–2993, Dec. 1995.

[12] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1989.

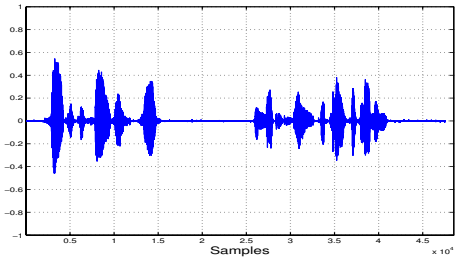
[13] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, April 1979.



(a) Original speech signal



(b) Speech signal received by the first microphone



(c) Speech signal after processed by the propose method

Fig. 3. Speech waveform comparison (0dB SIR for point interference)