行政院國家科學委員會補助專題研究計畫

研究成果報告

免持聽筒汽車套件之迴聲與噪音控制系統

計畫類別：■ 個別型計畫　　□ 整合型計畫

計畫編號：NSC 97-2221-E-009-010-MY3

執行期間：　97 年　8 月　1 日至　100 年　7 月　31 日

計畫主持人：白明憲

共同主持人：

計畫參與人員： 徐和生、趙婉芝、阮星瑋

成果報告類型(依經費核定清單規定繳交)：□精簡報告　■完整報告

處理方式：除產學合作研究計畫、提升產業技術及人才培育研究計畫、列管計畫及下列情形者外，得立即公開查詢

■涉及專利或其他智慧財產權，□一年□二年後可公開查詢

執行單位：國立清華大學動力機械工程研究所

中文摘要

本計劃提出一種能夠實現在電信通訊系統中的麥克風陣列技術運用聲學信號處理方法，此技術能自動偵測訊號位置以及增進語音辨識率，利用兩麥克風之間的相位差可獲得聲源角度進而決定波束開口大小以增進語音辨識率，此外利用波束轉向技術能應付語音訊號不在主軸位置上的情況。運用於車用電子上的電聲元件可能會成非線性迴聲的產生，針對此種無法由線性回聲消除系統消除的成分，我們主要利用沃特拉濾波器及漢默斯坦模式達到消除非線性迴聲的目的。

英文摘要

This study proposes microphone array techniques aimed at enhancing speech recognition. If the noise doesn't come from the rear, on the contrary, it comes from the direction closing to the target source, then the phase difference estimation is used to solve this problem, which can reduce the noise without distortion even when the angle between noise and target source is small. It is found that the ITD threshold in the phase difference estimation plays an important role in enhancing the speech recognition, and hence it has to be optimized. In this paper, GSS is used to search the optimal threshold. If the target source is not from the direction of main lobe, beam steering technique has to be applied to the system. Finally, experiment results are discussed to demonstrate that the performance of the proposed algorithm is better than conventional methods. The nonlinear distortions from low cost audio equipments degrade the performance of linear acoustic echo cancellation system. The approaches of nonlinear adaptive filter are have resorted to discrete-time transversal Volterra filters and block-based Hammerstein model.

I. INTRODUCTION

Automatic speech recognizers (ASRs) have significantly improved in recent years but the performance degrades rapidly in noisy or reverberant environments. Therefore, noisy speech needs to be processed by speech improvement algorithms. For instance, the delay-and-sum (DAS) beamformer is a well known algorithm which is computational efficiency. However, it only performed well for uncorrelated noise. The one-channel noise reduction (NR) technology has been widely applied in the communication community, and was expected to enhance speech recognition. Nevertheless, the improvement of one-channel NR in speech enhancement does not always translate into substantial gains in speech recognition performance, because too aggressive NR destroys the speech features. The one-channel NR encounters the dilemma of noise reduction or distortion. Therefore, microphone array is used in the proposed algorithm, which can ease the tradeoff of the above situation.

Lately, a missing-data approach was suggested to enhance speech recognition in noisy environments, based on designing whether data are reliable. The performance of the missing-data approach is significantly improved comparing to that of the DAS beamformer. Nevertheless, the success of this technique depends on the sufficiency of reliable data and errors in imputation procedures affect the performance. The speech recognition in the environments with non-stationary noise still remains a tough problem. An alternative is the binaural processing which is well known for separating speech signals. Several algorithms were discussed the phenomena of binaural system, such as interaural time difference (ITD) and interaural intensity difference (IID). Recently, computational auditory scene analysis (CASA) systems were developed to construct an ideal binary mask by comparing the signals at the two microphones in binaural systems. Both voice and unvoiced speech signals could be segregated by CASA systems from a noisy environment. However, the computation

of the CASA systems is quite complex.

In this study, microphone arrays are used for enhancing speech recognition in noisy and reverberant environments. Typically, there are two types of microphone array— the broadside and endfire arrays. When the maximum of the array beam pattern (the mainlobe) is alone a line perpendicular to the axial direction of the microphone array, the array is called a broadside array. On the contrary, an endfire array means that the mainlobe is in the direction diaphragm to the microphone axis, “off the end” rather than off the side and consequently the name is endfire array. Since the directivity of super-directive microphone arrays is higher than that of a uniformly summed array in the same condition, it can not only suppress noise and reverberation coming from all directions well but also keep the feature of the target signal from the principal direction. Furthermore, although in many applications the direction of the target signal can’t be predetermined, it is usually in front of the array and disturbances are at the rear. In these cases, the endfire array is suitable than a broadside array.

With the aid of phase-difference estimation, speech signal can be separated well without distortion and the recognition rate is enhanced. Because this algorithm is very sensitive to the choice of ITD threshold in binary masking criterion, how to choose ITD threshold becomes an important problem. An automatic selection of ITD threshold proposed by Kim et al is based on minimizing the cross correlation between the target and the interference signals. However, the performance of the automatic selection algorithm degrades significantly when signal-to-noise ratio (SNR) and the subtending angle between speech and noise signal are small. Hence, this paper proposes an optimal threshold varying with the subtending angle, which is based on finding the minimum of the WER by GSS. Using the optimal ITD threshold proposed in this paper, PDE algorithm can perform well with small SNR and

subtending angle. Furthermore, the selection of volume affects the performance and needs to be adjusted, which is also discussed in this paper. The speech recognition will decrease when the sound source isn't on the main lobe of microphone arrays, and the system needs the beam-steering technique to change the main lobe of array pattern by electronic compensation.

Current AECs which use linear adaptive filter are hinged on the assumption that an acoustic echo path can be modeled as the linear filter. However, there are nonlinear components in the echo path due to low-cost audio equipment. The main nonlinearities are generated from the loudspeaker and power amplifier in the transmission path. The performance of AEC is degraded with this nonlinear distortion. As a result, the nonlinear model has to be taken into account. In this study, we neglect the nonlinearities of A/D and D/A converter. Since the signal power is small in the microphone, the microphone behaves linearly even it is cheap. The nonlinear distortions are generated when the amplifier is overdriven, and they are memoryless. The Hammerstein model consisting of a cascade of memoryless polynomial filter and finite impulse response filter has been proposed to overcome this nonlinearity. Another sort of nonlinear distortion is caused by the loudspeaker since the time constants of their electro-mechanical systems are large compared to the sampling rate, and this nonlinearity behaves with memory. To compensate the nonlinearities with memory, the adaptive Volterra filter has been utilized to model the echo path in the acoustic echo cancellation system. In this study, we propose the modified system for Hammerstein model to avoid instability of convergence.

II. PHASE-DIFFERENCE ESTIMATION (PDE)

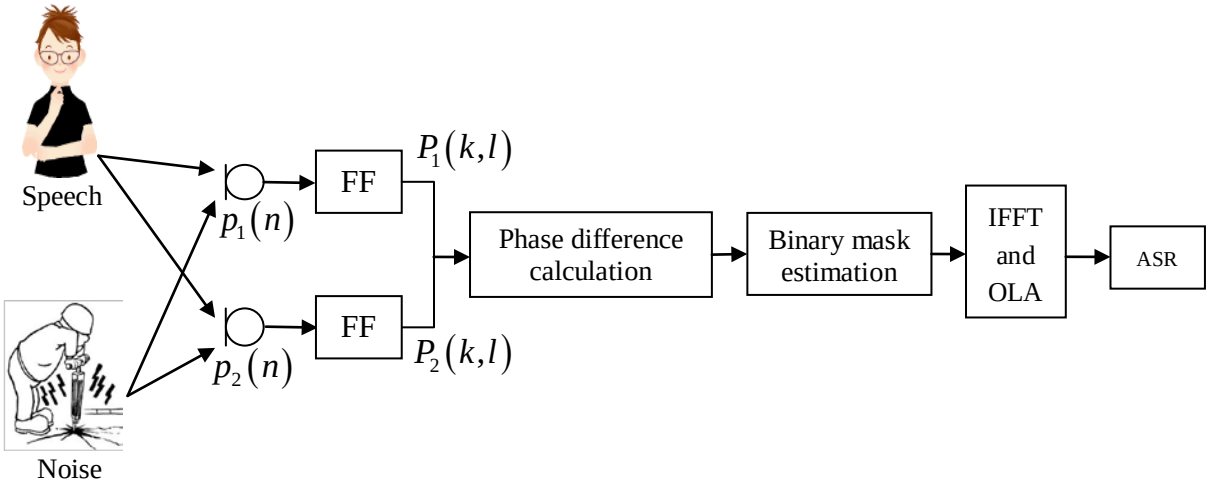


Fig. 1 The block diagram of phase-difference estimation.

The block diagram of PDE algorithm is illustrated in Fig. 1. The noisy signal received in two microphones is first segmented to frames by applying a moving Hamming window and then transferred to the time-frequency domain by Short-Time Fourier transform (STFT) as follows:

$$P_1(k,l) = X(k,l) + \sum_{i=1}^{V} N_i(k,l) \tag{1}$$

$$P_2(k,l) = X(k,l) + \sum_{i=0}^{V} e^{-j\omega_k d_i(k,l)} N_i(k,l) \tag{2}$$

where k is the frequency index and l is the frame index, $X(k,l)$ and $N_i(k,l)$ represent the speech and the ith noise signals, respectively, $P_1(k,l)$ and $P_2(k,l)$ are the signals at the first and second microphone, and $\omega_k = 2\pi k / N$ for $0 \le k \le N/2-1$, where N is the STFT size. The frame length here is 75ms and the hub size is half of frame length. It is assumed that the target signal is at the location alone the perpendicular bisector of the line between two microphones, and therefore its ITD is equal to zero. On the other hand, $d_i(k,l)$ is

the ITD of the *i*th noise signal dependent on time and frequency. If a time-frequency bin (k_m, l_m) is controlled by a strongest interference source *n*, the above equations can be approximated as

$$P_1(k_m, l_m) \approx N_n(k_m, l_m) \tag{3}$$

$$P_2(k_m, l_m) \approx e^{-j\omega_{k_m} d_n(k_m, l_m)} N_n(k_m, l_m) \tag{4}$$

and the ITD of this bin can be estimated by calculating the unwrapped phase difference between two microphones:

$$\left| d_n(k_m, l_m) \right| \approx \frac{1}{\left| \omega_{k_m} \right|} \min_r \left| \angle P_1(k_m, l_m) - \angle P_2(k_m, l_m) - 2\pi r \right| \tag{5}$$

Then, a binary mask can be formulated as

$$B(k_m, l_m) = \begin{cases} 1, & \text{if } \left| d_n(k_m, l_m) \right| \le \tau \\ 0.01, & \text{otherwise} \end{cases} \tag{6}$$

where τ is the ITD threshold. It means that only bins with its ITD smaller than τ are supposed to belong to the target signal. Correspondingly, the speech signal $S(k,l)$ is re-established from multiplying the average signals of the two microphones $\overline{P}(k,l)$ by the mask $B(k_j, l_j)$ got in above formula.

$$\overline{P}(k,l) = \frac{1}{2}\left\{ P_1(k,l) + P_2(k,l) \right\} \tag{7}$$

$$S(k,l) = B(k,l)\overline{P}(k,l) \tag{8}$$

Finally, the enhanced speech signal is converted to the time-domain with the aid of inverse fast Fourier transform (IFFT) and overlap addition (OLA) method. In this paper, three approaches of technical refinement are exploited to enhance the aforementioned PDE algorithm. As shown in Fig. 2, after the received signal is transformed to the time-frequency domain, the system estimates the speech and noise

location. The subtending angle between speech and noise is used to select the corresponding optimal ITD threshold searched by GSS. If the target source is not from the designed direction, beam steering technique is applied to orient the main lobe to the target source location. After IFFT and OLA, the time domain signal is scaled to the optimal volume to further increase the WRR.

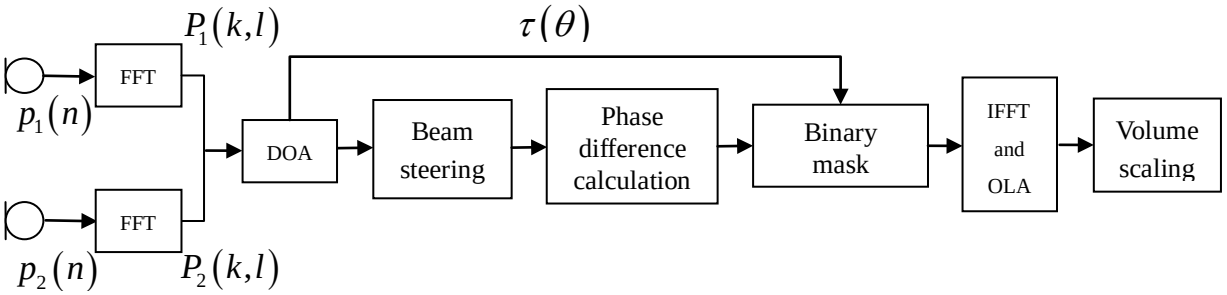


Fig. 2 The block diagram of the proposed PDE-based enhancement algorithm, where θ is the subtending angle estimated by DOA.

A. Optimization of the ITD threshold using GSS

As mentioned previously, the parameter τ is used in the binary mask principle as the ITD threshold having profound impact on mask estimation and hence on the performance of the speech recognition. As expected, it is found that this parameter is related to the included angle between speech and noise sources. Therefore, it is worth exploring how to adjust this parameter such that the recognition rate can be maximized. In the following, a procedure based on the GSS is presented for automated tuning of the ITD threshold.

1. Golden section search

The goals of GSS are to get an optimal reduction factor for a search interval and to minimize the number of the iterations. By GSS, the minimum can be searched efficiently within a finite number of steps, and do not need to evaluate numerical

gradients. Assume a function $f(x)$ is continuous and having only one minimum over the interval [*a*, *b*]. An interior point *c* is between *a* and *b*, and

$$\frac{c-a}{b-a}=w, \qquad \frac{b-c}{b-a}=1-w \tag{9}$$

where $0<w<\frac{1}{2}$. Suppose another interior point *d* is over [*c*, *b*], and

$$\frac{d-c}{b-a}=z \tag{10}$$

Notice that the choosing of *d* is applied the same strategy as that of *c*, which means

$$\frac{z}{1-w}=w \tag{11}$$

For minimizing the number of the iterations, the fraction 1-*w* musts equal to *w*+*z*, i.e. the new point *d* is the symmetric point of *c* in the interval[*a*, *b*], namely

$$z=1-2w \tag{12}$$

Comparison of Eqs. (11) and (12) yields the following quadratic equation

$$w^2-3w+1=0 \tag{13}$$

and the root

$$w=\frac{3-\sqrt{5}}{2}\approx 0.382 \tag{14}$$

is used. Note that the number is related to the golden ratio *g*, where

$$g=\frac{\sqrt{5}+1}{2}=\frac{1}{1-w} \tag{15}$$

Therefore it's called "golden section search". Now comparing $f(c)$ and $f(d)$, if $f(c)<f(d)$, then the new interval is [*a*, *d*]; otherwise, it becomes [*c*, *b*]. The rule at each stage is to keep a center point lower than the two outside points. The

process above iterates until the interval is tolerably small, and the question here is how to decide the time to stop the iteration. According to Taylor's theorem, the value of the function $f(x)$ near x_m is approximately

$$f(x) \approx f(x_m) + \frac{1}{2} f''(x_m)(x - x_m)^2 \tag{16}$$

If $f(x)$ is enough close to $f(x_m)$, then the second term can be quite small and negligible, which can be represented as

$$\frac{1}{2} f''(x_m)(x - x_m)^2 < \varepsilon \left| f(x_m) \right| \tag{17}$$

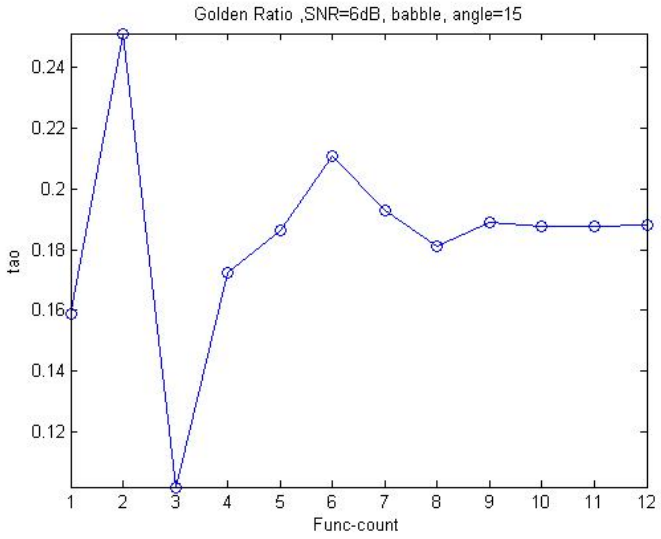
where εis usually set to 10^{-2} for single precision.

2. The Optimal ITD threshold varying with the included angle

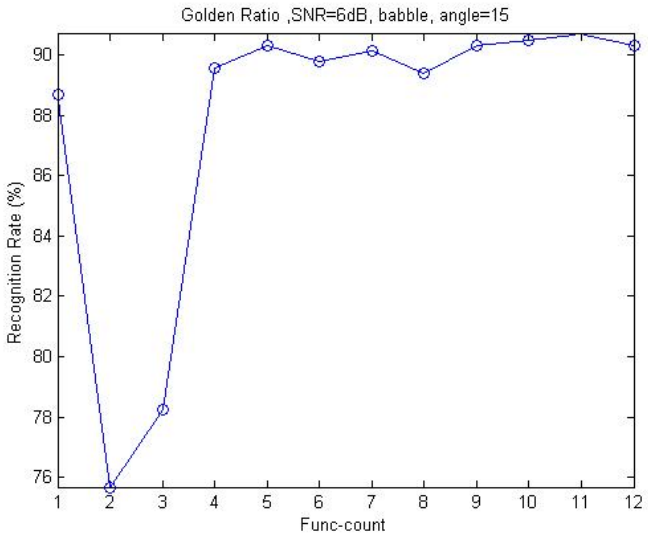
The searching process of the optimal ITD threshold by GSS is shown in Fig. 3, where the noise type is babble at SNR 6dB and the included angle is 15 degrees. The SNR here were conducted according to ITU P.56 standard, which defined as

$$SNR = 10 * \log_{10}\left(\frac{x^2}{n^2}\right) \tag{18}$$

where *x* and *n* represent the speech signal and noise respectively.



(a) The searching process of τ



(b) Relative recognition rate

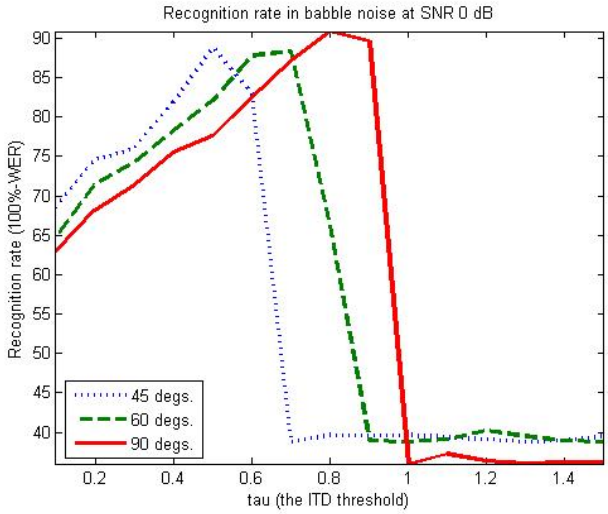
Fig. 3 The searching process of the ITD threshold by GSS.

Fig. 4(a) shows the performance of PD algorithm where the ITD threshold τ varies from 0.1 to 1.5. It can be found that the recognition rate gets better by increasing τ but decreases sharply whenτ exceeds a value which differs with the included angle. It turns out that there is a relation between τ and the included angle. To find the optimal ITD threshold, GSS is used in this paper, which can quickly search the local minimal of a function in an interval. The result of the optimal τ found by GSS is shown in Fig. 4(b). The included angle is from 15 degrees to 90 degrees at SNR 0dB and 6dB, and "babble" noise is used as the noise source. It indicates that the optimal thresholds τ at SNR 0dB and 6dB are similar to each other, which means the influence of SNR is small and can be disregard. Because the curve of the optimal τ has an obvious trend, it can be fitted by a polynomial of low degree easily. A polynomial fitting of degree 2 is shown in Fig. 4(b), which is found to be

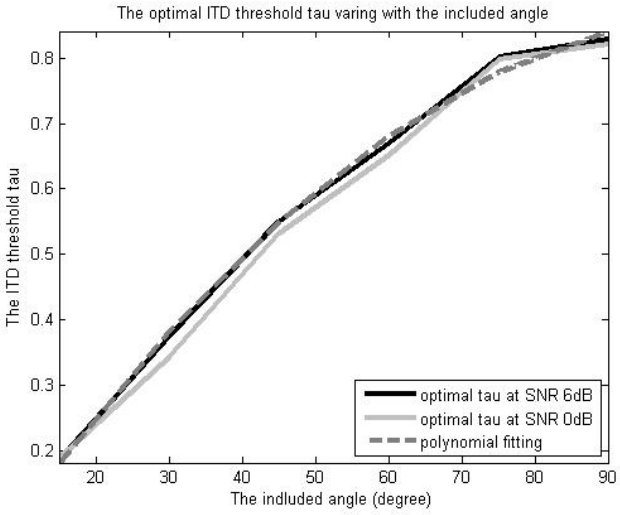
$$\tau(i) = \left(-7.76*10^{-5}\right)i^2 + \left(1.69*10^{-2}\right)i - \left(5.45*10^{-2}\right) \tag{19}$$

where *i* is the included angle. It revealed that, by using a polynomial fitting, it can use only 3 parameters to represent the optimal τ varying with the included angle very

well. By comparing the effective beamwidth and the real spanning angles, the effective beamwidth is smaller but the differences become smaller as the subtending angle decreases. The reason is that, the effective beamwidth has to be smaller than the real subtending angle, or the noise will be received in the binary mask, while if the effective beamwidth is too small, some speech signals will not be picked up in the binary mask and some feature will lose. For ASR, preservation of speech features is crucial. Loss of speech features causes the WRR to markedly decrease. Even if the noise is close to the target source, the effective beamwidth can not be too small.



(a)



(b)

Fig. 4 (a) Recognition rate in babble noise at SNR 0dB. (b) The optimal ITD threshold tau and the polynomial fitting.

B. Beam steering

The beam steering technique is discussed in this section to overcome the problem about the movement of the target source. With the aid of electronic compensation, the direction of the main lobe of the microphone array pattern can be changed. Assume the angle to be moved is θ_M, then the beam steering filters are given as

$$w_n = e^{-j\ n\ k\ d\sin\theta_M} = e^{\frac{-j\omega\ f_s\ nd\sin\theta_M}{c}} \tag{20}$$

where n is array index, ω is the frequency index, and f_s is the sampling rate, d is the spacing between microphones. In time domain, the beam steering filter can be written as a delay:

$$delay = \frac{f_s\ nd\sin\theta_M}{c} \tag{21}$$

That is, by applying different delays to the signal received in every microphone, the direction of main lobe can be controlled and steered to any desired directions. One thing has to be noticed is that these delays are not integer delays, hence Lagrange interpolation is used here to interpolate fractional delay values, which is easier to achieve and more flexible. Simplicity, it can approximate a fractional delay by a FIR filter,

$$h(n) = \prod_{\substack{k=0 \\ k\neq n}}^{N} \frac{D-k}{n-k} \quad \text{for } n = 0,1,2...N \tag{22}$$

where N is the order of the filter. The case N=1 corresponds to linear interpolation between two samples, which suffices when the sampling frequency is high enough. The result is in Fig. 5 with the target source angle from 15° to 75° aside the main lobe. When the target source is far from the main lobe, the recognition rate degrades

correspondingly. By using beam steering technique, the performance is enhanced obviously, as shown in Fig. 6.

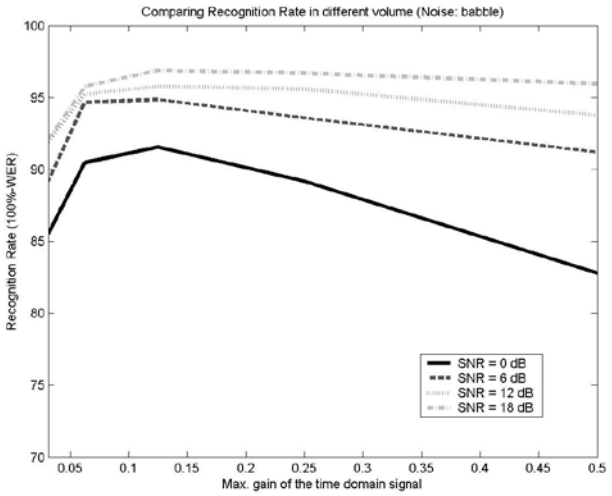


Fig. 5 Comparing recognition rate in different volume.

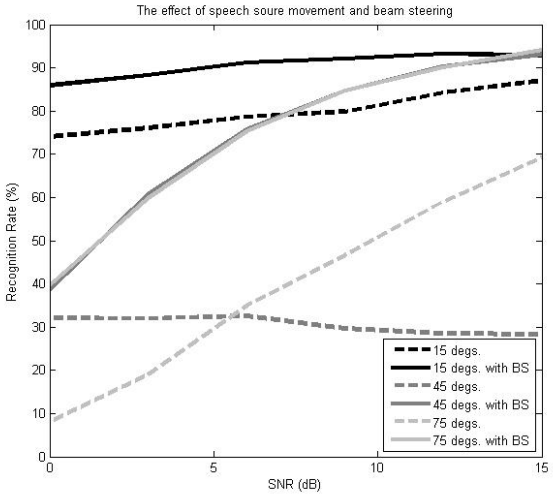


Fig. 6 Comparing the recognition rate when the source is not at the direction of the designed mainlobe and the effect of beam steering, where “15degs.” means the source is aside the desired main axis 15 degrees.

C. Simulation and Experiment

The simulated and experimental results are presented in this section. The input stimuli are 50 commends (547 wave files) rendered from a point source placed at 90 degrees (the look direction). The speech recognizer is based on continuous density Hidden Markov Model (HMM) with Mel-Frequency Cepstral Coefficients (MFCCs) as features. As shown in Fig. 7, the interelement spacing is 5 cm and the sampling rate is 8 KHz, the distance between microphone array and the speakers is 30 cm.

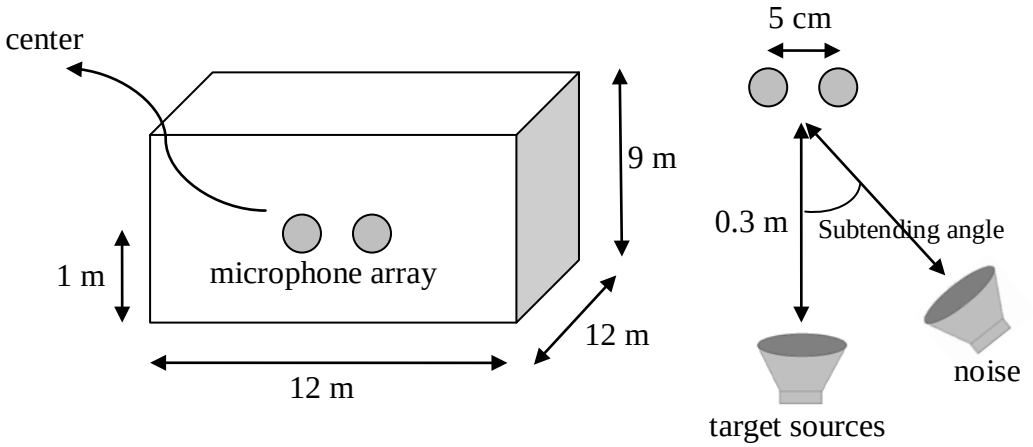
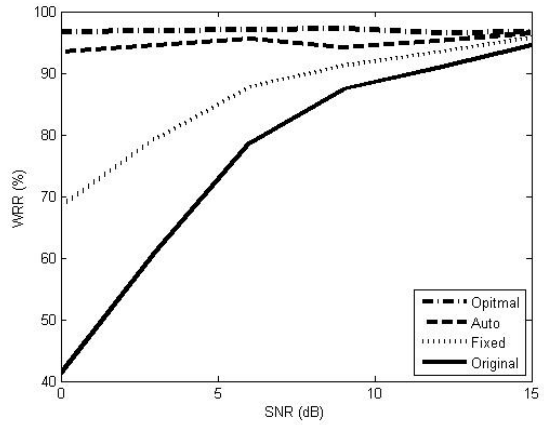
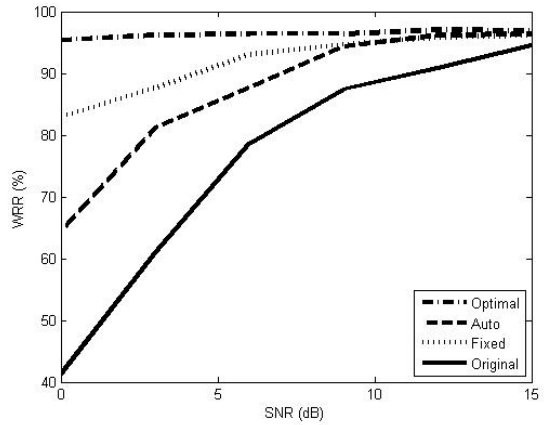


Fig. 7 The simulated and experimental environments.

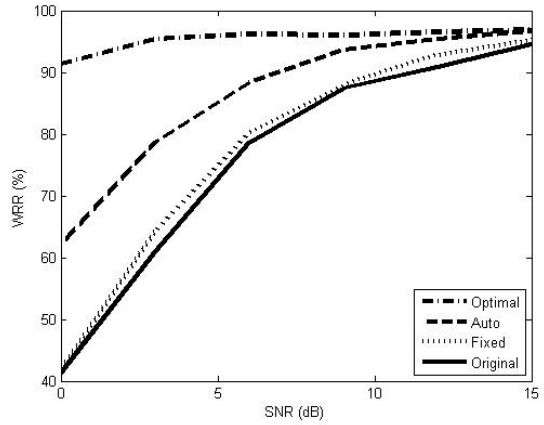
Assume a room of dimensions $12 \times 12 \times 9$ m, with the microphone located at the center of the room. The SNR is from 0 to 15dB and the subtending angle is from 15 to 90 degrees. Babble noise is used as the noise source. FIG. 8 compares the performance of the original noisy signal, PDE algorithm with fixed ITD threshold, automatic ITD threshold selection algorithm, and the proposed PDE-based enhancement algorithm. The subtending angle between target source and interference signal is 15°, 45°, and 75°, and there is no reberberation. The volume gain here is set to be 0.0945. The original noisy signal is the signal received in one microphone, and PDE algorithm with fixed ITD threshold is the result of the basic PDE system, where the ITD threshold is chose to be 0.4.



(a)



(b)



(c)

Fig. 8 Comparing the performance of the original noisy signal, PDE algorithm with fixed ITD threshold, automatic ITD threshold selection algorithm, and the

proposed PDE-based enhancement algorithm (a) Subtending angle = 75°. (b) Subtending angle = 45°. (c) Subtending angle = 15°.

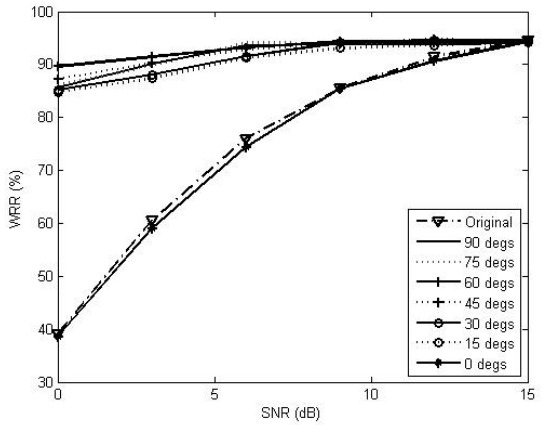
Automatic ITD threshold selection algorithm is organized as follows: First, two complementary masks are constructed using the binary threshold, one for the target signal, the other for interference signal. After that, the short-time power for the target and the interference is calculated. Finally, the ITD threshold is obtained by minimizing the cross-correlation of the target and interfering signals after a compressive nonlinearity, as shown below:

$$\hat{\tau}_0 = \arg\min_{\tau_0} \left| \frac{\frac{1}{N}\sum_{l=1}^{L} R_T[l \mid \tau_0) R_I[l \mid \tau_0) - \mu_{R_T}\mu_{R_I}}{\sigma_{R_T}\sigma_{R_I}} \right| \tag{23}$$

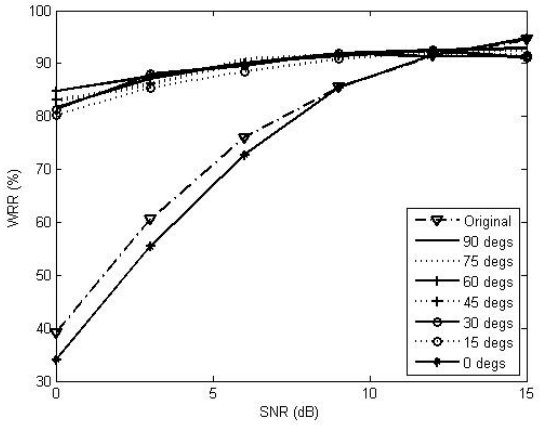
where $R_T[l \mid \tau_0)$ and $R_I[l \mid \tau_0)$ are the power of the target and the interference signals after nonlinearity, σ_{R_T} and σ_{R_I} are the standard deviations of $R_T[l \mid \tau_0)$ and $R_I[l \mid \tau_0)$, respectively, and μ_{R_T} and μ_{R_I} are the means of $R_T[l \mid \tau_0)$ and $R_I[l \mid \tau_0)$, respectively. From F4. 18 can find that, the proposed PDE-based enhancement algorithm gets excellent performance no matter what subtending angle it is, which enhances WRR about 50-60% at SNR 0dB and all the accuracies in different subtending angles are above 90% even if the noise is very close to the target source like 15 degrees, whereas the fixed-threshold PDE and the automatic-threshold selection algorithm degrade at low SNR. Furthermore, the automatic-threshold selection algorithm performs as well as the proposed algorithm when the subtending angle is large, like 75 degrees, but significantly degrades if the subtending angle is small and SNR is low.

The effect of reverberation presents in Fig. 9. The Room Impulse Response (RIR) software is used here to simulate reverberation effects. T60 represents the

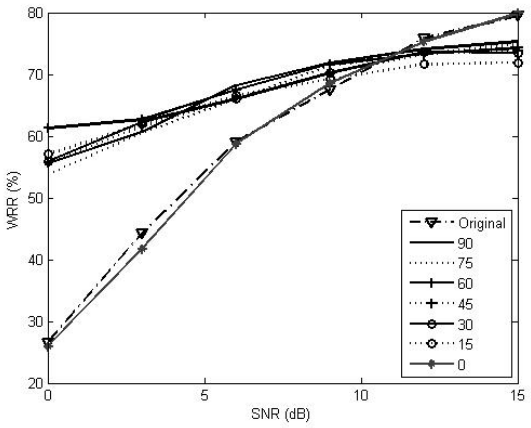
reverberation time, which is the time it takes for the reverberation level to drop by 60 dB. When the reverberation time T60 is small, the effect of reverberation is not obvious, and the performance after the proposed algorithm is almost above 85% at SNR 0dB. One thing to be noticed is that, PDE technique doesn't work if noise and speech come from the same direction, as shown in Fig. 9. It even gets worse WRR than the original signal when the reverberation time is long because of the distortion of speech signal. The performance decreases quickly when T60 is larger than 2 seconds. Even with the aid of the proposed PDE-based enhancement algorithm, WRR only increases to about 60% at SNR 0dB, and the result is worse than the original signal at high SNR because of the distortion of speech signal.



(a)



(b)



(c)

Fig. 9 The effect of reverberation, where the subtending angle is from 0 to 90 degrees. (a)T_{60}=0.138 secs. (b)T_{60}=0.966 secs. (c)T_{60}=2.898 secs.

FIG. 10 is recognition rate of record wave files. The recording is at an anchor chamber, and therefore the effect of reverberation can be neglected. SNR is 0dB in this case, and the noise source is babble noise. It indicates that, all WRR of original signals are low, between 10% and 30%, and after the proposed PDE-based enhancement algorithm, the performance is excellent even when SNR is low.

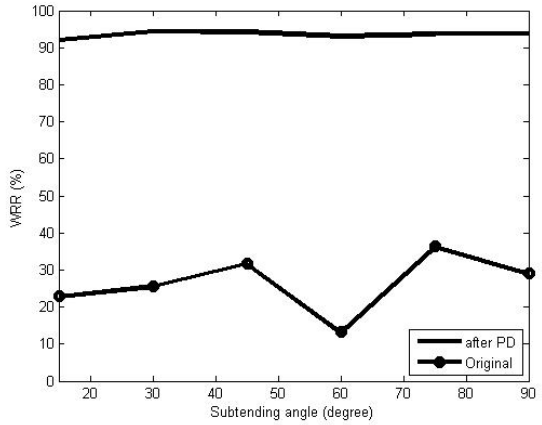


Fig. 10 The recognition rate with the optimal threshold of record wave file.

III. NONLINEAR ACOUSTIC ECHO CANCELLATION

The acoustic echo cancellers to date generally rely on the assumption of linear echo path. The nonlinear distortions from low-cost audio devices can adversely impact the performance of linear acoustic echo cancellation system. In this section, two methods are presented to deal with nonlinear echoes.

A. Second-order Volterra Filter

One of the sources of nonlinearity is the loudspeaker when overdriven beyond its linear region. When the loudspeaker is operated at the power limit, the nonlinear distortions will damage the linear echo cancellation. Since the time constants of their electro-mechanical system are large compared to the sampling rate, the loudspeaker causes nonlinearities with memory. For this nonlinear system with memory, the adaptive Volterra filters have been proposed to nonlinear echo cancellation system.

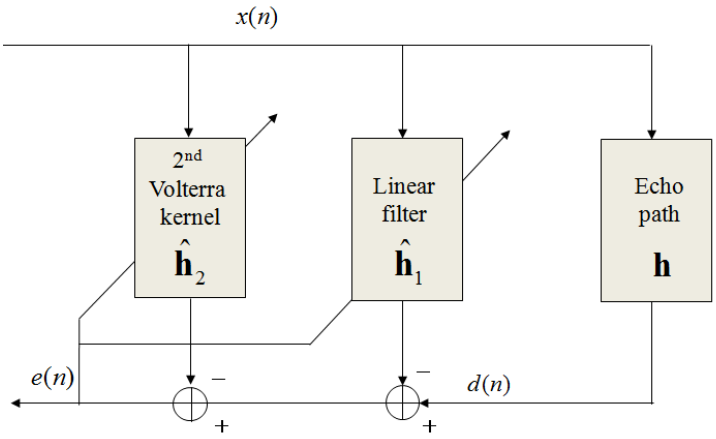


Fig. 11 Nonlinear acoustic echo canceller using 2nd order adaptive Volterra filter.

The Volterra filter, extension of the Taylor series, is a general type of nonlinear filters. Consider an *N*-th order discrete Volterra filter representation described as

$$y(n)=\sum_{r=1}^{N}\sum_{\kappa_1}^{M}\cdots\sum_{\kappa_r=\kappa_r-1}^{M}h_r(\kappa_1,\cdots,\kappa_r)\,x(n-\kappa_1)\cdots x(n-\kappa_r) \tag{24}$$

where M is the memory length and h_r are the r-th order Volterra kernels. $x(n)$ and $y(n)$ are the input and output signals, respectively. However, the numerical complexity of Volterra filter is too high for the practical acoustic echo cancellation system. The AEC using second order Volterra filter (Fig. 11) was proposed.[4] For the first order volterra kernel, the input vector is

$$\mathbf{x}_1(n) = \begin{bmatrix} x(n) & x(n-1) & \cdots & x(n-M+1) \end{bmatrix}^T \tag{25}$$

where M is memory length. The first-order filter can be represented as

$$\hat{\mathbf{h}}_1 = [\hat{h}_1(0)\ \hat{h}_1(1)\ \cdots\ \hat{h}_1(M-1)]^T \tag{26}$$

For the second-order volterra kernel, the input vector is

$$\begin{aligned} \mathbf{x}_2(n) = [&x^2(n)\ x(n)x(n-1)\ \cdots\ x(n)x(n-M+1) \\ &x(n-1)x(n-1)\ \cdots\ x(n-M+1)x(n-M+1)]^T \end{aligned} \tag{27}$$

and the second order filter is

$$\begin{aligned} \hat{\mathbf{h}}_2 = [&\hat{h}_2(0,0)\ \ \hat{h}_2(0,1)\ \cdots\ \hat{h}_2(0,M-1) \\ &\hat{h}_2(1,1)\ \cdots\ \hat{h}_2(M-1,M-1)]^T, \end{aligned} \tag{28}$$

The PNLMS adaptive Volterra filter can be formulated as

$$e(n) = d(n) - \hat{\mathbf{h}}_1\mathbf{x}_1^T(n) - \hat{\mathbf{h}}_2\mathbf{x}_2^T(n) \tag{29}$$

$$\hat{\mathbf{h}}_1(n+1) = \hat{\mathbf{h}}_1(n) + \frac{\alpha_1\mathbf{K}(n)e(n)\mathbf{x}_1(n)}{\mathbf{x}_1^T(n)\mathbf{K}(n)\mathbf{x}_1(n)} \tag{30}$$

$$\hat{\mathbf{h}}_2(n+1) = \hat{\mathbf{h}}_2(n) + \frac{\alpha_2\mathbf{K}(n)e(n)\mathbf{x}_2(n)}{\mathbf{x}_2^T(n)\mathbf{K}(n)\mathbf{x}_2(n)} \tag{31}$$

where α_1 and α_2 are the first and second kernel step size, respectively.

B. Hammerstein Model

The over driven amplifier mainly generates the memoryless nonlinear distortions. The nonlinear AEC dealing with memoryless nonlinearity is proposed in this section. The Hammerstein model consisting of linear FIR filter and the nonlinear function is illustrated in Fig. 12. The memoryless nonlinear function f is the polynomial model

and then can model saturation effects found in the amplifier. The PNLMS adaption for linear FIR filter based on the error signal $e(n)$ in Fig. 12 is used for two stages of Hammerstein model.

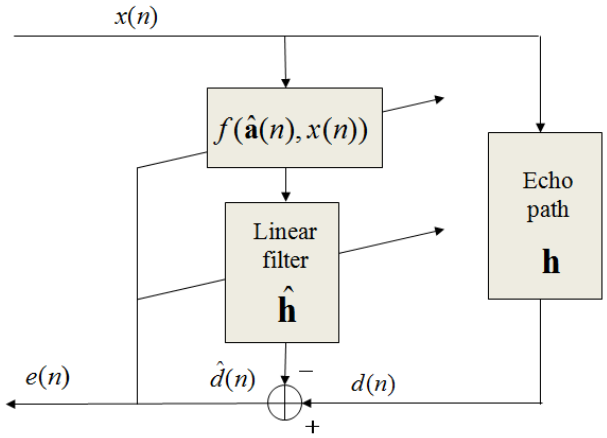


Fig. 12 Nonlinear acoustic echo canceller using Hammerstein model.

First, the function *f* can be represented by a *P*th-order polynomial, and the output of the nonlinear function is denoted by $\hat{s}(n)$

$$\hat{\mathbf{a}}(n) = [a(n)\ a_1(n)\ \cdots\ a_P(n)]^T \tag{32}$$

$$\hat{s}(n) = f(\hat{\mathbf{a}}(n), x(n)) = \sum_{p=1}^{P} \hat{a}_p(n) x^p(n) \tag{33}$$

where $\hat{\mathbf{a}}(n)$ is the $P \times 1$ column vector. The input vector $\hat{\mathbf{s}}(n)$ to the linear FIR filter $\hat{\mathbf{h}}(n) = [\hat{h}_0(n)\ \hat{h}_1(n)\ \cdots\ \hat{h}_{N-1}(n)]^T$ is formed by N latest values of nonlinear function output

$$\hat{\mathbf{s}}(n) = [\hat{s}(n)\ \hat{s}(n-1)\ \cdots\ \hat{s}(n-N+1)]^T \tag{34}$$

The error signal $e(n)$ is the difference between the linear FIR filter output and the microphone signal $d(n)$

$$e(n) = d(n) - \hat{\mathbf{h}}^T \mathbf{f}(\hat{\mathbf{a}}(n), \mathbf{x}(n)) = d(n) - \hat{\mathbf{h}}^T(n)\hat{\mathbf{s}}(n) \tag{35}$$

$$\mathbf{f}(\hat{\mathbf{a}}(n), \mathbf{x}(n)) = [f(\hat{\mathbf{a}}(n), x(n))\ f(\hat{\mathbf{a}}(n), x(n-1)\ \cdots\ f(\hat{\mathbf{a}}(n), x(n-N+1))]^T \tag{36}$$

The update equation of $\hat{\mathbf{h}}(n)$ is based on PNLMS algorithm

$$\hat{\mathbf{h}}(n+1) = \hat{\mathbf{h}}(n) + \frac{\alpha_1 \mathbf{K}(n) e(n) \hat{\mathbf{s}}(n)}{\hat{\mathbf{s}}^T(n) \mathbf{K}(n) \hat{\mathbf{s}}(n)} \tag{37}$$

In an LMS-type adaptation for adaptive filter is derived by forming the gradient of squared error simple with respect to the adaptive coefficients. From the Eq. (35) and Eq. (36),

$$\frac{\partial e^2(n)}{\partial \hat{\mathbf{a}}(n)} = -2e(n)\mathbf{f}'(\hat{\mathbf{a}}(n), \mathbf{x}(n))^T \hat{\mathbf{h}}(n) \tag{38}$$

and

$$\mathbf{f}(\hat{\mathbf{a}}(n), \mathbf{x}(n)) = [\mathbf{x}_p(n)\ \mathbf{x}_p(n-1) \cdots\ \mathbf{x}_p(n-N+1)]^T \hat{\mathbf{a}}(n) = \mathbf{X}_P(n)\hat{\mathbf{a}}(n) \tag{39}$$

$$\mathbf{x}_p(n) = [x(n)\ x^2(n)\ \cdots\ x^P(n)]^T \tag{40}$$

As a result, the Eq. (34) becomes

$$\frac{\partial e^2(n)}{\partial \hat{\mathbf{a}}(n)} = -2e(n)\mathbf{X}_P(n)^T \hat{\mathbf{h}}(n) \tag{41}$$

the update equation of $\hat{\mathbf{a}}(n)$ based on NLMS algorithm is

$$\hat{\mathbf{a}}(n+1) = \hat{\mathbf{a}}(n) + \frac{\alpha_a}{\| \mathbf{X}_P(n)^T \hat{\mathbf{h}}(n) \|_2^2} \mathbf{X}_P(n)^T \hat{\mathbf{h}}(n) e(n) \tag{42}$$

C. Modified Nonlinear Adaptive Algorithms

The divergence behavior has been observed in both nonlinear AEC systems. Although the Volterra kernels are adapted separately, the error introduced by a misadjusted linear kernel acts as a distortion for the adaption of the quadratic kernel. Therefore, the adaptive Volterra filter system is modified as shown in Fig. 13. The main idea of this system is to choose the smaller error signal for the linear kernel adaption. First, we define

$$e_1(n) = d(n) - \hat{\mathbf{h}}_1 \mathbf{x}_1^T(n), \quad \overline{e}_1^2(n) = \lambda \overline{e}_1^2(n-1) + (1-\lambda) e_1^2 \tag{43}$$

$$e(n) = d(n) - \hat{\mathbf{h}}_1 \mathbf{x}_1^T(n) - \hat{\mathbf{h}}_2 \mathbf{x}_2^T(n), \quad \overline{e}^2(n) = \lambda \overline{e}^2(n-1) + (1-\lambda) e^2 \tag{44}$$

where the forgetting factor λ is chosen 0.1 . The error signal

$$\hat{e}_1(n) = \begin{cases} e_1(n), \text{ if } \overline{e}_1^2(n) < \overline{e}^2(n) \\ e(n), \text{ if } \overline{e}_1^2(n) \geq \overline{e}^2(n) \end{cases} \tag{45}$$

is used for the adaption of the linear kernel in Eq. (31). The error signal $\hat{e}_1(n)$ also represents the residual echo signal. The unstable behavior also occurs due to misadjustment on the preprocessor and linear FIR filter. Consequently, we apply the preceding comparison idea to the Hammerstein model to avoid divergence, as shown in Fig. 14. Another linear FIR filter $\hat{\mathbf{w}}(n) = [\hat{w}_0(n), \hat{w}_1(n), \cdots, \hat{w}_{N-1}(n)]^T$ is parallel to the cascade connection of nonlinear function *f* and linear FIR filter $\hat{\mathbf{h}}(n)$. We define

$$e_w(n) = d(n) - \hat{\mathbf{w}}\mathbf{x}^T(n), \quad \overline{e}_w^2(n) = \lambda \overline{e}_w^2(n-1) + (1-\lambda)e_w^2 \tag{46}$$

$$e(n) = d(n) - \hat{\mathbf{h}}^T(n)\hat{\mathbf{s}}(n), \quad \overline{e}^2(n) = \lambda \overline{e}^2(n-1) + (1-\lambda)e^2 \tag{47}$$

where the forgetting factor λ is chosen to be 0.1 . The error signal

$$\hat{e}_w(n) = \begin{cases} e_w(n), \text{ if } \overline{e}_w^2(n) < \overline{e}^2(n) \\ e(n), \text{ if } \overline{e}_w^2(n) \geq \overline{e}^2(n) \end{cases} \tag{48}$$

is used for the adaption of the linear kernel in Eq. (34). The error signal $\hat{e}_w(n)$ also represents the residual echo signal.

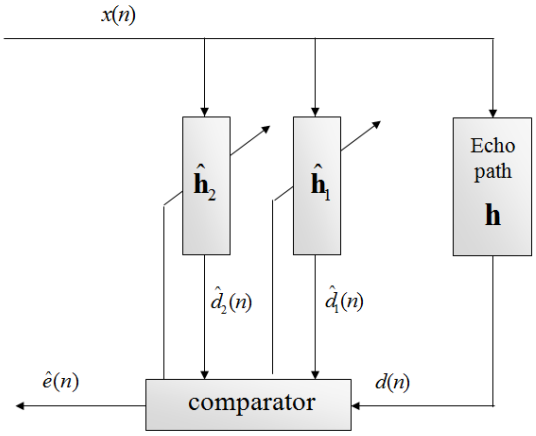


Fig. 13 Modified Volterra filter system.

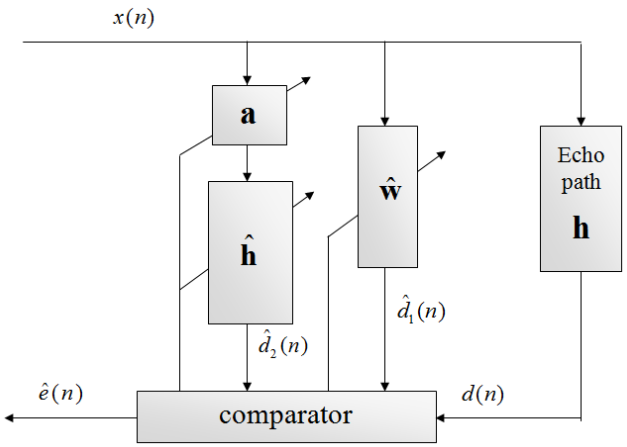


Fig. 14 Modified Volterra filter system.

D. Performance evaluation of nonlinear echo cancellation

In the aspect of experiment about nonlinear echo cancellation, in order to create nonlinearity into AEC system, we turn the level of loudspeaker larger. In the following experiments, we used female speech signal and CSS as the echo signals. The experimental configuration is the same as the previous settings for the linear AECs. The sample rate here is also 16 kHz. First we evaluated the nonlinear AEC performance using modified Volterra filter method. For the speech input signal, the ERLE obtained using linear PNLMS algorithm denoted by the solid line in Fig. 15 reaches approximately 18 dB. The ERLE of the modified Volterra filter (dotted line) is further increased by 3 dB with nonlinear processing. For the CSS input, the ERLE can be increased by 5 dB via the modified Volterra filter method. Next, by the same protocol, we examine the nonlinear AEC with the modified Hammerstein model. For the speech and the CSS as the input signals, the ERLE can be increased by 3 dB and 5 dB with nonlinear processing (Fig. 16).

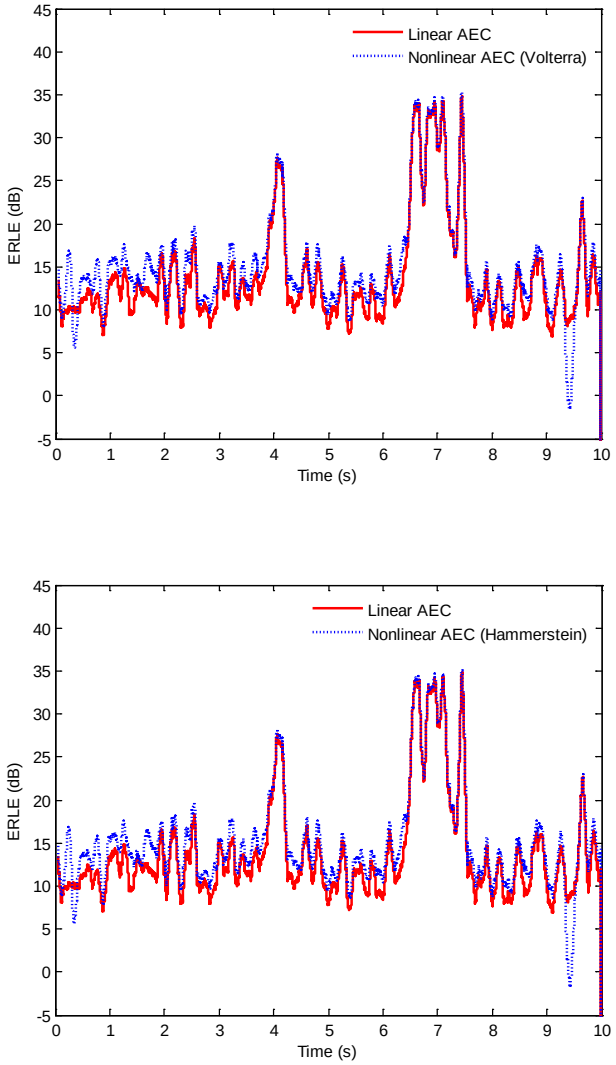


Fig. 15 The ERLE for a recorded echo for the speech input signal.

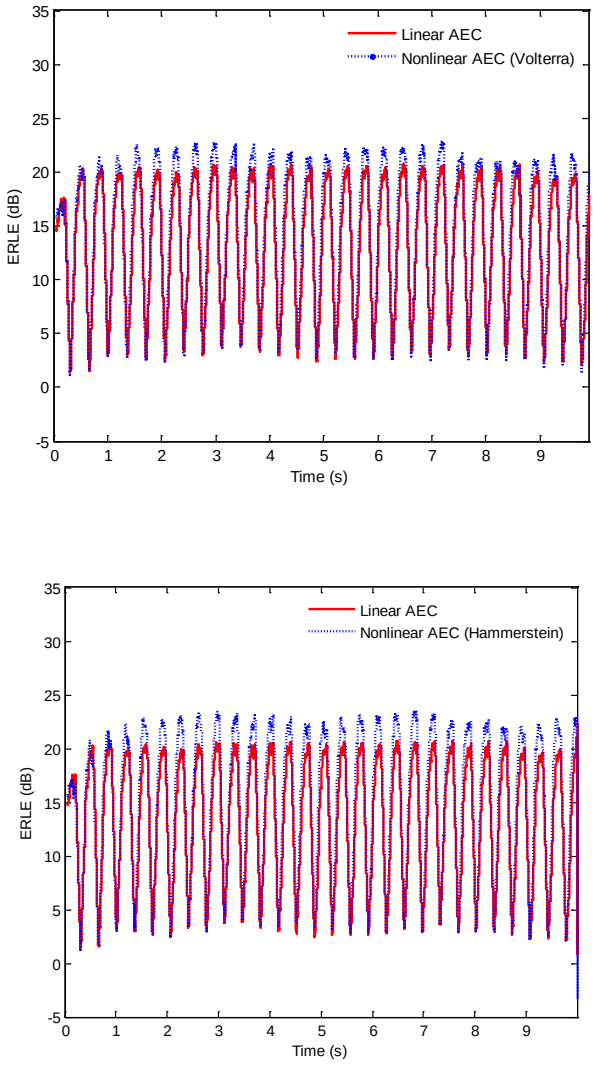


Fig. 16 The ERLE for a recorded echo for the CSS input signal.

IV. CONCLUSION

When the noise signal is close to the speech, PD is proposed to solve this problem. Using GSS to find the optimal ITD threshold differing with the included angle and the optimal volume can further improve the speech recognition. Finally, simulated and experimental results are discussed to prove effective in enhancement of speech recognition. The performance of linear acoustic echo cancellers is limited by nonlinear components in the echo path. We used a nonlinear AEC to deal with nonlinear echoes. Experiment results showed that the proposed nonlinear AEC

provided increased echo attenuation, as compared to a linear AEC applied to a nonlinear echo path. For recorded echoes, the ERLE can reach approximately 3 dB for the female speech signal and 5 dB for the CSS. In the modified nonlinear AEC, the Volterra filter and Hammerstein model are effective in dealing with nonlinear echoes with ensured convergence.

REFERENCES

[1] Y. Gong, "Speech recognition in noisy environments: a survey," *Speech Commun.*, vol. 16, pp. 261-291, 1995.

[2] M. R. Bai and P. J. Hsieh, "Optimal design of minimum mean-square error noise reduction algorithms using the simulated annealing technique," *J. Acoust. Soc. Am.*, vol. 125, no. 2, pp. 934-943, 2009.

[3] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Commun.*, vol. 34, pp. 267-285, 2001.

[4] S. Srinivasan, N. Roman, and D. L. Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Commun.*, vol. 48, pp. 1486-1501, 2006.

[5] R. M. Stern, E. Gouvea, C. Kim, K. Kumar, and H. Park, "Binaural and multiple-microphone signal processing motivated by auditory perception," HSCMA-2008, Trento, Italy, pp. 98–103, May. 2008.

[6] R. M. Stern and C. Trahiotis, "Models of binaural interaction," Hearing, B. C. J. Moore, ed., Academic Press, pp. 347–386, 2002.

[7] H. Park and R. M. Stern, “Spatial separation of speech signals using amplitude estimation based on interaural comparisons of zero crossings,” *Speech Commun.*, vol. 51, no. 1, pp. 15–25, 2009.

[8] K. J. Palomaki, G. J. Brown, and D. L. Wang, “A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation,” *Speech Commun.*, vol. 43, pp. 361-378, 2004.

[9] N. Roman, D. L. Wang, and G. J. Brown, “Speech segregation based on sound localization,” *J. Acoust. Soc. Am.*, vol. **114**, pp. 2236-2252, 2003.

[10] C. Kim, K. Kumar, B. Raj, and R. M. Stern, “Signal separation for robust speech recognition based on phase difference information obtained in the frequency domain,” INTERSPEECH-2009, Brighton, UK, pp. 2495–2498, Sept. 2009.

[11] C. Kim, R. M. Stern, K. Eom, and J. Lee, “Automatic selection of thresholds for signal separation algorithm based on interaural delay,” INTERSPEECH-2010, Makuhari, Japan, Sept. 2010).

[12] J. Bergqvist and F. Rudolf, “A silicon condenser microphone using bond and etch-back technology,” *Sensors and Actuators A*, vol. 45, pp. 115-124, 1994.

[13] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical recipes in C: the art of scientific computing*, 2nd ed, Cambridge University Press: New York, 1992, pp. 397-402.

[14] ITU-T P. 56, “Objective measurement of active speech level,” ITU-T Recommendation, 1993.

[15] J. G. Wilpon, L. R. Rabiner, C. H. Lee, and E. R. Goldmn, “Automatic recognition of keyword in unconstrained speech using hidden Markov models,” *IEEE Trans. Acoustics, Speech, Signal Proc.*, vol. 38, no.11, pp. 1870-1878, 1990.

[16] H. Ney, "The Use of a one-stage dynamic programming algorithm for connected word recognition," *IEEE Trans. Acoustics, Speech, Signal Proc.*, vol. 32, no. 2, 263-271, 1984

[17] S. G. McGovern, "A model for room acoustics," http://2pi.us/rir.html ,last viewed, 8/3/2010.

[18] A. Stenger, L. Trautmann and R. Rabenstein, " Nonlinear acoustic echo cancellation with 2nd order adaptive Volterra filters" *IEEE International Conference on Acoust. Speech, Signal Processing*, Phoenix, Vol. 2, pp.877-880 (1999).

[19] A. Stenger and W. Kellermann, "Adaptation of a memoryless preprocessor for nonlinear acoustic echo cancelling." *Signal Processing*, Vol. 80, No. 9, pp. 1747-1760(2000).

[20] J. Benesty and S. L. Gay, "An improved PNLMS algorithm," *IEEE International Conference on Acoust. Speech, Signal Processing*, Orlando, pp.1881-1884(2002).