Hindawi Publishing Corporation
EURASIP Journal on Applied Signal Processing
Volume 2006, Article ID 61214, Pages 1–10
DOI 10.1155/ASP/2006/61214

Speech Enhancement by Multichannel Crosstalk Resistant ANC and Improved Spectrum Subtraction

Qingning Zeng and Waleed H. Abdulla

Department of Electrical and Computer Engineering, The University of Auckland, Private Bag 92019, Auckland, New Zealand

Received 31 December 2005; Revised 3 August 2006; Accepted 13 August 2006

A scheme combining multichannel crosstalk resistant adaptive noise cancellation (MCRANC) algorithm and improved spectrum subtraction (ISS) algorithm is presented to enhance noise carrying speech signals. The scheme would permit locating the microphones in close proximity by virtue of using MCRANC which has the capability of removing the crosstalk effect. MCRANC would also permit canceling out nonstationary noise and making the residual noise more stationary for further treatment by ISS algorithm. Experimental results have indicated that this scheme outperforms many commonly used techniques in the sense of SNR improvement and music effect reduction which is an inevitable byproduct of the spectrum subtraction algorithm.

1. INTRODUCTION

Many speech enhancement algorithms have been developed in the previous years as speech enhancement is a core target in many demanding areas such as telecommunications, and speech and speaker recognitions. Among them, spectrum subtraction (SS) [1–3] and adaptive noise cancellation (ANC) [4] are the most practical and effective algorithms.

SS algorithm needs only one channel signal and can be easily implemented with the existing digital hardware. It has been embedded in some high-quality mobile phones. Nevertheless, SS is only appropriate for stationary noise environments. Furthermore, it inevitably introduces "music noise" problem. In fact, the higher the noise is suppressed, the greater the distortion is brought to the speech signal and accordingly the poorer the intelligibility of the enhanced speech is obtained. As a result, ideal enhancement can hardly be achieved when SNR of the noisy speech is relatively low; below 5 dB. In contrast, it has quite good result when SNR of the noisy speech is relatively high; above 15 dB.

On the other hand, ANC algorithm can be used to enhance speech signals in many noisy environments situations. However, it requires two channels to acquire signals for processing; the main channel and the referential channel. In addition, the referential channel signal should contain only noise signal. This implies that the referential microphone should be somewhat far from the main microphone. It has been proven that because of the propagation complexity of the audio signal in the practical environment, the farther the referential microphone from the main microphone, the smaller the correlation of the referential signal with the main signal and accordingly less noise could be cancelled. Thus, the enhancement effect of ANC algorithm is in fact also quite limited. Fortunately, multichannel version of ANC algorithm can increase the cancellation effect since two or more referential signals implicate greater correlation with the main signal [5–7].

Multichannel ANC (MANC) employs more than one referential sensor in addition to the main sensor and thus generally makes the sensor array quite big. But in many applications such as in mobile and hands-free phones, microphone array of the speech enhancement system is expected to be small in size [8, 9]. This implies that the distances between any two of the employed microphones must be very small. On the other hand, sensors such as microphones located in close proximity undergo serious crosstalk effect. This effect violates the operating condition of MANC algorithm [5, 10] because the referential signals in MANC must not contain any speech signal. Otherwise, the speech signal is simultaneously cancelled with the noises.

Various two-channel crosstalk resistant ANC (CRANC) methods have been well introduced in the literature [11–16]. They make use of the principal of adaptive noise cancellation but permit the main channel sensor and the referential channel sensor to be closely located. However, some of these methods are unstable and some are computationally expensive. Among them, the algorithms of [12, 15] are quite stable. Both of them deal with biomedical signal extraction and the algorithm of [15] is obviously the simplified version of [12].

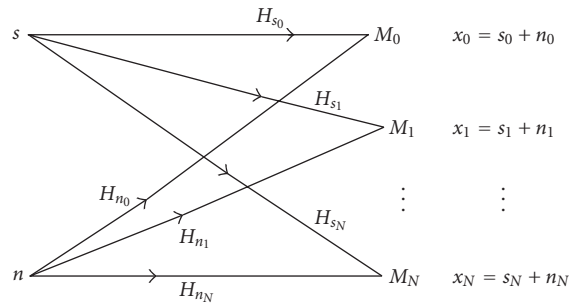

Figure 1: Speech and noise propagations between the emitting sources and the acquiring microphones.

In this paper we further simplify the algorithm in [15] and extend it to multichannel signals. The extended algorithm is named as multichannel crosstalk resistant ANC (MCRANC). Then MCRANC is augmented with an improved SS (ISS) algorithm to further improve the enhanced speech. The proposed MCRANC has the advantages of MANC and CRANC. It increases the noise cancellation performance as well as permits locating the microphones in close proximity. As the SNR of the enhanced speech by MCRANC has increased and the residual noise becomes more stationary, the augmented ISS algorithm will definitely have better performance. Experiments showed that the proposed scheme has made the speech enhancement system more efficient in suppressing noise, and small in size while preserving the speech quality. In addition, as ISS is easy to implement, and the present MCRANC employs only two adaptive FIR filters and a simple voice detector (VD), the proposed scheme in this paper can be realized in real time with the common DSP chips.

2. SIGNAL PROPAGATION MODELING

Assume $N+1$ microphones are used and closely placed. These microphones form an array. The array layout might be in any structure; such as uniform linear array, planar array, or solid array. We have no strict limitations on the physical layout of the microphones.

Suppose a digital speech signal $s(k)$ and noise $n(k)$ are generated by independent sources, as indicated in Figure 1. These signals arrive at microphone M_i through multipaths and are acquired as $s_i(k)$ and $n_i(k)$. The impulse responses of the intermediate media between the speech and noise sources and the acquiring microphone M_i are $h_{si}(k)$ and $h_{ni}(k)$, respectively. The audio signal acquired by microphone M_i can be represented by $x_i(k) = s_i(k) + n_i(k)$, where $i = 0, 1, \ldots, N$; $N + 1$ is the number of microphones employed; k is the discrete time index. Since the acquired signals by the microphones contain noise and speech concurrently, crosstalk between noise and speech happens [12, 16].

Let us consider $x_0(k)$ as the main channel signal acquired by microphone M_0, and $x_i(k)$ $(i = 1, \ldots, N)$ as the referential signals acquired by the other N microphones. Assume that the main channel signal is correlated with the referential channel signals, which is a valid assumption as the microphones are located in close proximity. Since the referential signals contain both speech and noise, common adaptive noise cancellation (ANC) and multichannel ANC (MANC) methods will not be appropriate methods for speech enhancement. That is because crosstalk effect violates their working conditions and consequently both speech and noise will be cancelled out.

From Figure 1, we have

$$x_i(k) = s_i(k) + n_i(k), \tag{1}$$

$$s_i(k) = h_{si}(k) * s(k), \tag{2}$$

where $*$ is the convolution sign, $h_{si}(k)$ and $h_{ni}(k)$ is the time domain impulse response correspondence of the z-domain response $H_{si}(z)$ and $H_{ni}(z)$.

Let the impulse response of the intermediate environment between the input signal s_i and the output signal s_j be $h_{s_j s_i}(k)$, then

$$s_j(k) = h_{s_j s_i}(k) * s_i(k), \quad i, j = 0, 1, \ldots, N. \tag{3}$$

Through (2)-(3),

$$H_{s_j s_i}(z) = \frac{H_{sj}(z)}{H_{si}(z)}, \quad i, j = 0, 1, \ldots, N. \tag{4}$$

In the practical environment, noise emitted from a certain source may propagate to microphone M_i through multiple paths including direct propagations, reflections, and refractions. The noise may also be emitted from multiple sources. We consider that those noises are from a combined source and all propagation paths are included in the combined transfer function $H_{ni}(z)$, which has an impulse response $h_{ni}(k)$.

3. PROPOSED SCHEME

As shown in Figure 2, the proposed scheme of the speech enhancement system is MCRANC cascaded with ISS. Its subsystem on the left of the dotted line indicates the diagram of MCRANC algorithm while that on the right is the ISS subsystem. Both subsystems employ a voice detector (VD) [17] to adapt the system, which will be described after MCRANC is introduced and ISS is summarized.

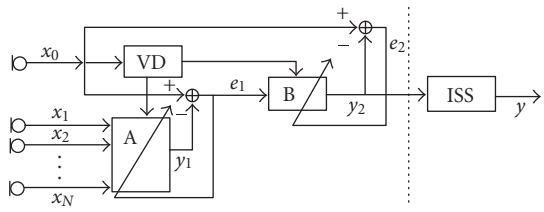

Figure 2: MCRANC-based speech enhancement system.

3.1. MCRANC formulation

MCRANC-based system consists of a VD module and two FIR filters A and B. During nonvoice periods (NVPs), where the noise dominates, the referential signals are used to cancel out the main signal through filter A. In this case, as $s_0(k) = 0$ in the main channel and $s_i(k) = 0$ $(i = 1,\ldots,N)$ in the referential channels, we have

$$\begin{aligned} x_0(k) &= y_1(k) + e_1(k), \\ n_0(k) &= \overline{w}\,\overline{n}(k) + \mathrm{err}(k), \end{aligned} \tag{5}$$

where $e_1(k) = \mathrm{err}(k)$ is the prediction error, $\overline{w}$ is the weight vector of the FIR filter A, that is,

$$\overline{w} = (\overline{w}_1, \overline{w}_2, \ldots, \overline{w}_N), \tag{6}$$

where $\overline{w}_i = (w_{i0}, w_{i1}, \ldots, w_{iL})$, $\overline{n}(k)$ is the vector of noise signal,

$$\overline{n}(k) = [\overline{n}_1(k), \overline{n}_2(k), \ldots, \overline{n}_N(k)]^T, \tag{7}$$

where $\overline{n}_i(k) = [n_i(k), n_i(k-1), \ldots, n_i(k-L)]^T$, and L is the number of delay units in the FIR filter of each referential channel.

Let the minimal prediction error power be denoted by $P[\mathrm{err}^0(k)]$ and the corresponding optimal weight vector by

$$\begin{aligned} \overline{w}^0 &= \left(\overline{w}_1^0, \overline{w}_2^0, \ldots, \overline{w}_N^0\right) \\ &= \left(w_{10}^0, w_{11}^0, \ldots, w_{1L}^0, w_{20}^0, w_{21}^0, \ldots, \right. \\ &\qquad \left. w_{2L}^0, \ldots, w_{N0}^0, w_{N1}^0, \ldots, w_{NL}^0\right). \end{aligned} \tag{8}$$

We need only to adjust the weights of filter A to minimize the square sum of $e_1(k)$ in Figure 2 to obtain $\overline{w}^0$. Theoretically $P[\mathrm{err}^0(k)]$ is inversely proportional to the number of the referential channels used.

In our approach, it has been assumed that the environment is changing slowly or it is pseudostationary. Accordingly, during the voice period (VP) which is the time interval from the end of the current NVP to the beginning of next NVP, we may keep the optimized weights $\overline{w}^0$ of filter A unchanged. Thus the output of filter A in this VP period is represented by

$$\begin{aligned} y_1(k) &= \overline{w}^0\,\overline{x}(k) \\ &= \overline{w}^0[\overline{s}(k) + \overline{n}(k)] \\ &= \overline{w}^0\,\overline{s}(k) + [n_0(k) - \mathrm{err}^0(k)], \end{aligned} \tag{9}$$

where $\overline{x}(k)$ and $\overline{s}(k)$ represent the acquired speech plus noise and the pure speech vectors, respectively. It may be expressed in a similar way to $\overline{n}(k)$ in (7). Then from (1) and (9),

$$\begin{aligned} e_1(k) &= x_0(k) - y_1(k) \\ &= [s_0(k) + n_0(k)] - [\overline{w}^0\,\overline{s}(k) + n_0(k) - \mathrm{err}^0(k)] \\ &= s_0(k) - \overline{w}^0\,\overline{s}(k) + \mathrm{err}^0(k) \\ &= p(k) + \mathrm{err}^0(k), \end{aligned} \tag{10}$$

where

$$p(k) = s_0(k) - \overline{w}^0\,\overline{s}(k). \tag{11}$$

Obviously $p(k)$ is the distorted signal of the speech $s_0(k)$. If the main microphone is reasonably separated from the referential microphones, the distortion will not be serious and thus $e_1(k)$ could be used as the enhanced speech in some applications. But if the microphones are very closely placed or the distortion is unacceptable for some applications, we can recover the clean signal using the following way.

Take the z-transform of (10) and (11) to get

$$\begin{aligned} E_1(z) &= P(z) + \mathrm{err}^0(z), \\ P(z) &= S_0(z) - Z\left[\sum_{i=1}^{N}\sum_{j=0}^{L} w_{ij}^0 s_i(k-j)\right] \\ &= S_0(z) - Z\left[\sum_{i=1}^{N}\sum_{j=0}^{L} w_{ij}^0 h_{s_i s_0}(k-j) * s_0(k-j)\right] \\ &= S_0(z) - \sum_{i=1}^{N}\sum_{j=0}^{L} w_{ij}^0 Z[h_{s_i s_0}(k-j)]Z[s_0(k-j)] \\ &= \left[1 - \sum_{i=1}^{N}\sum_{j=0}^{L} w_{ij}^0 z^{-2j} H_{s_i s_0}(z)\right] S_0(z) \\ &= \widetilde{H}(z) S_0(z), \end{aligned} \tag{12}$$

where

$$\widetilde{H}(z) = 1 - \sum_{i=1}^{N}\sum_{j=0}^{L} w_{ij}^0 z^{-2j} H_{s_i s_0}(z). \tag{13}$$

If the transfer function of filter B is $\widetilde{H}^{-1}(z) = [\widetilde{H}(z)]^{-1}$, then by using (12) we get

$$\begin{aligned} Y_2(z) &= \widetilde{H}^{-1}(z)E_1(z) \\ &= \widetilde{H}^{-1}(z)[\widetilde{H}(z)S_0(z) + \mathrm{err}^0(z)] \\ &= S_0(z) + \widetilde{H}^{-1}(z)\,\mathrm{err}^0(z). \end{aligned} \tag{14}$$

Thus

$$y_2(k) = s_0(k) + e(k), \tag{15}$$

$$e(k) = \widetilde{h}^{-1}(k) * \mathrm{err}^0(k), \tag{16}$$

where $e(k)$ is the residual noise in the output signal $y_2(k)$, $\widetilde{h}^{-1}(k)$ is the inverse z-transform of $\widetilde{H}^{-1}(z)$, and $*$ is the convolution symbol.

As commonly assumed in ANC, the noise $n_0(k)$ is uncorrelated with the speech signal $s_0(k)$ and the mean value of $n_0(k)$ is zero [4]. Thus in order that the system transfer function of filter B approximates $\widetilde{H}^{-1}(z)$, we need only to adjust the coefficients of filter B to minimize the square sum of $e_2(k)$. This is because

$$\begin{aligned} ||e_2(k)||^2 &= ||x_0(k) - y_2(k)||^2 \\ &= ||s_0(k) + n_0(k) - y_2(k)||^2 \\ &= ||n_0(k)||^2 + ||s_0(k) - y_2(k)||^2 \\ &\quad + 2n_0(k)(s_0(k) - y_2(k)), \end{aligned} \tag{17}$$

$$E[e_2^2(k)] = E[n_0^2(k)] + E[s_0(k) - y_2(k)]^2. \tag{18}$$

From (17), we may conclude that to minimize $E[e_2^2(k)]$ we need to minimize $E[s_0(k) - y_2(k)]^2$ which implies minimizing the error between $y_2(k)$ and $s_0(k)$.

The power of residual noise $e(k) = \widetilde{h}^{-1}(k) * \mathrm{err}^0(k)$ in the output enhanced speech $y_2(k)$ (15) is generally, though not always, smaller than the noise $n_0(k)$ in the original noisy speech signal $x_0(k) = s_0(k) + n_0(k)$. We might explain this as follows.

During NVP, the power of $e_1(k)$ would be quite small because the noise is efficiently cancelled through filter A. Then during the next VP, noise is still effectively cancelled while speech signal is minimally attenuated. This is because the speech source is located at a different location from the noisy source. The amplitude response of the noise cancellation subsystem would form notches in the noises propagation paths and accordingly the noises are successfully cancelled. However, the speech propagation directions do not mainly fall within these notches due to the assumption that speech source location deviates from the noise sources locations. As a result, $e_1(k)$ will have higher signal-to-noise ratio (SNR), where $p(k)$ is considered as the signal and $\mathrm{err}^0(k)$ is the noise, as indicated in (10). The purpose of filter B is to recover the original clean speech $s_0(k)$ from the distorted speech $p(k)$. If the correlation between the speech signals $s_0(k)$ and $p(k)$ is high, then the SNR of $y_2(k)$ will be higher than that of the original signal $x_0(k)$ acquired by the main microphone.

3.2. Improved spectrum subtraction

Despite the SNR of the enhanced speech $y_2(k)$ is highly improved through the MCRANC algorithm, enhanced speech still contains residual noise. If the noise $n_0(k)$ is stationary, the residual noise $e(k)$ in $y_2(k)$ will also be stationary. Additionally, if $n_0(k)$ is not stationary, $e(k)$ may well be quasi-stationary noise since the nonstationarity of the noise is cancelled to a certain degree by MCRANC algorithm. Thus, generally speaking $e(k)$ will have better stationarity than the original noise $n_0(k)$. So it will be more suitable to use improved spectrum subtraction (ISS) algorithm [1–3] to further enhance the preliminary enhanced speech $y_2(k)$. If we apply ISS algorithm directly to the original noisy speech $x_0(k)$, we may get poor enhancement result if the noise $n_0(k)$ is nonstationary or the SNR of $x_0(k)$ is low. In such cases, the music noise effect introduced by the spectrum subtraction algorithm will seriously harm the quality of the enhanced speech. As MCRANC can improve both the SNR of the noisy speech and the stationarity of the residual noise, ISS algorithm is more suitable to operate with $y_2(k)$ rather than $x_0(k)$.

ISS algorithm can be briefly described by the following. Divide $y_2(k)$ signal into suitable 50% overlapped frames. Hamming window is used to smooth each frame and to reduce spectrum leakage. Then apply DFT operation to each frame to obtain the power spectrum estimation of $y_2(k)$,

$$|Y_2(l)|^2 \approx |S_0(l)|^2 + |E(l)|^2, \tag{19}$$

where

$$Y_2(l) = \sum_{k=0}^{K-1} y_2(k)e^{-j(2\pi lk/K)} = |Y_2(l)|\,e^{j\varphi(l)}, \tag{20}$$

where K is the length of the frame, and $\varphi(l)$ is the phase of $Y_2(l)$.

Use the weighted average of several frames of the residual noise power spectrum $|\hat{E}(l)|^2$ during NVP as the estimation of $|E(l)|^2$.

Speech power spectrum is estimated by

$$|\hat{S}_0(l)|^2 = |Y_2(l)|^2 - \alpha|\hat{E}(l)|^2, \tag{21}$$

where α is called over-subtraction factor and is expressed by

$$\alpha = \alpha_0 - \frac{3}{20}\,\mathrm{SNR}, \quad 5\,\mathrm{dB} \le \mathrm{SNR} \le 20\,\mathrm{dB}, \tag{22}$$

where α_0 is the value of the over-subtraction factor α when SNR = 0 dB. Generally we take $\alpha_0 = 3$.

Half-wave rectification is used and is expressed as

$$|\hat{S}_0(l)|^2 = \begin{cases} |\hat{S}_0(l)|^2 & \text{if } |\hat{S}_0(l)|^2 \ge \beta|\hat{E}(l)|^2, \\ \beta|\hat{E}(l)|^2 & \text{otherwise,} \end{cases} \tag{23}$$

where β is a small positive number called spectrum base.

At last, the enhanced speech is

$$y(k) = \hat{s}_0(k) = \text{IDFT}\left(\left|\hat{S}_0(l)\right| e^{j\varphi(l)}\right). \tag{24}$$

3.3. System adaptation

In the proposed scheme a VD is needed to detect the NVP and VP intervals in the processed utterances [17]. MCRANC updates the optimal weights of filter A during the NVP intervals while the optimal weights of filter B are updated during the VP intervals. ISS updates the noise power spectrum estimation during NVP intervals. These updates would allow the speech enhancement system track the changes in the environment.

The problem here is that it is neither easy nor accurate to detect the VP and NVP intervals in noisy speech. To overcome this problem, these periods are substituted by easy to detect subperiods called voiced segment (VS) and nonvoiced segment (NVS) to replace VP and NVP intervals, respectively. Thus the adaptation of filter A will be processed during NVS rather than NVP whereas the adaptation of filter B will be conducted during VS rather than VP.

The adaptation rules can be formulated as follows.

Let us divide the discrete time axis as

$$[0,\infty) = \bigcup_{j=1}^{\infty} \left\{\left[t'_{1j}, t''_{1j}\right) \cup \left[t'_{2j}, t''_{2j}\right)\right\}, \tag{25}$$

where the discrete time interval $[t'_{1j}, t''_{1j})$ is an NVP of the main channel signal $x_0(k)$ while $[t'_{2j}, t''_{2j})$ is a VP of $x_0(k)$, and $t'_{1j} < (t''_{1j} = t'_{2j}) < t''_{2j}$. Select NVS $[\hat{t}'_{1j}, \hat{t}''_{1j}) \subseteq [t'_{1j}, t''_{1j})$ and VS $[\hat{t}'_{2j}, \hat{t}''_{2j}) \subseteq [t'_{2j}, t''_{2j})$.

Filter A weights are updated during the NVS $[\hat{t}'_{1j}, \hat{t}''_{1j})$ intervals and filter B weights are updated during the VS $[\hat{t}'_{2j}, \hat{t}''_{2j})$ intervals. During time intervals apart from VS and NVS, filters A and B only perform as normal filters with fixed weights. For ISS, the residual noise power spectrum $\hat{E}(l)$ is estimated during the NVS $[\hat{t}'_{1j}, \hat{t}''_{1j})$ intervals.

We confirm here again that the above adaptation rules are based on the assumption that we have stable or slowly varying environments.

During NVP $[t'_{1j}, t''_{1j})$ to VP $[t'_{2(j+1)}, t''_{2(j+1)})$, if the environment does not change, the impulse responses $h_{ni}(k)$ and $h_{si}(k)$ $(i = 1, \ldots, N)$ will remain unchanged. Thus the optimal weights of filter A derived during NVS $[\hat{t}'_{1j}, \hat{t}''_{1j})$ may also be kept fixed during the next NVP $[t'_{1(j+1)}, t''_{1(j+1)})$. Also, the optimal weights of filter B derived during VS $[\hat{t}'_{2j}, \hat{t}''_{2j})$ may also be considered optimal weights during the next VP $[t'_{2(j+1)}, t''_{2(j+1)})$. Accordingly, even if the speech enhancement system misses to find NVS $[\hat{t}'_{1(j+1)}, \hat{t}''_{1(j+1)})$ or VS $[\hat{t}'_{2(j+1)}, \hat{t}''_{2(j+1)})$ it will still perform well. If the environment changes during this time period but the system misses to find NVS $[\hat{t}'_{1(j+1)}, \hat{t}''_{1(j+1)})$ or VS $[\hat{t}'_{2(j+1)}, \hat{t}''_{2(j+1)})$, it will not perform perfectly in this short time period. However, once the next NVS $[\hat{t}'_{1(j+2)}, \hat{t}''_{1(j+2)})$ and VS $[\hat{t}'_{2(j+2)}, \hat{t}''_{2(j+2)})$ are detected, the system will perform perfectly again.

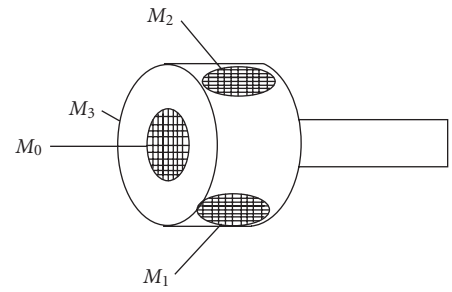

Figure 3: A solid microphone array.

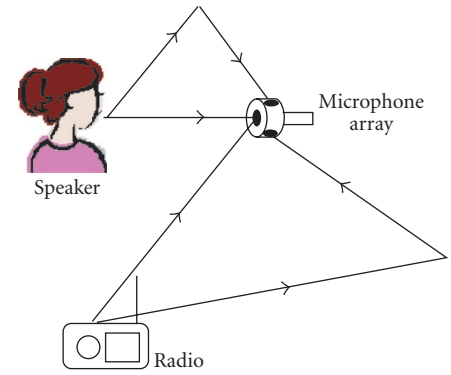

Figure 4: A scenario of noisy speech environment.

To adaptively find the optimal weights of FIR filters A and B, we may use any algorithm such as LMS, NLMS, RLS, BFTF, LSLL, GRBLS, [4, 6, 18–21]. The algorithms with quick convergence will better track changes in the environment. But they usually have higher computational complexity. For hardware implementation, one should select the algorithm that suits the computational power of the platform used.

4. EXPERIMENTS

Several experiments have been conducted to benchmark the performance of the proposed system against some commonly used systems with parallel paradigms.

4.1. Experiment 1

One of our experiments is carried out in a common research room about $8 \times 5 \times 3$ meters. In the experiment, four small microphones $M_0, M_1, \ldots, M_3$ are employed and closely placed on a cylindrical shape structure with 1 cm radius as shown in Figure 3. M_0 is placed onto the top surface of the cylinder while the referential microphones are embedded into the side surface. The noise is generated from an improperly tuned radio located at about 1.5 meter from the microphone array, as shown in Figure 4. The speech is coming from a person at 0.5 meter from the microphones. The sampling rate is 8 KHz.

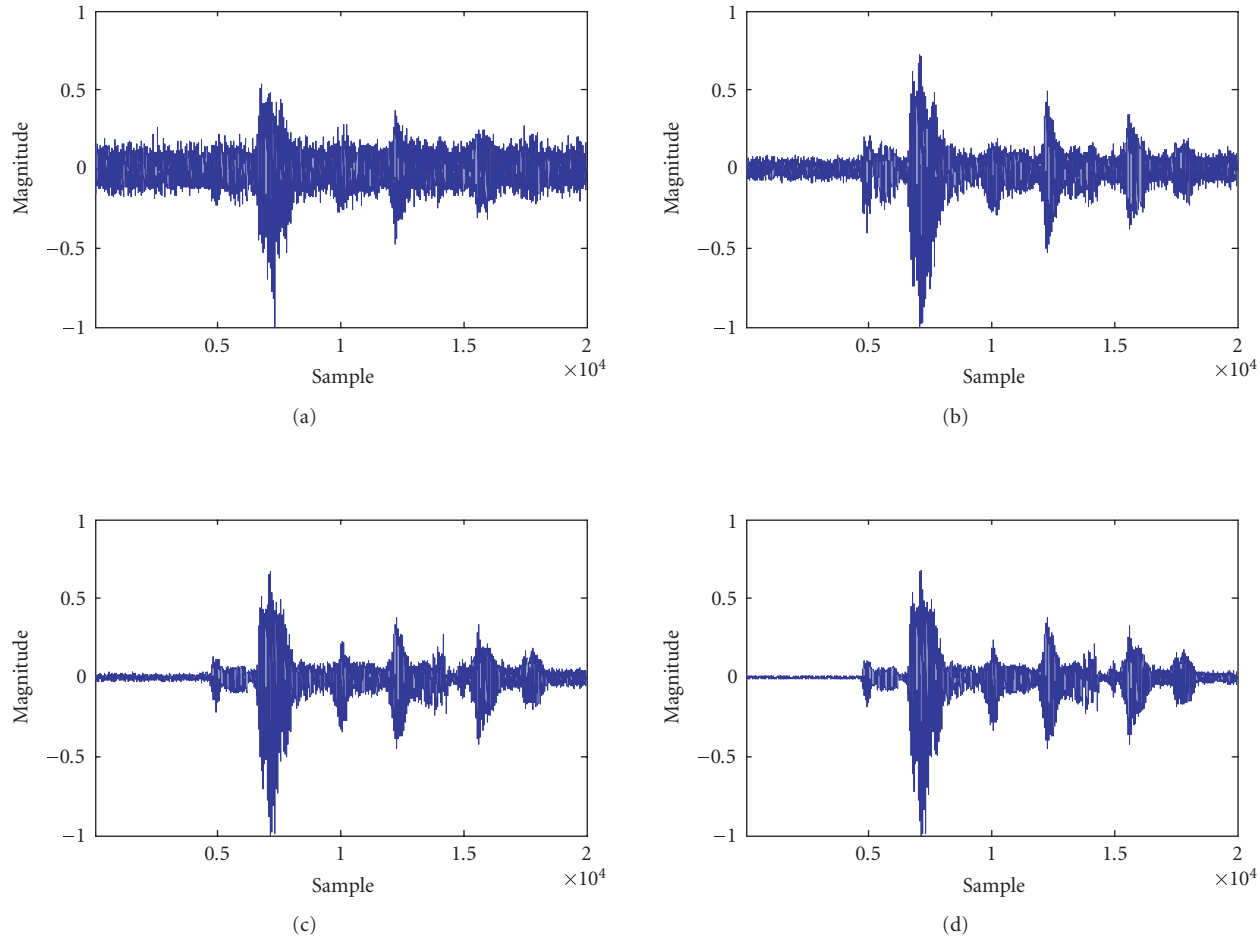

Figure 5: Results of Experiment 1: (a) noisy speech signal; (b) enhanced speech by two-channel CRANC; (c) enhanced speech by MCRANC; (d) enhanced speech by MCRANC and ISS.

For parameter adaptation, the normalized least mean square (NLMS) algorithm is employed to find the optimum weights of FIR filters A and B. For filter A, the tapped delay line per channel uses $L = 32$ delay units and hence filter A has 99 coefficients. The number of coefficients of filter B is selected to be 48.

In ISS, the window frame length $K = 256$ with 50% overlapped and using Hamming window for smoothing. We average the power spectrum over 3 frames of pure noise during NVS to estimate the residual noise power spectrum $|E(l)|^2$. Over-subtraction factor estimation, shown in (22), uses $\alpha_0 = 4$ and the spectrum-base factor, appears in (23), $\beta = 0.1$.

For the speech signal under investigation, the first NVS interval is detected with the samples $[1, 2, \ldots, 2000)$ and the subsequent VS interval is detected with the samples $[5001, 5002, \ldots, 20000)$.

Figure 5 shows visually the performance of the proposed speech enhancement system. Figure 5(a) is the noisy speech signal $x_0(k)$ acquired by the main microphone with SNR of 2.8 dB. Signals acquired by the referential microphones are visually similar to $x_0(k)$ and they do not need to be replicated. Figure 5(b) is the enhanced speech using two-channel CRANC algorithm, with SNR improvement of 9.2 dB. Figure 5(c) is the enhanced speech by the proposed MCRANC algorithm with SNR improvement of 18.0 dB. Figure 5(d) is the enhanced speech using a system based on MCRANC augmented with ISS which achieves an SNR improvement of 27.0 dB. Since it is impossible to get the clean speech signal in this experiment the SNR here is computed by

$$\text{SNR} = 10 \log \frac{(K''/K') \sum_{k \in K_1} x^2(k) - \sum_{k \in K_2} x^2(k)}{\sum_{k \in K_2} x^2(k)}, \tag{26}$$

where $x(k)$ is the noisy speech signal concerned, K_1 is the set of speech signal samples (speech section) while K_2 is the set of noise samples (noise section), K' and K'' are the total number of samples within K_1 and K_2, respectively.

Figure 6 is a zoomed view of a short noise segment from Figure 5. Figure 7 is also a zoomed view of a short speech segment from Figure 5.

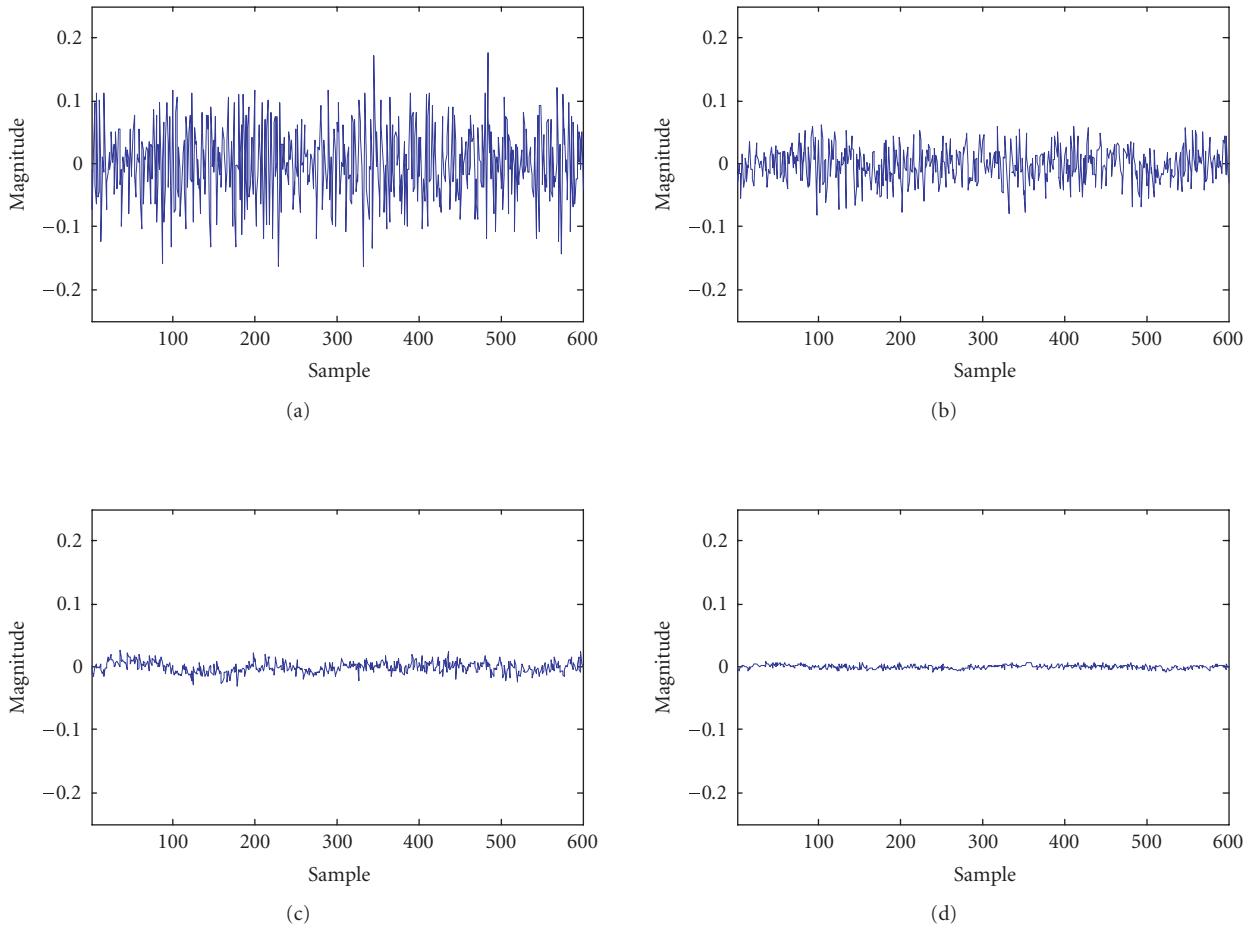

Figure 6: Zoomed view of a short noise segment from Figure 5 (pure noise): (a) pure noise segment; (b) output noise by two-channel CRANC; (c) output noise by MCRANC; (d) output noise by MCRANC and ISS.

4.2. Experiment 2

The second experiment is carried out in a Mitsubishi ETERNA car. A uniform linear array with four microphones is placed in front of the driver. Small microphones are collinearly placed with each neighboring microphones and are separated by 3 cm. The aperture of the array is about 13 cm. One of the two microphones near the center of the array is used as the main microphone while the rest are considered as referential microphones. The coexisting noises are generated by the car engine, air condition, and car radio. The noise from the radio is a piece of musical song. The speech is from the driver about 60 cm directly from the microphone array. The sampling rate is also 8 KHz.

For MCRANC and ISS used in the enhancement process, all parameters are as the same as those described in Experiment 1.

The NVP is detected with the samples $[1, 2, \ldots, 10500)$ and $[27001, 27002, \ldots, 30000)$, while VP is detected in between with the samples $[10501, 10502, \ldots, 27000)$. The samples $[1, 2, \ldots, 8000)$ are labeled as NVS and $[10501, 10502, \ldots, 27000)$ as VS.

Figure 8 shows the results of enhancements obtained from this experiment. Figure 8(a) is the noisy speech signal $x_0(k)$ acquired by the main microphone, with SNR $= -8.4$ dB. Figure 8(b) is the enhanced speech using the ISS algorithm only and giving SNR improvement of 14.5 dB. Figure 8(c) is the enhanced speech obtained by using the proposed MCRANC algorithm, with SNR improvement of 15.1 dB. Figure 8(d) is the enhanced speech by joining MCRANC and ISS algorithms, which offers an SNR improvement of 25.4 dB. The SNR is also estimated by applying (26).

4.3. Discussions

In Experiment 1, the noise source is near the microphone array and speech enhancement is mainly achieved by MCRANC. In experiment 2, the noise source is relatively far from the microphone array since the loudspeaker is in the rear part of the car, and the SNR improvement by MCRANC decreases. In fact, the amount of cancelled noises by MCRANC is highly related to the correlations between the main microphone and any of the referential microphones. In real environment, the closer the noise sources to the array,

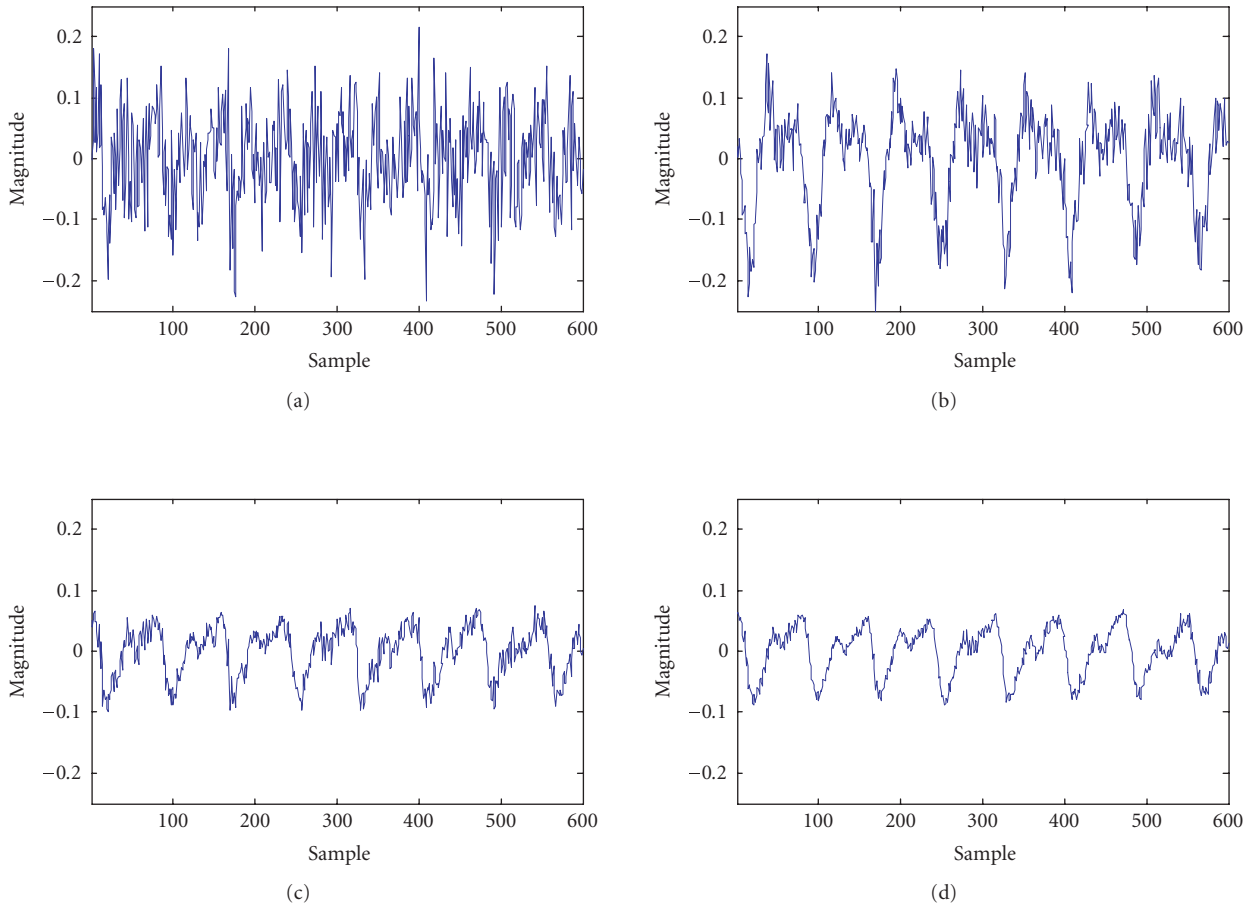

Figure 7: Zoomed view of a short speech segment from Figure 5 (noisy speech): (a) noisy speech segment; (b) enhanced speech by two-channel CRANC; (c) enhanced speech by MCRANC; (d) enhanced speech by MCRANC and ISS.

the higher the correlations, and so the greater the amount of noise cancelled.

As pointed out in [15], the signal enhancement achieved by using CRANC algorithm is sensitive to the positions of the sensors. From our experiments, we also find that the SNR of the enhanced speech by MCRANC is sensitive to the position of the microphone array. The speech enhancement performance depends on the positions of the speaker and noise sources, the surrounding space environment, and the type of noise. As a matter of fact, these factors have great influence on all ANC related algorithms. For MCRANC, the direction of the speaker with respect to the microphone array is better being different from the directions of the noise sources to the array. In other words, the speaker should not be very near from any of the noise sources. Despite these drawbacks, MCRANC still provides quite good speech enhancement in many cases. When ISS is cascaded with MCRANC, the whole system performs better than any of them alone.

5. CONCLUSIONS

In this paper a scheme is presented for speech enhancement, in which MCRANC algorithm is used to obtain a primary enhancement of noisy speech signals then followed by ISS stage to further improve the enhancement performance.

The MCRANC stage partially cancels out the introduced noise in the acquired speech signal. Thus it improves the SNR of the speech signal whereas minimum distortion incurred due to the enhancement process. This would almost assure preserving the speech quality. The MCRANC stage thus provides a more appropriate signal to the ISS stage for further improvement in the SNR while keeping the introduced spectrum subtraction byproduct (music-noise) to a minimum level.

As per implementation, the MCRANC technique employs only two FIR filters and a common voice detector. It has very good stability and low computational complexity, as well as it is easy to realize.

It also permits the microphones to be closely placed. As a result, the speech enhancement system based on the proposed scheme may use a small size microphone array and can achieve better speech enhancement than ISS, CRANC, or MCRANC algorithms alone. It is also quite easy for implementation.

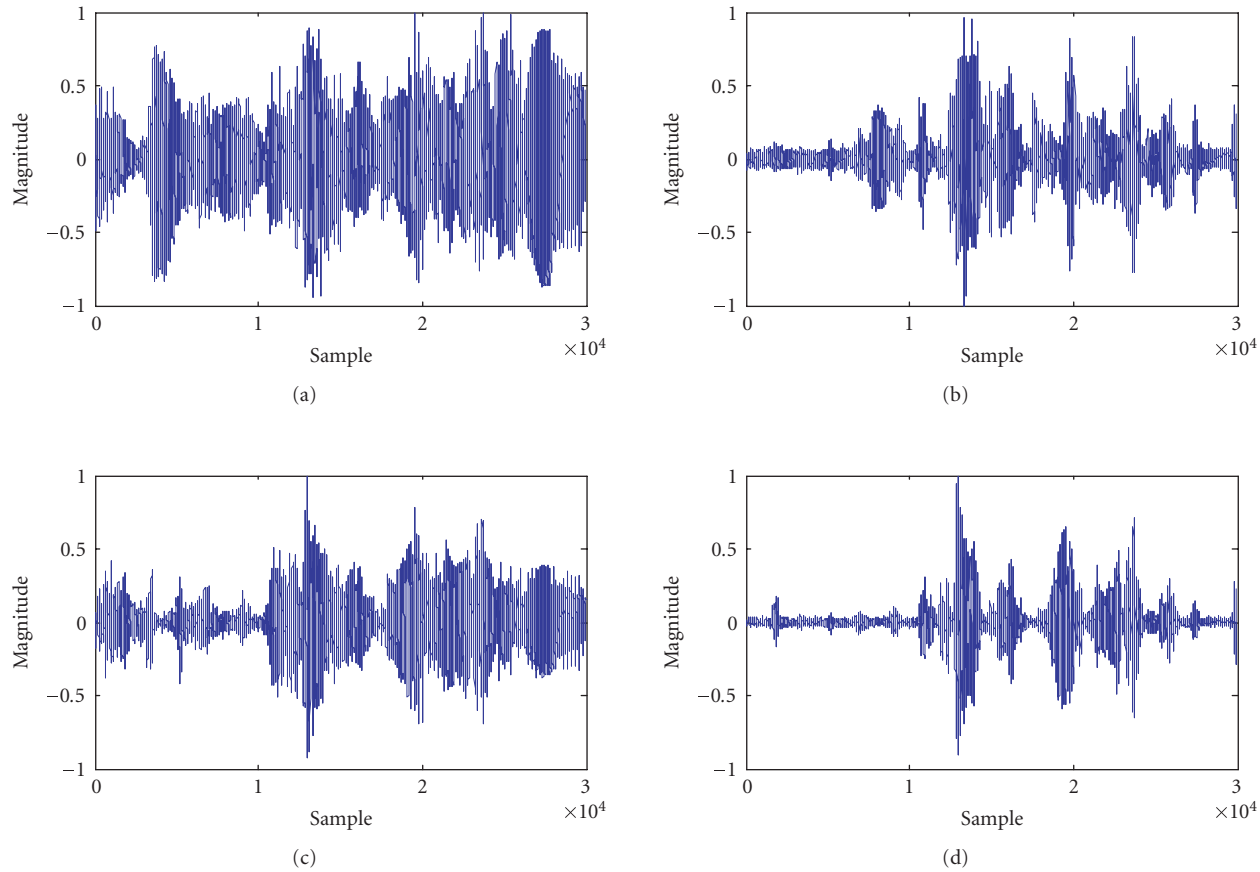

Figure 8: Results of Experiment 2: (a) noisy speech; (b) enhanced speech by ISS; (c) enhanced speech by MCRANC; (d) enhanced speech by MCRANC and ISS.

ACKNOWLEDGMENTS

This research is funded by The University of Auckland Research Committee Grant no.3603819 and partially by the National Nature Science Foundation of China Grant no.60272038.

REFERENCES

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[2] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proceedings of 4th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '79)*, vol. 4, pp. 208–211, Washington, DC, USA, April 1979.

[3] S. Ogata and T. Shimamura, "Reinforced spectral subtraction method to enhance speech signal," in *Proceedings of IEEE Region 10 International Conference on Electrical and Electronic Technology*, vol. 1, pp. 242–245, Singapore, August 2001.

[4] S. Haykin, *Adaptive Filter Theory*, Prentice-Hall, Upper Saddle River, NJ, USA, 1996.

[5] A. Hussain, "Multi-sensor adaptive speech enhancement using diverse sub-band processing," *International Journal of Robotics and Automation*, vol. 15, no. 2, pp. 78–84, 2000.

[6] O. Hoshuyama, A. Sugiyama, and 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. 10, pp. 2677–2684, 1999.

[7] R. Zelinski, "Noise reduction based on microphone array with LMS adaptive post-filtering," *Electronics Letters*, vol. 26, no. 24, pp. 2036–2037, 1990.

[8] R. Le Bouquin, "Enhancement of noisy speech signals: application to mobile radio communications," *Speech Communication*, vol. 18, no. 1, pp. 3–19, 1996.

[9] R. Martin, "Small microphone arrays with postfilters for noise and acoustic echo reduction," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds., pp. 255–276, Springer, Berlin, Germany, 2001.

[10] M. Dahl, I. Claesson, and S. Nordebo, "Simultaneous echo cancellation and car noise suppression employing a microphone array," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '97)*, vol. 1, pp. 239–242, Munich, Germany, April 1997.

[11] S. M. Kuo and W. M. Peng, "Principle and applications of asymmetric crosstalk-resistant adaptive noise canceler," in *Proceedings of IEEE Workshop on Signal Processing Systems (SiPS '99)*, pp. 605–614, Taipei, Taiwan, October 1999.

[12] G. Madhavan and H. De Bruin, "Crosstalk resistant adaptive noise cancellation," *Annals of Biomedical Engineering*, vol. 18, no. 1, pp. 57–67, 1990.

[13] G. Mirchandani, R. C. Gaus Jr., and L. K. Bechtel, "Performance characteristics of a hardware implementation of the cross-talk resistant adaptive noise canceller," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '86)*, pp. 93–96, Tokyo, Japan, April 1986.
[14] G. Mirchandani, R. Zinser Jr., and J. Evans, "A new adaptive noise cancellation scheme in the presence of crosstalk," *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, vol. 39, no. 10, pp. 681–694, 1992.
[15] V. Parsa, P. A. Parker, and R. N. Scott, "Performance analysis of a crosstalk resistant adaptive noise canceller," *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, vol. 43, no. 7, pp. 473–482, 1996.
[16] R. Zinser Jr., G. Mirchandani, and J. Evans, "Some experimental and theoretical results using a new adaptive filter structure for noise cancellation in the presence of cross-talk," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '85)*, vol. 10, pp. 1253–1256, Tampa, Fla, USA, April 1985.
[17] S. Jongseo and S. Wonyong, "A voice detector employing soft decisio based noise spectrum adaptation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '98)*, vol. 1, pp. 365–368, Seattle, Wash, USA, May 1998.
[18] B. Friedlander, "Lattice filters for adaptive processing," *Proceedings of IEEE*, vol. 70, no. 8, pp. 829–867, 1982.
[19] M. L. Honig and D. G. Messerschmitt, "Convergence properties of an adaptive digital lattice filter," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 3, pp. 642–653, 1981.
[20] F. Ling, D. Manolakis, and J. Proakis, "Numerically robust least-squares lattice-ladder algorithms with direct updating of the reflection coefficients," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 4, pp. 837–845, 1986.
[21] F. Ling, "Givens rotation based least squares lattice and related algorithms," *IEEE Transactions on Signal Processing*, vol. 39, no. 7, pp. 1541–1551, 1991.

Qingning Zeng received the B.S. degree from the Harbin Institute of Technology, China, in 1982, and the M.S. degrees from the Xidian University, China, in 1987, both in applied mathematics. From 1995 to 1997, he was a Visiting Scholar in the Department of Information and Systems, University of Rome "La Sapienza," Italy. Now he is doing research work in The University of Auckland, New Zealand. He has published more than 40 papers including an invention patent and organized more than 8 research projects. His research interests are in the areas of audio signal processing, image recognition, mathematic programming, and Markov decision process.

Waleed H. Abdulla has a Ph.D. degree from the University of Otago, Dunedin, New Zealand. He was awarded Otago University Scholarship for 3 years and the Bridging Grant. He has been working since 2002 as a Senior Lecturer in the Department of Electrical and Computer Engineering, The University of Auckland. He was a Visiting Researcher to Siena University, Italy, in 2004. He has collaborative work with Essex University in UK, IDIAP Research Centre in Switzerland, Tsinghua University, and Guilin University of Electronic Technology in China. He is the Head of the Speech Signal Processing and Technology Group. He has more than 40 publications including a patent and a book. He has supervised more than 20 postgraduate students. He has many awards and funded projects. He is a Reviewer of many conferences and journals. He is the Deputy Chair of the Scientific Committee of the ASTA 2006 Conference and Member of the Advisory Board of IE06 Conference. His research areas are in developing generic algorithms, speech signal processing, speech recognition, speaker recognition, speaker localization, microphone arrays modeling, speech enhancement and noise cancelation, statistical modeling, human biometrics, EEG signal analysis and modeling, time-frequency analysis, and neural networks applications. He is a Member of ISCA, IEE, and IEEE.