Adaptive Co-Channel Speech Separation and Recognition

Kuan-Chieh Yen and Yunxin Zhao, *Senior Member, IEEE*

***Abstract*— An improved technique of co-channel speech separation, S-AADF/LMS, and its integration with automatic speech recognition is presented. The S-AADF/LMS technique is based on the algorithms of accelerated adaptive decorrelation filtering (AADF) and LMS noise cancellation, where a switching between the two algorithms is made depending upon the active/inactive status of the co-channel signal sources. The AADF improves the previous adaptive decorrelation algorithm in terms of system stability and estimation efficiency, and leads to better estimation of time-varying and reverberant channels. The S-AADF/LMS further improves the estimation accuracy when only one source signal remains active during certain periods of time. A coherence-function based source signal detection algorithm is also presented, which is successfully used in the switching between AADF and LMS and in extracting speech signals from leakage-corrupted background. Experiments were conducted under a simulated environment based on the measurements made of certain real room-acoustic conditions, and the results demonstrated the effectiveness of the proposed technique for co-channel speech separation and recognition.**

***Index Terms*— Acoustic channel estimation, active source signal detection, adaptive decorrelation filtering, automatic speech recognition, co-channel speech separation.**

I. Introduction

THE state-of-the-art speech recognition technology is still vulnerable to the presence of interfering signals [1]. Many research efforts have focused on the stationary and broadband noise sources [2]–[4]. These studies either assume that the noise statistics are known *a priori*, or that they can be estimated from certain inactive period of speech. In active environments where interfering signals are inherently time-varying, such as the co-channel interference from a competing talker, the noise characteristics estimated at one instant might not be applicable at a later time. Furthermore, a single microphone is normally used for speech acquisition, which limits the effectiveness of the techniques intended for handling time-varying interference [5], [6]. While this is usually the result of system constraints (speech acquired on the telephone line, for example), in other applications where multimicrophone acquisition is feasible (such as teleconferencing or speech controlled devices), the additional information makes more effective processing possible.

Several techniques based on multimicrophone processing, such as speech enhancement based on subband adaptive processing [7]–[9] and blind separation in multipath environment [10], have been explored in the recent years. Among the techniques using two-microphone speech acquisition, Widrow's LMS noise cancellation algorithm [11] has been widely used. This algorithm focuses on restoring only the primary signal, and has difficulties when the primary signal is also picked up by the reference microphone. In recent literature, a few researchers proposed algorithms for signal separation via the adaptive decorrelation filtering (ADF) between two simultaneously acquired co-channel signals [12], [13]. These algorithms are shown to be capable of reducing the cross-channel interference and are more general than Widrow's LMS algorithm.

In the current work, we propose several improvements to the ADF algorithm in the aspects of estimation stability and efficiency and describe an integrated co-channel speech separation and recognition system [14], [15]. In this system, two co-existent and independent speech sources are considered, and their convolutive mixtures are acquired via two microphones. The acquired signals are first processed to separate out the co-channel speech signals, and the separated signals are then analyzed by a coherence-function based source detection algorithm to determine the active regions of each source. The separated speech signals in their respective active regions are recognized by a hidden Markov model (HMM) based speaker-independent continuous speech recognition (SICSR) system [16].

This paper is organized into eight sections. In Section II, the background of the co-channel system and the adaptive decorrelation filtering algorithm are briefly described. In Section III, several improvements to the ADF algorithm are discussed, including a blockwise implementation, an upper bound of adaptation gain for stability, a power normalization on the adaptation gain, and an accelerated version of ADF. In Section IV, a source detection algorithm based on the cross-spectral coherence function of the processed signals is developed in order to determine the active and inactive regions of each signal source. In Section V, a strategy for switching between the ADF or AADF algorithm and the traditional LMS algorithm is devised based on the active/inactive regions of the source signals. The specifications of the automatic speech recognition (ASR) system and experimental results are given in Section VI, and a conclusion is made in Section VII.

Manuscript received April 11, 1997; revised February 27, 1998. This work was supported by the National Science Foundation under Grant IRI-95-02074 and by a grant from the Whitaker Foundation. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jean-Claude Junqua.

K. C. Yen is with the Department of Electrical and Computer Engineering and Beckman Institute, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA (e-mail: yen@ifp.uiuc.edu).

Y. Zhao is with the Department of Computer Engineering and Computer Science, University of Missouri, Columbia, MO 65211 USA.

Publisher Item Identifier S 1063-6676(99)01633-8.

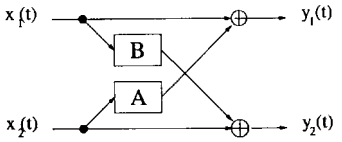

Fig. 1. Block diagram of the co-channel system.

II. Co-Channel System and Signal Separation

A. Co-Channel Speech Acquisition System

In a co-channel speech acquisition system, each microphone acquires not only its target signal, but also the interfering signals from the other sources. For simplicity, our discussion is limited to the two-source two-microphone case. Let $x_1(t)$ and $x_2(t)$ be the signals generated by sources 1 and 2, respectively, which are assumed to be independent of each other. The signal acquired by the microphone that targets the source 1 is denoted by $y_1(t)$, and that acquired by the microphone that targets the source 2 is denoted by $y_2(t)$. Using the linear filters A and B to model the channel coupling effects and assuming no distortion between each microphone and its target source, the co-channel system can be described in the frequency domain as

$$\begin{aligned} Y_1(f) &= X_1(f) + A(f)X_2(f) \\ Y_2(f) &= X_2(f) + B(f)X_1(f). \end{aligned} \tag{1}$$

This co-channel system is illustrated in Fig. 1.

B. Signal Separation by Adaptive Decorrelation Filtering

Let $\hat{A}(f)$ and $\hat{B}(f)$ be the estimates of the channel filters A and B, respectively. Define the filter $C(f) = 1 - \hat{A}(f)\hat{B}(f)$, and define the Fourier transforms of the signals $v_1(t)$ and $v_2(t)$ as

$$\begin{aligned} V_1(f) &= Y_1(f) - \hat{A}(f)Y_2(f) \\ V_2(f) &= Y_2(f) - \hat{B}(f)Y_1(f). \end{aligned} \tag{2}$$

It is easy to verify that if $A(f) = \hat{A}(f)$ and $B(f) = \hat{B}(f)$, then

$$V_i(f) = C(f)X_i(f), \qquad i = 1, 2.$$

Therefore, if the filters A and B are known, the signals from the sources 1 and 2 can be separated from the acquired signals by (2). Furthermore, if $C(f)$ is invertible, the source signals can be perfectly reconstructed by

$$\hat{X}_i(f) = C^{-1}(f)V_i(f), \qquad i = 1, 2. \tag{3}$$

The (2) and (3) provide the basis for separating the source signals $x_1(t)$ and $x_2(t)$ from the acquired signals $y_1(t)$ and $y_2(t)$, and a block diagram of such a separation system is illustrated by Fig. 2.

Since in most applications the coupling channels are time-varying and unknown, the filters A and B need to be adaptively estimated. It was shown in [12] that if the source signals are zero-mean and uncorrelated, and if the filters $\hat{A}$ and $\hat{B}$ are finite impulse response (FIR) filters represented by $\underline{a} = [a_0, \cdots, a_{N_a-1}]^T$ and $\underline{b} = [b_0, \cdots, b_{N_b-1}]^T$, where the superscript T denotes vector transpose, then the filter

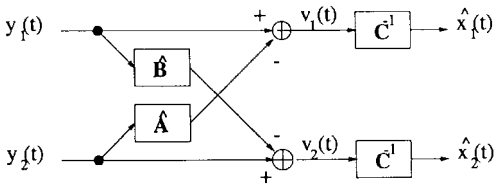

Fig. 2. Block diagram of the source separation system.

coefficients can be estimated recursively by the following equations:

$$\begin{aligned} \underline{a}^{(t)} &= \underline{a}^{(t-1)} + \mu(t)\underline{v}_2^{(t-1)}(t)v_1^{(t-1)}(t) \\ \underline{b}^{(t)} &= \underline{b}^{(t-1)} + \mu(t)\underline{v}_1^{(t-1)}(t)v_2^{(t-1)}(t) \end{aligned} \tag{4}$$

where $\underline{a}^{(t)}$ and $\underline{b}^{(t)}$ denote the estimates of $\underline{a}$ and $\underline{b}$ at time t, and $v_1^{(t-1)}(\tau)$ and $v_2^{(t-1)}(\tau)$ denote the values of signals $v_1(\tau)$ and $v_2(\tau)$ calculated according to $\underline{a}^{(t-1)}$ and $\underline{b}^{(t-1)}$:

$$\begin{aligned} v_1^{(t-1)}(\tau) &= y_1(\tau) - \underline{y}_2(\tau)^T\underline{a}^{(t-1)} \\ v_2^{(t-1)}(\tau) &= y_2(\tau) - \underline{y}_1(\tau)^T\underline{b}^{(t-1)}. \end{aligned} \tag{5}$$

The vectors $\underline{y}_1(t), \underline{y}_2(t), \underline{v}_1^{(t-1)}(t)$ and $\underline{v}_2^{(t-1)}(t)$ are defined as

$$\begin{aligned} \underline{y}_1(t) &= [y_1(t) \cdots y_1(t - N_b + 1)]^T \\ \underline{y}_2(t) &= [y_2(t) \cdots y_2(t - N_a + 1)]^T \\ \underline{v}_1^{(t-1)}(t) &= [v_1^{(t-1)}(t) \cdots v_1^{(t-1)}(t - N_b + 1)]^T \\ \underline{v}_2^{(t-1)}(t) &= [v_2^{(t-1)}(t) \cdots v_2^{(t-1)}(t - N_a + 1)]^T. \end{aligned}$$

The adaptation gain $\mu(t)$ will be discussed in the next section.

III. Modification and Enhancement on Adaptive Decorrelation Filtering

In [12], it is recommended to use γ or γ/t as the adaptation gain $\mu(t)$, where γ is a constant. However, if $\mu(t) = \gamma$ is used, γ has to be very small in order to avoid instability, which limits the efficiency of the system. On the other hand, if $\mu(t) = \gamma/t$ is used, the adaptation gain will diminish toward zero as t increases, and hence is not suitable for time-varying environments. Therefore, a blockwise implementation based on $\mu(t) = \gamma/t$ is chosen. Furthermore, an upper bound for γ is first derived for ensuring system stability, and then, a method of choosing γ based on power normalization is developed accordingly. An accelerated adaptation gain sequence is further proposed to replace $\mu(t) = \gamma/t$ for enhanced efficiency, which is very important for the estimation of time-varying coupling channels.

A. Blockwise Implementation

In the blockwise implementation, the co-channel speech signals acquired simultaneously by two microphones are blocked into two sequences of frames, where the frames are synchronized between the two channels. Each frame has N samples, and the shift between successive frames is M samples. Here N is usually a multiple of M. The frames are labeled by $m = 0, 1, \cdots$, and the acquired signal sample in the frame m from

the channel i is denoted as $y_{i,[m]}(t), t=0,\cdots,N-1, i=1,2$, i.e.,

$$y_{i,[m]}(t)=y_i(Mm+t), \qquad i=1,2, \quad m=0,1,\cdots, \quad t=0,1,\cdots,N-1.$$

The processing in each frame involves the following four steps, where the subscript $[m]$ denotes the frame index:

Step 1: Initialize $\underline{a}_{[m]}$ and $\underline{b}_{[m]}$ by $\underline{a}_{[m]}^{(0)}=\underline{a}_{[m-1]}^{(N-1)}$ and $\underline{b}_{[m]}^{(0)}=\underline{b}_{[m-1]}^{(N-1)}$, since the estimates of $\underline{a}$ and $\underline{b}$ from the previous frame are usually good initial values for the current frame. For $m=0$, simply set $\underline{a}_{[0]}^{(0)}=\underline{0}$ and $\underline{b}_{[0]}^{(0)}=\underline{0}$.

Step 2: Adapt $\underline{a}_{[m]}^{(t)}$ and $\underline{b}_{[m]}^{(t)}$ from $t=1$ to $t=N-1$ according to (4), with $\mu(t)=\gamma/t$.

Step 3: Use $\underline{a}_{[m]}=\underline{a}_{[m]}^{(N-1)}$ and $\underline{b}_{[m]}=\underline{b}_{[m]}^{(N-1)}$ as the estimates of A and B for frame m, and compute $v_{1,[m]}(t)$ and $v_{2,[m]}(t), t=0,\cdots,N-1$, according to (5).

Step 4: Compute the restored signals, $\hat{x}_{1,[m]}(t)$ and $\hat{x}_{2,[m]}(t)$, recursively from $t=0$ to $t=N-1$ according to [12, eqs. (37) and (38)]:

$$\hat{x}_{i,[m]}(t)=\frac{1}{c_{[m]}(0)}\left[v_{i,[m]}(t)-\sum_{k=1}^{N_a+N_b-2}c_{[m]}(k)\hat{x}_{i,[m]}(t-k)\right]$$

where

$$c_{[m]}(k)=\delta(k)-\sum_{l=0}^{k}a_{[m]}(l)b_{[m]}(k-l), \qquad k=0,\cdots,N_a+N_b-2.$$

By using the blockwise method, the signals can be processed within time intervals of several frames, i.e., long before the end of the utterances, hence, processing delay can be reduced, which is especially important for real-time processing. Note that there are overlaps between successive frames when $N>M$. In this case, the multiple values corresponding to the same signal sample in different frames are averaged to produce the restored sample. In general, allowing more overlaps yields better restoration with the cost of increased computation.

B. Choice of Adaptation Gain

It can be shown that the following bound can be used for γ to maintain stability (see Appendix A for derivation):

$$0<\gamma<\frac{2}{N_a\ \mathrm{var}\{y_2(t)\}+N_b\ \mathrm{var}\{y_1(t)\}}=\Gamma. \qquad (6)$$

Since the variances of $y_1(t)$ and $y_2(t), \mathrm{var}\{y_1(t)\}$, and $\mathrm{var}\{y_2(t)\}$, can be evaluated in each frame, the corresponding Γ can be calculated for each frame as in (6). Therefore, the adaptation gain in a frame is chosen as

$$\mu(t)=\alpha\Gamma/t, \qquad 0<\alpha<1 \qquad (7)$$

where α is a positive constant chosen according to the expected time variation rate of the acoustic environment. As such, $\gamma=\alpha\Gamma$ is normalized by the power of the incoming signals in each frame and satisfies the condition stated in (6). This method reduces the dependency of system stability on the power of the incoming signals, and hence makes the system more efficient.

The following two examples demonstrate the influence of the adaptation gain on the system stability and the effectiveness of the derived upper bound for γ. In each example, source signals chosen from TIMIT database were mixed by a pair of fixed channel filters to generate the co-channel signals. The major delay caused by the channel filters was approximately 2 ms, and the attenuation was approximately 8 dB.

Example 1: In this experiment, a set of co-channel signals were processed with $M=N=200$. Three methods of choosing adaptation gains were used:

$$\mu_1(t)=\frac{2}{\sqrt{N_aN_b\ \mathrm{var}\{y_1\}\ \mathrm{var}\{y_2\}}}\,1/t$$
$$\mu_2(t)=\frac{2}{\max(N_a\ \mathrm{var}\{y_2\},N_b\ \mathrm{var}\{y_1\})}\,1/t$$
$$\mu_3(t)=\frac{2}{N_a\ \mathrm{var}\{y_2\}+N_b\ \mathrm{var}\{y_1\}}^{1/t}=\Gamma/t.$$

The stability of the system was examined after each iteration of adaptation (i.e., every sample). Once the system became unstable, the filter coefficients were reset to zeros and the process continued until all the signal samples were processed. In more than 40 million iterations, the system was reset 23 times when using $\mu_1(t)$; three times when using $\mu_2(t)$; and zero times when using $\mu_3(t)$. This example shows that the bound Γ works well for the ADF algorithm in most situations.

Example 2: In this experiment, the co-channel signals were processed using $\mu_1(t)=5\Gamma/t, \mu_2(t)=\Gamma/t$, and $\mu_3(t)=0.2\Gamma/t$ as the adaptation gain, respectively. To evaluate the performance of ADF, the squared estimation error, $E(t)$, was defined as

$$E(t)=[\Delta\underline{a}^{(t)}]^T\Delta\underline{a}^{(t)}+[\Delta\underline{b}^{(t)}]^T\Delta\underline{b}^{(t)}$$
$$\Delta\underline{a}^{(t)}=\underline{a}^{(t)}-\underline{a}^*$$
$$\Delta\underline{b}^{(t)}=\underline{b}^{(t)}-\underline{b}^* \qquad (8)$$

where $*$ denotes the true filter coefficients. The relation between $E(t)$ and the number of processed frames is plotted in Fig. 3 for the three choices of the adaptation gains. Since the filter coefficients were all initialized as zeros, the beginning part of each curve represents the system behavior for a fast-changing channel, and the ending part represents the system behavior for time-invariant channel. It is shown in Fig. 3(b) that $E(t)$ was reduced faster at the beginning with $\mu_2(t)$, but was more stable at the end with $\mu_3(t)$. Also as shown in Fig. 3(a), the system quickly became unstable with $\mu_1(t)$, which has $\gamma>\Gamma$.

C. Accelerated ADF (AADF)

In the adaptive estimation of the co-channel system, it is desirable to apply a larger adaptation gain when the previous filter estimates differ from the current channel filters significantly; on the other hand, a smaller adaptation gain is desirable when the previous filter estimates are close to the current channel filters. A good way of implementing such an adaptation strategy is to use Kesten's procedure of accelerating convergence for stochastic approximation [17],

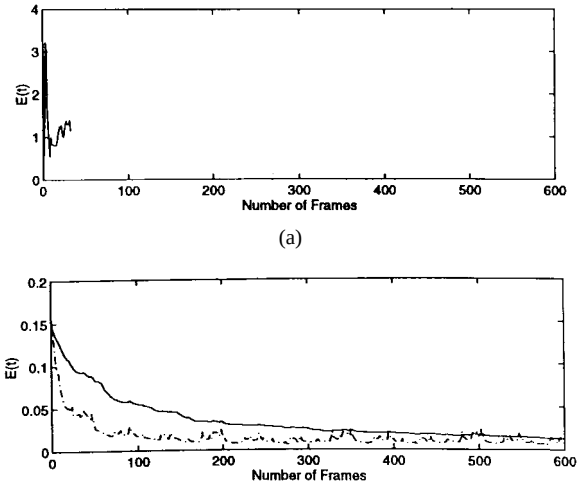

Fig. 3. Squared estimation error versus the number of frames using three different adaptation gains: (a) $\mu_1(t) = 5\Gamma/t$; (b) dashed curve: $\mu_2(t) = \Gamma/t$, solid curve: $\mu_3(t) = 0.2\Gamma/t$, where Γ is the derived upper-bound of adaptation gain constant γ for system stability.

where, instead of using the adaptation gain $\mu(t) = \alpha\Gamma/t$ for all filter coefficients, the ADF estimation equations are modified as

$$a_k^{(t)} = a_k^{(t-1)} + \frac{\alpha\Gamma}{i_{a,k}(t)}\, v_2^{(t-1)}(t-k)v_1^{(t-1)}(t)$$
$$b_k^{(t)} = b_k^{(t-1)} + \frac{\alpha\Gamma}{i_{b,k}(t)}\, v_1^{(t-1)}(t-k)v_2^{(t-1)}(t) \qquad (9)$$

where k is the filter index, and

$$i_{a,k}(1) = 1; \qquad k = 0, 1, \cdots, N_a - 1$$
$$i_{b,k}(1) = 1; \qquad k = 0, 1, \cdots, N_b - 1$$
$$i_{a,k}(t+1) = \begin{cases} i_{a,k}(t), & \text{if } \tilde{a}_k^{(t)}\tilde{a}_k^{(t-1)} > 0 \\ i_{a,k}(t)+1, & \text{if } \tilde{a}_k^{(t)}\tilde{a}_k^{(t-1)} \le 0 \end{cases}$$
$$i_{b,k}(t+1) = \begin{cases} i_{b,k}(t), & \text{if } \tilde{b}_k^{(t)}\tilde{b}_k^{(t-1)} > 0 \\ i_{b,k}(t)+1, & \text{if } \tilde{b}_k^{(t)}\tilde{b}_k^{(t-1)} \le 0 \end{cases}$$

and

$$\tilde{a}_k^{(t)} = v_2^{(t-1)}(t-k)v_1^{(t-1)}(t)$$
$$\tilde{b}_k^{(t)} = v_1^{(t-1)}(t-k)v_2^{(t-1)}(t).$$

As defined above, the signs (positive, negative) of the consecutive correlation terms control the adjustment of the adaptation gain for each filter coefficient, and the gain decreases only when the sign changes. In fact, the two extreme cases of (9) are equivalent to using $\mu(t) = \alpha\Gamma$ and using $\mu(t) = \alpha\Gamma/t$, respectively. This modified algorithm is referred to as accelerated ADF (AADF) in the following discussion.

The following example gives a performance comparison between the ADF and the AADF algorithms. The co-channel signals in the examples of Section III-B was processed using the following three schemes, with $M = N = 200$.

1) the ADF; $\alpha = 1$.
2) the ADF; $\alpha = 0.5$.
3) the AADF; $\alpha = 0.5$.

Similarly, the estimated filter coefficients were initialized as zeros. The relation between the squared estimation error $E(t)$ and the number of processed frames is plotted in Fig. 4 for all three cases. The results illustrate that the AADF with $\alpha = 0.5$ could track the channel variation almost as fast as the ADF with $\alpha = 1$, and the AADF was almost as stable as the ADF with $\alpha = 0.5$ when the channel became steady. As such, the accelerated algorithm enhances the system ability to accommodate a wider range of channel variations.

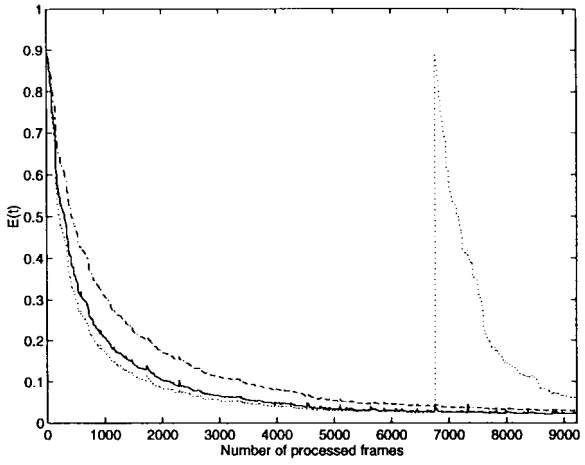

Fig. 4. Squared estimation error versus the number of processed frames: dotted curve: ADF with $\alpha = 1$; dashed curve: ADF with $\alpha = 0.5$; solid curve: AADF with $\alpha = 0.5$.

IV. Source Signal Detection

Although the ADF and AADF algorithms can significantly attenuate the interfering signals in each channel, the residual interfering signals, referred to as leakage signals, still pose problems for automatic speech recognition: in regions where the target speech source is inactive (silent) for an extended period of time, even a weak leakage signal could cause recognition errors (mainly insertions) and hence deteriorate the recognition accuracy. If the active regions of each source can be identified so that the ASR is constrained to be performed within the active regions of each target signal, the recognition accuracy may be improved. For such a purpose, a source detection algorithm is developed based on the coherence function of the restored signals to determine the active and inactive regions for each source.

A. The Use of Coherence Function for Source Detection

From (1)–(3), the relationship between the restored signals and the source signals in the frequency domain can be represented by

$$\hat{X}_1(f) = \frac{1 - \hat{A}(f)B(f)}{1 - \hat{A}(f)\hat{B}(f)}\, X_1(f) + \frac{A(f) - \hat{A}(f)}{1 - \hat{A}(f)\hat{B}(f)}\, X_2(f)$$
$$\hat{X}_2(f) = \frac{1 - A(f)\hat{B}(f)}{1 - \hat{A}(f)\hat{B}(f)}\, X_2(f) + \frac{B(f) - \hat{B}(f)}{1 - \hat{A}(f)\hat{B}(f)}\, X_1(f) \qquad (10)$$

where the first term in the right hand side of each equation is the linearly distorted source signal, and the second term is the leakage signal. Assuming that the filter estimates $\hat{A}$ and

$\hat{B}$ are very close to the true filters, i.e., $A(f) \approx \hat{A}(f)$ and $B(f) \approx \hat{B}(f)$, then the linear distortion is ignorable and (10) can be simplified as

$$\begin{aligned} \hat{X}_1(f) &\approx X_1(f) + G(f)X_2(f); \qquad |G(f)| \ll 1, \quad \forall f \\ \hat{X}_2(f) &\approx X_2(f) + H(f)X_1(f); \qquad |H(f)| \ll 1, \quad \forall f. \end{aligned} \tag{11}$$

From (11), each restored signal is approximately the sum of its source signal and a small leakage from the other source. Using a short-time discrete Fourier transform (DFT) of length L, (11) can be written as

$$\begin{aligned} \hat{X}_1(k) &\approx X_1(k) + G(k)X_2(k); \qquad |G(k)| \ll 1; \\ &\qquad k = 0, 1, \cdots, L-1 \\ \hat{X}_2(k) &\approx X_2(k) + H(k)X_1(k); \qquad |H(k)| \ll 1; \\ &\qquad k = 0, 1, \cdots, L-1. \end{aligned}$$

The coherence function $\rho(k)$ for each frequency bin k is defined as

$$\rho(k) = \sqrt{\frac{E\{\hat{X}_1(k)\hat{X}_2^*(k)\}E\{\hat{X}_1^*(k)\hat{X}_2(k)\}}{E\{|\hat{X}_1(k)|^2\}E\{|\hat{X}_2(k)|^2\}}}.$$

Since $x_1(t)$ and $x_2(t)$ are both zero-mean and independent of each other, $X_1(k)$ and $X_2(k)$ are also zero-mean and independent in each frequency bin. Define the short-time energy of the source signal i in the frequency bin k as

$$E_{i,k} = E\{|X_i(k)|^2\}$$

and it follows that

$$\begin{aligned} E\{|\hat{X}_1(k)|^2\} &= E_{1,k} + |G(k)|^2 E_{2,k} \\ E\{|\hat{X}_2(k)|^2\} &= E_{2,k} + |H(k)|^2 E_{1,k} \\ E\{\hat{X}_1(k)\hat{X}_2^*(k)\} &= H^*(k)E_{1,k} + G(k)E_{2,k} \\ E\{\hat{X}_1^*(k)\hat{X}_2(k)\} &= H(k)E_{1,k} + G^*(k)E_{2,k}. \end{aligned}$$

It can be shown that if $E_{1,k} \approx E_{2,k}$, i.e., the short-time energies of the two sources are comparable in the kth frequency bin, then the coherence function of this bin will be close to zero, i.e., $\rho(k) \approx 0$. On the other hand, if $E_{1,k} \ll E_{2,k}$ or $E_{1,k} \gg E_{2,k}$, i.e., the short-time energy of one source signal is much greater than that of the other signal in the kth frequency bin, then the coherence function of this bin will be close to unity, i.e., $\rho(k) \approx 1$. A source signal is very likely to be inactive if its short-time energy is much weaker than that of the other source for an extended period of time. Define the decision variables $\delta_i, i = 1, 2$, as

$$\delta_i = \begin{cases} 1: & \text{Source } i \text{ is active} \\ 0: & \text{Source } i \text{ is inactive} \end{cases}$$

and define

$$\begin{aligned} \hat{E}_i &= \sum_{k=0}^{L-1} E\{|\hat{X}_i(k)|^2\} \\ P &= \sum_{k=0}^{L-1} \rho^2(k). \end{aligned}$$

Then by choosing a threshold value T for the variable P, an active/inactive decision on each source can be made according to the following rule:

Case 1: If $P \le T$, both sources are active; set $\delta_1 = \delta_2 = 1$.

Case 2: If $P > T$ and $\hat{E}_1 > \hat{E}_2$, source 1 is active and source 2 is inactive; set $\delta_1 = 1$ and $\delta_2 = 0$.

Case 3: If $P > T$ and $\hat{E}_1 < \hat{E}_2$, source 1 is inactive and source 2 is active; set $\delta_1 = 0$ and $\delta_2 = 1$.

B. Implementation of the Source Detection Algorithm

To determine the presence/absence of each source signal at time t, the coherence function at time t, denoted as $\rho(k;t)$, needs to be evaluated. The short-time DFT coefficients of the restored signals at time t is defined as

$$\hat{X}_i(k;t) = \sum_{l=1}^{L} \hat{x}_i(t-L+l)e^{-j(2\pi kl/L)}$$

which can be computed efficiently by the recursion:

$$\hat{X}_i(k;t) = \hat{X}_i(k;t-1)e^{j(2\pi k/L)} + \hat{x}_i(t) - \hat{x}_i(t-L). \tag{12}$$

Since the second-order statistics of the DFT coefficients are not available, they can be approximated by

$$E\{\hat{X}_i(k;t)\hat{X}_j^*(k;t)\} \approx \frac{1}{2D} \sum_{u=-D}^{D-1} \hat{X}_i(k;t-u)\hat{X}_j^*(k;t-u) \tag{13}$$

where $2D$ is the size of the averaging window. Based on (13), $\hat{E}_1(t), \hat{E}_2(t)$, and $P(t)$ are computed, and the decision scores $\delta_1(t)$ and $\delta_2(t)$ are determined. The decision scores are further smoothed and thresholded to avoid the fragmentation of the speech utterances to facilitate automatic speech recognition.

A good value for the threshold T depends on the acoustic environment, and T is chosen experimentally for the studied environment in the current work. When the noise level is small or ignorable, the values of $P(t)$ differ significantly in the region where both sources are active and in the region where only one source is active. In this case, T can be chosen from a wider range of values while still resulting in similar performance. However, the difference of $P(t)$ in the two types of regions decreases as the noise level rises, and hence the choice of T becomes more critical. In this case, a longer averaging window can help reducing noise effect, and therefore improve the detection accuracy.

C. Performance

The following example illustrates the performance of the source detection algorithm. The two source signals from the TIMIT database were made to have different and yet partially overlapping active regions. The source signals, the acquired co-channel signals, and the AADF-restored signals in both channels are shown in Figs. 5–7, respectively. The signal-to-interference ratio (SIR) is about 10 dB in both channels before processing, and is about 22 dB after processing. After computing the length-32 DFT coefficients of the restored signals according to (12), two different sized averaging windows

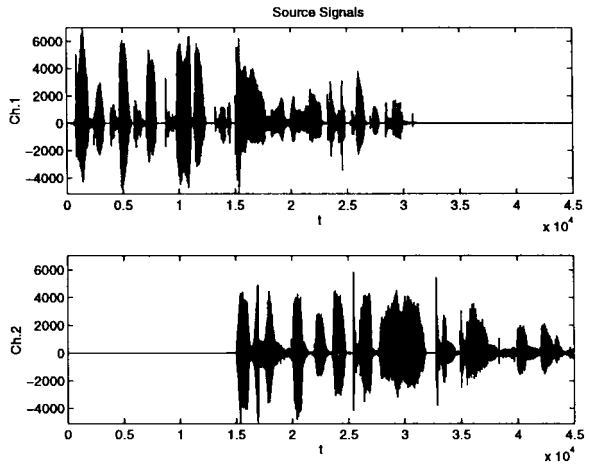

Fig. 5. Source signals—an example: the unit of t is in sample(s), and the sampling rate is 10.67 kHz.

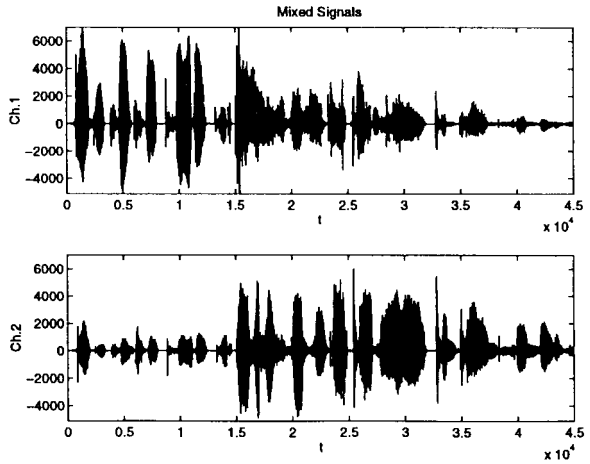

Fig. 6. Co-channel convolutive mixture of the two source signals in Fig. 5.

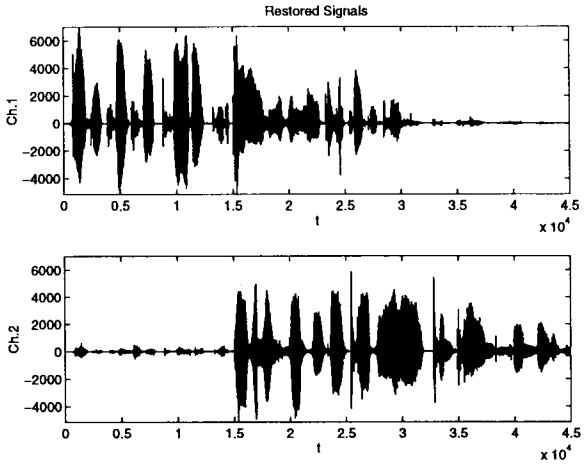

Fig. 7. AADF-restored signals of the mixed signals in Fig. 6.

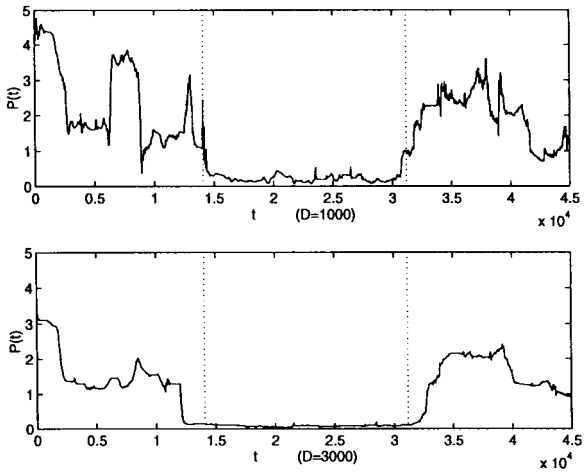

Fig. 8. $P(t)$ curves of the restored signals in Fig. 7. In the first curve, an averaging window of $D = 1000$ was used and in the second curve, a window of $D = 3000$ was used. The vertical dotted lines indicate the true boundaries between the active and inactive regions.

($D = 1000$ and 3000) were used in (13) to calculate $P(t)$, which are plotted in Fig. 8 for both cases.

As can be observed from Fig. 8, both $P(t)$ curves dropped toward zero shortly before the source signal 2 became active, and then rose back again shortly after the source signal 1 became inactive. There were certain excess regions before and after the true active regions being detected as active regions. These excess regions were ignorable for the shorter window, but were not ignorable for the longer window which introduced too much temporal smoothing. On the other hand, the $P(t)$ curve of the shorter window was noisier than that of the longer window where the latter reduced randomness in estimating the second order statistics of the short-time DFT coefficients. Therefore, the size of the averaging window needs to be chosen based on the trade-off between the temporal resolution of the detection boundary and the reliability of the estimate of the coherence function.

V. Algorithm Switching for Improved Filter Estimation

From experiments, we observed that the ADF and AADF algorithms worked well when the source signals in both channels were active; however, the algorithms encountered problems when either of the source signals became inactive. From the co-channel model of (1), if the source 2 is inactive, i.e., $x_2(t) = 0$ for an extended period of time, the system becomes

$$\begin{aligned} Y_1(f) &= X_1(f) \\ Y_2(f) &= B(f)X_1(f). \end{aligned} \tag{14}$$

Since there is no information of filter A in $y_1(t)$ and $y_2(t)$, the adaptive estimation based on the decorrelation of $y_1(t)$ and $y_2(t)$ may result in large estimation error of filter A while the estimate of filter B continues to converge. Therefore, when the source 2 is detected as inactive, which can be accomplished by the source detection algorithm discussed in Section IV, it is desired to stop the adaptive estimation of filter A and to continue the estimation of filter B. In such a case, an obvious choice for the adaptive estimation of filter B is the LMS algorithm of Widrow *et al.* [11].

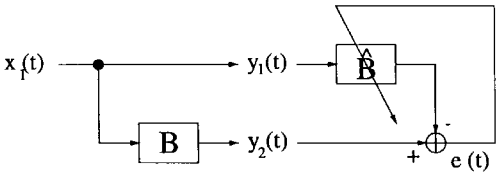

Fig. 9. Filter estimation model when one of the source signals is inactive. In this case, the Widrow's LMS algorithm can be applied.

A. The LMS Algorithm

Assume the source 2 as inactive and denote the filter B and its estimate at time t by $\underline{b}$ and $\underline{b}^{(t)}$, respectively. From (14), the relationship between $y_1(t)$ and $y_2(t)$ can be written as

$$y_2(t) = \underline{y}_1(t)^T \underline{b}$$

Define the error function as

$$\varepsilon(t) = y_2(t) - \underline{y}_1(t)^T . \underline{b}^{(t)}$$

The filter coefficients $\underline{b}$ can be estimated by minimizing $E\{\varepsilon^2(t)\}$. According to the LMS algorithm of Widrow *et al.* [11], $\underline{b}$ can be estimated iteratively by

$$\underline{b}^{(t+1)} = \underline{b}^{(t)} + 2\mu\varepsilon(t)\underline{y}_1(t). \tag{15}$$

The block diagram of the LMS algorithm is illustrated in Fig. 9.

Similarly, if the source 1 is inactive, $\underline{a}$ can be estimated iteratively by

$$\underline{a}^{(t+1)} = \underline{a}^{(t)} + 2\mu\zeta(t)\underline{y}_2(t) \tag{16}$$

where

$$\zeta(t) = y_1(t) - \underline{y}_2(t)^T \underline{a}^{(t)}.$$

B. Implementation and Performance

From the above discussion, depending on the outcome of source detection, the adaptive estimation algorithm is switched between the ADF or AADF and the LMS by the following rule:

Case 1: If both sources are active, update the filter coefficients according to (4) (for ADF) or (9) (for AADF).

Case 2: If the source 1 is active and the source 2 is inactive, update only the filter coefficients of B according to (15).

Case 3: If the source 2 is active and the source 1 is inactive, update only the filter coefficients of A according to (16).

Below is an example which shows the different behaviors, in terms of filter estimation error, of the AADF algorithm in two types of regions: both sources are active vs. only one source is active. In both types of regions, the estimation errors without and with the algorithm switching are compared. A pair of co-channel signals generated from the TIMIT database were processed using AADF with $\alpha = 0.5, N = 1000$, and $M = 200$. The SIR is about 10 dB in both channels. The normalized estimation errors (NEE) for both filters were

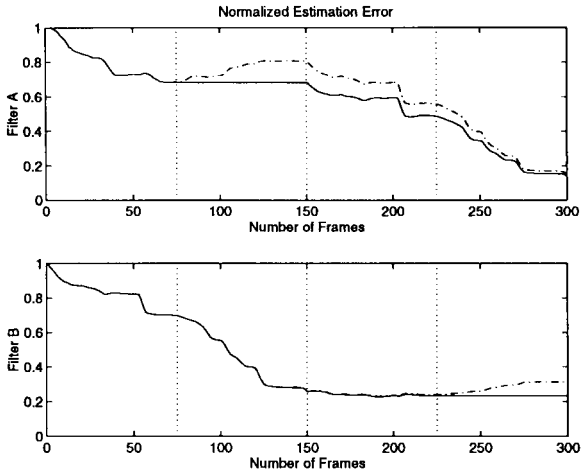

Fig. 10. Normalized estimation error, with $\underline{a}^{(0)} = \underline{0}$ and $\underline{b}^{(0)} = \underline{0}$. The source 2 is inactive in the time region 2, and the source 1 is inactive in the time region 4. The dashed curve is without algorithm switching and the solid curve is with algorithm switching.

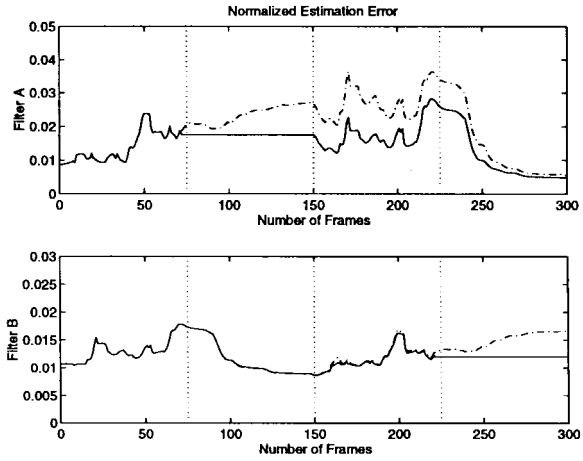

Fig. 11. The normalized estimation error, with $\underline{a}^{(0)} \approx \underline{a}^*$ and $\underline{b}^{(0)} \approx \underline{b}^*$. The source 2 is inactive in the time region 2, and the source 1 is inactive in the time region 4. The dashed curve is without algorithm switching and the solid curve is with algorithm switching.

defined as

$$E_A(t) = \frac{|\underline{a}^{(t)} - \underline{a}^*|^2}{|\underline{a}^*|^2}$$

$$E_B(t) = \frac{|\underline{b}^{(t)} - \underline{b}^*|^2}{|\underline{b}^*|^2}$$

and were measured after each frame of signals was processed. Two different initial conditions were tested. In the first condition, the filter coefficients were initialized as zeros to simulate the significant difference between the current filter estimates and the true filter coefficients. In the second condition, the filter coefficients were initialized to be close to the true filter coefficients. The NEE's for the two cases are plotted in Figs. 10 and 11, respectively, where the dashed curves represent the case of without algorithm switching, and the solid curves represent the case of with algorithm switching. According to the presence or absence of one of the source signals, the time axis of each plot is divided into four regions:

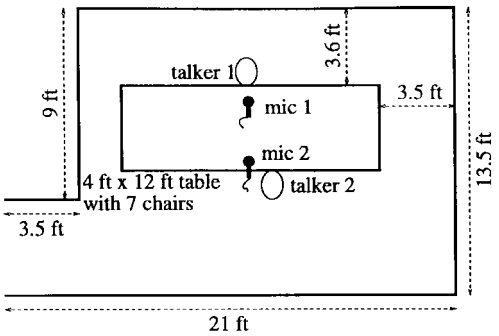

Fig. 12. Room-acoustic environment of House Ear Institute, Los Angeles, which was used for measuring the cross-channel acoustic paths.

Region 1: frames 1 through 75, both sources were active.
Region 2: frames 76 through 150, only the source 1 was active.
Region 3: frames 151 through 225, both sources were active.
Region 4: frames 226 through 300, only the source 2 was active.

As can be observed in Figs. 10 and 11, without the algorithm switching, the NEE of filter A in Region 2 and the NEE of filter B in Region 4 both increased, as compared to the NEE's with algorithm switching. The results demonstrate the effectiveness of the algorithm switching in preventing the filter estimates from drifting away from the true filters. The combined AADF and LMS based on the algorithm switching is referred to as the switched AADF/LMS (S-AADF/LMS).

VI. Experiments and System Performance

In the experiments, the co-channel signal separation techniques discussed in the previous sections were integrated with an ASR system and were evaluated under a simulated environment based on measurements made of a real room-acoustic condition. In this section, the acoustic environment is first described. Next, a strategy of accelerating estimation convergence for high-order finite impulse response (FIR) filters is discussed. The ASR system used in the experiment is then briefly described. Furthermore, the linear distortion removing part of the signal separation algorithm [specified by (3)] is not performed in the system, and the reason is discussed. Finally, the experimental results are summarized in terms of the SIR, the word recognition accuracy (WRA), and the word error rate before and after the adaptive co-channel processing.

A. Simulation of Acoustic Environment

The cross-channel acoustic paths were measured in the room environment described in Fig. 12 at the House Ear Institute, Los Angeles, and were represented by FIR filters A and B of order 200, where

Filter A: the acoustic path from talker 2 to microphone 1;
Filter B: the acoustic path from talker 1 to microphone 2.

In order to simplify the estimation problem, the channel distortions from talker 1 to microphone 1 and from talker 2 to microphone 2 were assumed negligible. This distortion will be addressed in the future work. The impulse responses of the two filters are plotted in Fig. 13, where the length of the filters were both 200 samples, that correspond to 18.7 ms at a sampling rate of 10 667 Hz. The frequency responses of the filters are also plotted in Fig. 14. The average attenuations introduced by the filters A and B are 10.34 dB and 11.23 dB, respectively. The filters A and B were used to generate the co-channel signals from the source signals according to (1).

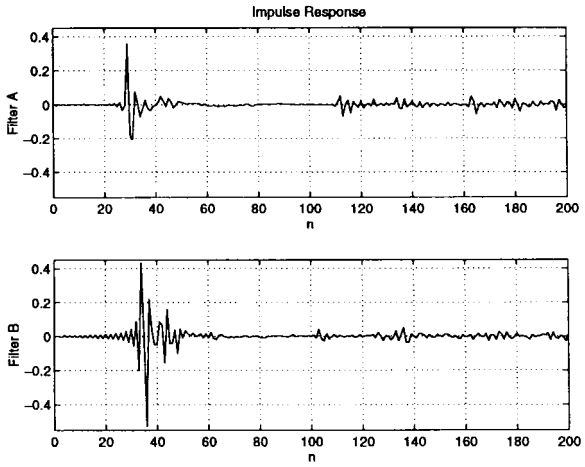

Fig. 13. Impulse responses of the cross-channel acoustic paths measured in Fig. 12: the filter A corresponds to the acoustic path from the talker 2 to the microphone 1, and the filter B corresponds to the acoustic path from the talker 1 to the microphone 2. The sample rate is 10.67 kHz.

B. Estimation of Reverberant Channels

From Fig. 13, most energy in each impulse response concentrates in a small number of samples with a short delay, which corresponds to the direct acoustic path. However, due to reverberation, the impulse response lasts for an extended period of time and hence requires much longer FIR filters in adaptive estimation. Long FIR filters not only require more computation, but also slow down the convergence of the filter estimates, and therefore deteriorate the system performance. Since the energy of an impulse response concentrates in the first 50 some samples, we experimented with a filter-order building-up scheme that starts the estimation with a low-order filter and then switch to a high-order filter when the low-order estimation was about to converge. This method significantly shortened the convergence time required for long filters and hence improved the system performance.

The following example shows the difference in the speed of convergence between using a low estimation order and a high estimation order. A set of co-channel signals were generated using the filters A and B as described in the above section, with the source signals chosen from the TIMIT database. The filter coefficients were initialized as zeros. The co-channel signals were processed using AADF with two choices of the filter orders N_a and N_b: the first choice was $N_a = N_b = 100$ and the second choice was $N_a = N_b = 200$. The NEE's for both cases are plotted in Fig. 15. The results show that although using $N_a = N_b = 200$ achieved lower NEE's eventually, using $N_a = N_b = 100$, the NEE's decreased much faster at the beginning stage of estimation even though half of the filter coefficients were not estimated.

TABLE I
SIGNAL-TO-INTERFERENCE RATIO, WORD RECOGNITION ACCURACY, AND WORD ERROR RATE BEFORE AND AFTER PROCESSING, WHEN ACOUSTIC NORMALIZATION WAS USED IN THE SICSR SYSTEM

The word recognition accuracy for source signals = 91.2 % (error rate = 8.8 %)

		Before Processing			After Processing		
Relative Source Energy Condition	Channel	SIR	WRA (Error Rate) Original	Extracted	SIR	WRA (Error Rate) Original	Extracted
Both sources heve the same energy	Ch. 1	10.34 dB	18.3 % (81.7 %)	36.3 % (63.7 %)	23.79 dB	64.1 % (35.9 %)	84.8 % (15.2 %)
	Ch. 2	11.23 dB	16.7 % (83.3 %)	32.8 % (67.2 %)	24.02 dB	62.5 % (37.5 %)	81.5 % (18.5 %)
Source 1 is 10 dB stronger	Ch. 1	20.34 dB	59.2 % (40.8 %)	79.4 % (20.6 %)	33.37 dB	68.9 % (31.1 %)	91.3 % (8.7 %)
	Ch. 2	1.23 dB	-18.9 % (118.9 %)	-9.1 % (109.1 %)	16.67 dB	45.8 % (54.2 %)	65.2 % (34.8 %)
Source 2 is 10 dB stronger	Ch. 1	0.34 dB	-20.6 % (120.6 %)	-10.7 % (110.7 %)	16.56 dB	43.1 % (56.9 %)	64.2 % (35.8 %)
	Ch. 2	21.23 dB	54.0 % (46.0 %)	74.8 % (25.2 %)	33.46 dB	71.7 % (28.3 %)	89.5 % (10.5 %)

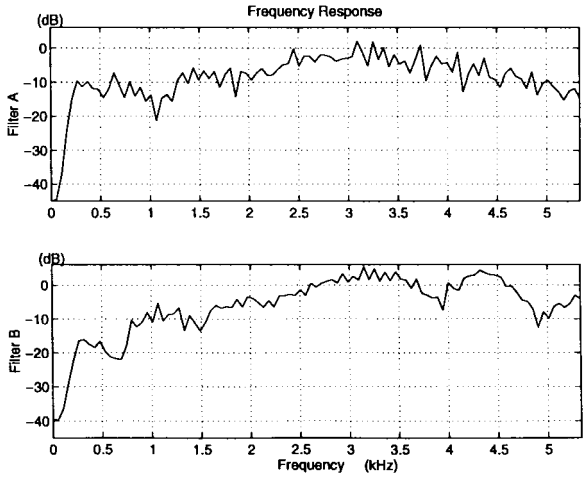

Fig. 14. Frequency responses of the cross-channel acoustic paths measured in Fig. 12.

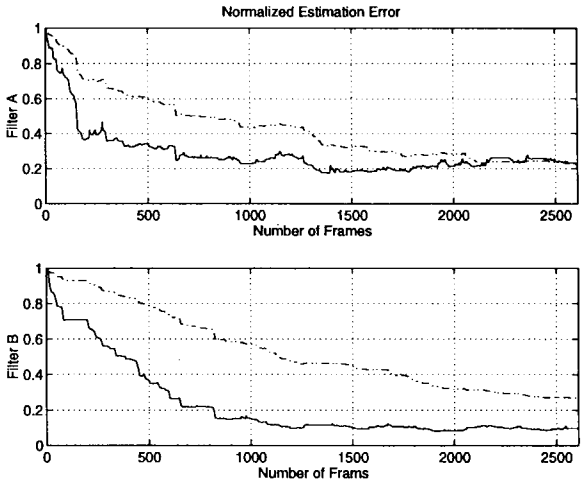

Fig. 15. Normalized estimation errors, with $\underline{a}^{(0)} = \underline{0}$ and $\underline{b}^{(0)} = \underline{0}$. The dashed curve corresponds to $N_a = N_b = 200$ and the solid curve corresponds to $N_a = N_b = 100$.

C. Automatic Speech Recognition

A speaker-independent continuous-speech recognition system was used to recognize the speech signals. The SICSR system is based on the HMM's of phone units: each phone-unit HMM has three tied-states; each state is modeled by a Gaussian mixture density. For each mixture density, the basis Gaussian densities are context-independent; the mixture weights are triphone context-dependent; the mixture size and the Gaussian density parameters were determined via a bottom-up merging algorithm. The phone models were trained from a subset of 717 sentences in the TIMIT database, with a total of 62 units as defined in the TIMIT (excluding "h#") [18]. The average mixture size per mixture density is approximately 19, and the total number of triphone contexts is about 4500. The TIMIT speech data were downsampled from 16 to 10.67 kHz. The cepstrum coefficients of the PLP analysis (eighth-order) and log energy were taken as instantaneous features, and their first-order 50-ms temporal regression coefficients as dynamic features. The recognition task has a vocabulary size of 853 and a grammar perplexity of 105. Further details of the background materials can be found in [16]. The recognition system allows an optional acoustic

TABLE II
SIGNAL-TO-INTERFERENCE RATIO, WORD RECOGNITION ACCURACY, AND WORD ERROR RATE BEFORE AND AFTER PROCESSING, WHEN ACOUSTIC NORMALIZATION WAS NOT USED IN THE SICSR SYSTEM

The word recognition accuracy for source signals = 88.6 % (error rate = 11.4 %)

		Before Processing			After Processing		
Relative Source Energy Condition	Channel	SIR	WRA (Error Rate) Original	Extracted	SIR	WRA (Error Rate) Original	Extracted
Both sources heve the same energy	Ch. 1	10.34 dB	27.7 % (72.3 %)	39.5 % (60.5 %)	23.79 dB	70.8 % (29.2 %)	76.9 % (23.1 %)
	Ch. 2	11.23 dB	21.2 % (78.8 %)	32.2 % (67.8 %)	24.02 dB	68.6 % (31.4 %)	72.2 % (27.8 %)
Source 1 is 10 dB stronger	Ch. 1	20.34 dB	63.2 % (36.8 %)	77.9 % (22.1 %)	33.37 dB	77.1 % (22.9 %)	83.2 % (16.8 %)
	Ch. 2	1.23 dB	-12.2 % (112.2 %)	-3.2 % (103.2 %)	16.67 dB	48.1 % (51.9 %)	58.1 % (41.9 %)
Source 2 is 10 dB stronger	Ch. 1	0.34 dB	-14.6 % (114.6 %)	-8.0 % (108.0 %)	16.56 dB	47.0 % (53.0 %)	60.4 % (39.6 %)
	Ch. 2	21.23 dB	62.9 % (37.1 %)	74.3 % (25.7 %)	33.46 dB	76.0 % (24.0 %)	81.1 % (18.9 %)

normalization on the cepstral features of test speech, where the acoustic normalization was based on cepstral bias estimation which was bootstrapped by cepstral mean estimation, and details can be found in [19].

D. Linear Distortion and Signal-to-Interference Ratio

In Section II, it is shown that the linear filter $C^{-1}(f)$ serves to transform $v_1(t)$ and $v_2(t)$ into $\hat{x}_1(t)$ and $\hat{x}_2(t)$ by removing the linear distortion $C(f)$. Similar to the treatment of $\hat{x}_i(t)$'s in (10), the $v_i(t)$'s can also be decomposed into the source parts and the leakage parts. It can be shown that the linear filter $C^{-1}(f)$ usually introduces larger gains on the leakage parts than on the source parts (see Appendix B for explanation), and therefore the improvement in removing distortion usually comes at the cost of reducing the SIR. Due to the built-in adaptation to channel distortion and the fact that the channel filters A and B do not have a strong distortion effect, the gain in removing distortion usually does not justify the loss in SIR. Therefore, the linear filtering in (3) is not performed in the system.

E. System Performance

To evaluate the effectiveness of the co-channel signal separation system, 156 sentence pairs were chosen from the TIMIT database as the source signals, and the filters A and B were used in generating the co-channel signals. The magnitudes of these sentences were scaled so that the SIR of the co-channel signals could be controlled. Three sets of co-channel signals were generated, with the relative source energy levels defined as the following:

Set 1: Both sources have the same energy;
Set 2: Source 1 is 10 dB stronger than source 2;
Set 3: Source 2 is 10 dB stronger than source 1.

The S-AADF/LMS with $\alpha = 0.5, N = 1000$, and $M = 200$ was used in the channel estimation. The source detection algorithm was performed using length-32 DFT, with $D = 1600$ and $T = 0.67$. Six sentence pairs were first processed with $N_a = N_b = 100$ and $\underline{a}^{(0)} = \underline{b}^{(0)} = \underline{0}$. Using the estimation results as the initial estimates, adaptive processing was made with the filter orders set to $N_a = N_b = 200$ on each set of co-channel signals. The separated signals within the detected active regions were then extracted and recognized by the SICSR system discussed in Section IV-C. The average SIR's and the WRA's for the three sets of co-channel signals before and after processing are summarized in Table I for the case of using acoustic normalization. The evaluation results without using acoustic normalization are summarized in Table II for comparison. The word recognition error rates (100%−WRA) are also included in the two tables inside the parentheses for reference. In both tables, the case of "extracted signals before processing" corresponds to extracting the mixed co-channel signals, $y_1(t)$ and $y_2(t)$, within the respective active regions of the source signals, $x_1(t)$ and $x_2(t)$, for the purpose of comparison with the extracted signals after processing. The following can be observed from Tables I and II.

1) Even a weak interference can deteriorate the recognition accuracy seriously.

2) In all cases, both SIR's and WRA's were improved significantly after processing.
3) The importance of the source signal detection can be shown by comparing the WRA's of the original and extracted signals, especially when the acoustic normalization technique is used.
4) The acoustic normalization used in the SICSR system can effectively handle the distortion introduced by the signal separation processing, and hence further improved recognition accuracy on the extracted signals after separation. On the other hand, the estimation of cepstral bias in acoustic normalization appears to be sensitive to the cross-channel interference, which led to decreased recognition accuracy on the unprocessed or unextracted speech data.
5) In Table I, except for the cases where the target source was 10 dB weaker than the interfering source, the WRA's of the extracted signals are all above 80%.

VII. Conclusion

In this paper, an improved technique of co-channel speech separation based on the algorithm switching between AADF and LMS is presented and is used as a front-end module for robust co-channel speech recognition. The S-AADF/LMS technique has been proven to be capable of reducing the cross-talk between simultaneously acquired co-channel speech signals. The AADF algorithm developed in the current work is based on the adaptive decorrelation filtering algorithm of Weinstein *et al.* and it improves the ADF in both aspects of system stability and efficiency. Specifically, an upper bound of adaptation gain has been derived for system stability; power normalization on the adaptation gain is developed to reduce the dependency of system stability on the input power level; a flattening sequence of adaptation gain has been used for accelerated convergence; and a filter order building-up scheme has been devised for the estimation of long FIR filters. A coherence-function based source signal detection algorithm has also been developed and is shown effective in determining the active/inactive regions of each signal source. The source detection algorithm is successfully used in the switching between the AADF and LMS algorithms which yielded better channel estimation performance than AADF alone, and in the extraction of speech from leakage-corrupted background which resulted in much higher word recognition accuracy of speech signals.

In order to handle more complicate acoustic environments, more effective channel estimation algorithms are needed and are critical to the improvement of both the SIR and the accuracy of automatic speech recognition. The interaction between the ASR system and the co-channel speech separation front-end should also be further studied.

Appendix A

The following equation is derived for updating the filter coefficients by expanding (4), replacing $\mu(t)$ by γ/t, and ignoring the quadratic terms of $\underline{a}$ and $\underline{b}$:

$$\underline{w}^{(t)} = \underline{w}^{(t-1)} + \frac{\gamma}{t}\,\underline{h}(t) - \frac{\gamma}{t}\,R(t)\underline{w}^{(t-1)} \tag{17}$$

where

$$\underline{w}^{(t)} = \begin{bmatrix} \underline{a}^{(t)} \\ \underline{b}^{(t)} \end{bmatrix}$$
$$\underline{h}(t) = \begin{bmatrix} \underline{y}_2(t)y_1(t) \\ \underline{y}_1(t)y_2(t) \end{bmatrix}$$
$$R(t) = \begin{bmatrix} \underline{y}_2(t)\underline{y}_2(t)^T & y_1(t)C_1(t) \\ y_2(t)C_2(t) & \underline{y}_1(t)\underline{y}_1(t)^T \end{bmatrix}$$

with

$$C_1(t) = [\underline{y}_1(t), \cdots, \underline{y}_1(t - N_a + 1)]^T$$
$$C_2(t) = [\underline{y}_2(t), \cdots, \underline{y}_2(t - N_b + 1)]^T.$$

Assuming that $R(t)$ and $\underline{h}(t)$ are stationary and independent of $\underline{w}^{(t)}$, and defining $\Psi = E\{R(t)\}$ and $\underline{\psi} = E\{\underline{h}(t)\}$, from (17), we have

$$E\{\underline{w}^{(t)}\} = E\{\underline{w}^{(t-1)}\} + \frac{\gamma}{t}\,\underline{\psi} + \frac{\gamma}{t}\,\Psi E\{\underline{w}^{(t-1)}\}. \tag{18}$$

If $E\{\underline{w}^{(t)}\} \to \underline{w}^*$ as $t \to \infty$, then

$$\underline{w}^* = \underline{w}^* + \frac{\gamma}{t}\,\underline{\psi} - \frac{\gamma}{t}\,\Psi\underline{w}^*$$
$$\Rightarrow \quad \underline{w}^* = \Psi^{-1}\underline{\psi}.$$

Substituting $\underline{w}^{(t)}$ by $\Delta\underline{w}(t) + \underline{w}^*$, (18) can be rewritten as

$$E\{\Delta\underline{w}(t)\} + \underline{w}^* = E\{\Delta\underline{w}(t-1)\} + \underline{w}^* + \frac{\gamma}{t}\,\underline{\psi} - \frac{\gamma}{t}\,\Psi E\{\Delta\underline{w}(t-1)\} - \frac{\gamma}{t}\,\Psi\underline{w}^*$$

i.e.,

$$E\{\Delta\underline{w}(t)\} = \left(I - \frac{\gamma}{t}\,\Psi\right) E\{\Delta\underline{w}(t-1)\}. \tag{19}$$

Defining $\underline{\varphi}(t) = E\{\Delta\underline{w}(t)\}$, (19) can then be rewritten as

$$\underline{\varphi}(t) = \left(I - \frac{\gamma}{t}\,\Psi\right)\underline{\varphi}(t-1).$$

Since Ψ is positive semidefinite, it can be decomposed as

$$\Psi = U^H \Lambda U$$

where U is unitary, the superscript H denotes Hermitian transpose, and

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_{N_a+N_b} \end{bmatrix}$$

with $\lambda_k \geq 0$ for $k = 1, 2, \cdots, N_a + N_b$. Therefore

$$\underline{\varphi}(t) = U^H\left(I - \frac{\gamma}{t}\,\Lambda\right) U\underline{\varphi}(t-1).$$

Define

$$\underline{\xi}(t) = U\underline{\varphi}(t)$$

then

$$\underline{\xi}(t) = \left(I - \frac{\gamma}{t}\,\Lambda\right)\underline{\xi}(t-1)$$

i.e.,

$$\xi_k(t) = \left(1 - \frac{\gamma}{t}\,\lambda_k\right)\xi_k(t-1), \qquad k = 1, 2, \cdots, N_a + N_b.$$

For convergence, it is necessary that

$$\|\underline{\varphi}(t)\| \to 0 \qquad t \to \infty.$$

Since U is unitary, it is equivalent to

$$\|\underline{\xi}(t)\| \to 0 \qquad \text{as} \quad t \to \infty$$

i.e.,

$$\xi_k(t) \to 0 \qquad \text{as} \quad t \to \infty$$

for $k = 1, 2, \cdots, N_a + N_b$. Hence

$$\begin{aligned} &\left|1 - \frac{\gamma}{t}\,\lambda_k\right| < 1 \\ \Rightarrow \quad &0 < \frac{\gamma}{t}\,\lambda_k < 2 \\ \Rightarrow \quad &0 < \gamma < \frac{2t}{\lambda_k}. \end{aligned} \tag{20}$$

Since (20) should be satisfied for all k and all t, the bounds of γ can be set as

$$0 < \gamma < 2/\lambda_{\max}$$

where

$$\lambda_{\max} = \max_{k=1,2,\cdots,N_a+N_b} \lambda_k.$$

Because $\text{trace}(E\{R(t)\}) = N_a \ \text{var}\{y_2(t)\} + N_b \ \text{var}\{y_1(t)\} \geq \lambda_{\max}$, (6) can be used for γ to avoid the calculation of the eigenvalues.

APPENDIX B

Following (1), (2), and (3), $V_1(k)$ and $\hat{X}_1(k)$ can be decomposed as

$$V_1(k) = [1 - \hat{A}(k)B(k)]X_1(k) + [A(k) - \hat{A}(k)]X_2(k) \tag{21}$$

$$\hat{X}_1(k) = \frac{1 - \hat{A}(k)B(k)}{1 - \hat{A}(k)\hat{B}(k)}\,X_1(k) + \frac{A(k) - \hat{A}(k)}{1 - \hat{A}(k)\hat{B}(k)}\,X_2(k). \tag{22}$$

Defining the following notations:

$$\begin{aligned} D_{V_1}(k) &= 1 - \hat{A}(k)B(k) \\ L_{V_1}(k) &= A(k) - \hat{A}(k) \\ D(k) &= 1 - \hat{A}(k)\hat{B}(k) \\ D_{\hat{X}_1}(k) &= \frac{1 - \hat{A}(k)B(k)}{1 - \hat{A}(k)\hat{B}(k)} = \frac{D_{V_1}(k)}{D(k)} \\ L_{\hat{X}_1}(k) &= \frac{A(k) - \hat{A}(k)}{1 - \hat{A}(k)\hat{B}(k)} = \frac{L_{V_1}(k)}{D(k)} \end{aligned}$$

and

$$\begin{aligned} \boldsymbol{D}_{V_1} &= \sum_{k=0}^{L-1} |D_{V_1}(k)|^2 \\ \boldsymbol{L}_{V_1} &= \sum_{k=0}^{L-1} |L_{V_1}(k)|^2 \\ \boldsymbol{D} &= \sum_{k=0}^{L-1} |D(k)|^2 \\ \boldsymbol{D}_{\hat{X}_1} &= \sum_{k=0}^{L-1} |D_{\hat{X}_1}(k)|^2 \\ \boldsymbol{L}_{\hat{X}_1} &= \sum_{k=0}^{L-1} |L_{\hat{X}_1}(k)|^2 \end{aligned}$$

equations (21) and (22) can be rewritten as

$$\begin{aligned} V_1(k) &= D_{V_1}(k)X_1(k) + L_{V_1}(k)X_2(k) \\ \hat{X}_1(k) &= D_{\hat{X}_1}(k)X_1(k) + L_{\hat{X}_1}(k)X_2(k) \end{aligned}$$

and the SIR's of $v_1(t)$ and $\hat{x}_1(t)$ are

$$\text{SIR}_{V_1} = \frac{\sum_{k=0}^{L-1} |D_{V_1}(k)|^2 |X_1(k)|^2}{\sum_{k=0}^{L-1} |L_{V_1}(k)|^2 |X_2(k)|^2}$$

and

$$\text{SIR}_{\hat{X}_1} = \frac{\sum_{k=0}^{L-1} |D_{\hat{X}_1}(k)|^2 |X_1(k)|^2}{\sum_{k=0}^{L-1} |L_{\hat{X}_1}(k)|^2 |X_2(k)|^2},$$

respectively. Since $x_1(t)$ and $x_2(t)$ are independent of the channel filters and assuming $X_1(k)$ and $X_2(k)$ are approximately flat, the good indicators for the SIR's of $v_1(t)$ and $\hat{x}_1(t)$ become $\boldsymbol{D}_{V_1}/\boldsymbol{L}_{V_1}$ and $\boldsymbol{D}_{\hat{X}_1}/\boldsymbol{L}_{\hat{X}_1}$, respectively. Define

$$\begin{aligned} d_{V_1}(k) &= \frac{|D_{V_1}(k)|^2}{\boldsymbol{D}_{V_1}} \\ l_{V_1}(k) &= \frac{|L_{V_1}(k)|^2}{\boldsymbol{L}_{V_1}} \\ d(k) &= \frac{|D(k)|^2}{\boldsymbol{D}} \end{aligned}$$

with

$$\sum_{k=0}^{L-1} d_{V_1}(k) = \sum_{k=0}^{L-1} l_{V_1}(k) = \sum_{k=0}^{L-1} d(k) = 1.$$

It can be shown that

$$\frac{\boldsymbol{D}_{\hat{X}_1}}{\boldsymbol{L}_{\hat{X}_1}} = \frac{\boldsymbol{D}_{V_1}}{\boldsymbol{L}_{V_1}}\,\frac{\boldsymbol{G}_D}{\boldsymbol{G}_L}$$

where

$$\boldsymbol{G}_D = \sum_{k=0}^{L-1} \frac{d_{V_1}(k)}{d(k)}$$
$$\boldsymbol{G}_L = \sum_{k=0}^{L-1} \frac{l_{V_1}(k)}{d(k)}.$$

Given fixed $\hat{A}(k)$ and $\hat{B}(k)$, since $d(k)$ only depends on $\hat{A}(k)$ and $\hat{B}(k)$, the expected values for $\boldsymbol{G}_D$ and $\boldsymbol{G}_L$ can be expressed as

$$E\{\boldsymbol{G}_D\} = \sum_{k=0}^{L-1} \frac{E\{d_{V_1}(k)\}}{d(k)}$$
$$E\{\boldsymbol{G}_L\} = \sum_{k=0}^{L-1} \frac{E\{l_{V_1}(k)\}}{d(k)}.$$

If $A(k) \approx \hat{A}(k)$ and $B(k) \approx \hat{B}(k)$,

$$E\{d_{V_1}(k)\} \approx d(k)$$
$$\Rightarrow \quad E\{\boldsymbol{G}_D\} \approx L.$$

Furthermore, since the filter estimation errors $A(k) - \hat{A}(k)$ is white and independent of $\hat{A}(k)$ and $\hat{B}(k), l_{V_1}(k)$'s can be assumed i.i.d. Therefore

$$E\{l_{V_1}(0)\} = E\{l_{V_1}(1)\} = \cdots = E\{l_{V_1}(L-1)\} = \frac{1}{L}$$
$$\Rightarrow E\{\boldsymbol{G}_L\} = \frac{1}{L} \sum_{k=0}^{L-1} \frac{1}{d(k)}$$
$$= \frac{1}{L} \left\{ \frac{1}{1 - \sum_{k=1}^{L-1} d(k)} + \sum_{k=1}^{L-1} \frac{1}{d(k)} \right\}.$$

In order to find the minimum value of $E\{\boldsymbol{G}_L\}$, we take the derivatives of $E\{\boldsymbol{G}_L\}$ w.r.t. $d(k)$'s:

$$\frac{\partial E\{\boldsymbol{G}_L\}}{\partial d(k)} = \frac{1}{L} \left\{ \frac{1}{\left[1 - \sum_{l=1}^{L-1} d(l)\right]^2} - \frac{1}{d^2(k)} \right\}$$

and solve for the equations

$$\frac{\partial E\{\boldsymbol{G}_L\}}{\partial d(1)} = \frac{\partial E\{\boldsymbol{G}_L\}}{\partial d(2)} = \cdots \frac{\partial E\{\boldsymbol{G}_L\}}{\partial d(L-1)} = 0$$

which yields

$$d(k) = 1 - \sum_{l=1}^{L-1} d(l), \quad k = 1, \cdots, L-1.$$

The following linear equations can be derived from the above relation and must be satisfied by $d(k)$:

$$\begin{bmatrix} 2 & 1 & \cdots & 1 \\ 1 & 2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 1 & \cdots & 1 & 2 \end{bmatrix} \begin{bmatrix} d(1) \\ d(2) \\ \vdots \\ d(L-1) \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}.$$

The solution is

$$d(0) = d(1) = \cdots = d(L-1) = \frac{1}{L}.$$

From the above derivation, we see that

$$\min_{d(0),\cdots,d(L-1)} E\{\boldsymbol{G}_L\} = E\{\boldsymbol{G}_L\}\Big|_{d(0)=\cdots=d(L-1)=(1/L)} = L$$

i.e.,

$$E\{\boldsymbol{G}_L\} \geq L \approx E\{\boldsymbol{G}_D\}.$$

As such, the gain on the leakage part is usually larger than the gain on the signal part, hence the SIR of $\hat{x}_1(t)$ is usually lower than the SIR of $v_1(t)$. The same relation holds between the SIR's of $v_2(t)$ and $\hat{x}_2(t)$ and it can be derived in the same way.

ACKNOWLEDGMENT

The measurement of room acoustics provided by Dr. S. Soli of House Ear Institute, Los Angeles, CA, and the suggestion by S. Wang of the Beckman Institute, UIUC, are also acknowledged.

REFERENCES

[1] R. Cole *et al.*, "The challenge of spoken language systems: Research directions for the nineties," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 1–21, Jan. 1995.
[2] B.-H. Juang, "Speech recognition in adverse environments," *Comput. Speech Lang.*, May 1991, pp. 275–294.
[3] A. Nadas, D. Nahamoo, and M. A. Picheny, "Speech recognition using noise-adaptive prototypes," *IEEE Trans. Speech Audio Processing*, vol. 37, pp. 1495–1503, Oct. 1989.
[4] M. J. F. Gales and S. J. Young, "An improved approach to the hidden Markov model decomposition," in *Proc. ICASSP*, 1992, pp. 729–734.
[5] A. Varga and R. K. Moore, "Simultaneous recognition of concurrent speech signals using hidden Markov model decomposition," in *Proc. EuroSpeech*, 1991, pp. 1175–1178.
[6] R. C. Rose, E. M. Hofstetter, and D. A. Reynolds, "Integrated models of signal and background with application to speaker identification in noise," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 245–258, Apr. 1994.
[7] E. Toner and D. R. Campbell, "Speech enhancement using sub-band intermittent adaptation," *Speech Commun.*, vol. 12, pp. 253–259, 1993.
[8] D. R. Campbell, "Sub-band adaptive speech enhancement for hearing aids," in *Proc. ICSLP*, Oct. 1996, vol. 1, pp. 180–183.
[9] P. W. Shields and D. R. Campbell, "Multi-microphone sub-band adaptive signal processing for improvement of hearing aid performance: Preliminary results using normal hearing volunteers," in *Proc. ICASSP*, Apr. 1997, vol. 1, pp. 415–418,
[10] R. H. Lambert and A. J. Bell, "Blind separation of multiple speakers in a multipath environment," in *Proc. ICASSP*, Apr. 1997, vol. 1, pp. 423–426.
[11] B. Widrow *et al.*, "Adaptive noise cancelling: Principles and applications," *Proc. IEEE*, vol. 63, pp. 1692–1716, Dec. 1975.
[12] E. Weinstein, M. Feder, and A. V. Oppenheim, "Multi-channel signal separation by decorrelation," *IEEE Trans. Speech Audio Processing*, vol. 1, pp. 405–413, Oct. 1993.
[13] S. Van Gerven and D. Van Compernolle, "Signal separation by symmetric adaptive decorrelation: Stability, convergence, and uniqueness," *IEEE Trans. Signal Processing*, vol. 43, pp. 1602–1612, July 1995.
[14] K. Yen and Y. Zhao, "Robust automatic speech recognition using a multi-channel signal separation front-end," in *Proc. ICSLP*, Oct. 1996, vol. 3, pp. 1337–1340.
[15] ——, "Co-channel speech separation for robust automatic speech recognition: Stability and efficiency," in *Proc. ICASSP*, Apr. 1997, vol. 2, pp. 859–862.
[16] Y. Zhao, "A speaker-independent continuous speech recognition system using continuous mixture Gaussian density HMM of phoneme-sized units," *IEEE Trans. Speech Audio Processing*, vol. 1, pp. 345–361, July 1993.

[17] H. Kesten, "Accelerated stochastic approximation," *Ann. Math. Stat.*, pp. 41–59, 1958.
[18] L. F. Lamel, R. H. Kassel, and S. Seneff, "Speech database development: Design and analysis of the acoustic-phonetic corpus," in *Proc. Speech Recognition Workshop (DARPA)*, 1986.
[19] Y. Zhao, "Self-learning speaker and channel adaptation based on spectral variation source decomposition," *Speech Commun.*, vol. 18, pp. 65–77, Jan. 1996.

Kuan-Chieh Yen received the B.S. degree in electrical engineering from National Taiwan University, Taipei, Taiwan, R.O.C., in 1990, and the M.S.E.E. degree from Purdue University, West Lafayette, IN, in 1994.

He is currently a Research Assistant at Beckman Institute and the Department of Electrical and Computer Engineering, University of Illinois, Urbana-Champaign, where he is working toward the Ph.D. degree in electrical engineering.

Yunxin Zhao (S'86–M'88–SM'94) received the Ph.D. degree in 1988 from University of Washington, Seattle.

She was Senior Research Staff and Project Leader of the Speech Technology Laboratory, Panasonic Technologies Inc., from 1988 to 1994. She was Assistant Professor in the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign (UIUC), from 1994 to 1998. She is currently Associate Professor with the Department of Computer Engineering and Computer Science, University of Missouri, Columbia. Her research interests are in spoken language processing, multimedia interface, multimodal human–computer interaction, statistical pattern recognition, statistical blind identification and estimation, speech and signal processing, and biomedical applications.

Dr. Zhao currently serves as an associate editor for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING. She received a 1995 NSF Career Award, and is listed in *American Men and Women of Science*.