IMPROVEMENTS ON CO-CHANNEL SPEECH SEPARATION USING ADF: LOW COMPLEXITY, FAST CONVERGENCE, AND GENERALIZATION

Kuan-Chieh Yen *Yunxin Zhao*

Beckman Institute and Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA
yen@ifp.uiuc.edu *yxz@ifp.uiuc.edu*

ABSTRACT

Three modifications on the adaptive decorrelation filtering (ADF) algorithm are proposed to improve the performance of a co-channel speech separation system. Firstly, a simplified ADF (SADF) is suggested to reduce the computational complexity of ADF from $O(N^2)$ to $O(N)$ per sample, where N is the filter length used in the channel estimation. Secondly, a transform-domain ADF (TDADF) is developed to accelerate the convergence of the filter estimates while maintaining computational complexity at $O(N)$. Thirdly, a generalized ADF (GADF) is derived to handle the noncausal filter estimation problem often encountered in co-channel speech separation. Experimental results showed that when the average signal-to-interference ratios (SIRs) in the co-channel signals were 6.15 and 5.38 dB, respectively, both the SADF and TDADF improved the SIRs to around 18 to 19 dB, and the GADF further improved the SIRs to around 19 to 24 dB.

1. INTRODUCTION

The state-of-the-art techniques in automatic speech recognition (ASR) are still vulnerable in the presence of interferences [1]. One of the difficult problems is the interference speech from competing talkers, or even worse, if the talkers are moving around. In these scenarios, robust speech recognition remains a challenging task.

In our recent work [2][3], adaptive decorrelation filtering (ADF) [4][5] was used as the core of a signal-separation front-end for improving the signal-to-interference ratio (SIR) in the input speech to an ASR system. In this scheme, two coexistent and independent speech sources are considered, and their convolutive mixtures are acquired via two microphones. Our experiments in [2] and [3] showed satisfactory improvements on both SIR and recognition accuracy when the distortion introduced by the acoustic paths between each microphone and its targeted source was neglectable. However, the system performance was seriously deteriorated when such distortion became significant. In addition, the computational complexity of ADF is $O(N^2)$, which prohibits real-time implementation of the co-channel speech separation system. In the current work, several key improvements on the ADF algorithm are made which yields a faster converging, $O(N)$ algorithm, and the algorithm is generalized for the estimation of noncausal long filters.

This paper is organized into six sections. In Section 2, the co-channel speech separation system using ADF are briefly described. Then a modification on ADF is proposed to reduce the computational complexity of ADF. In Section 3, the ADF is formulated in the Fourier transform domain improve the efficiency of the system. In Section 4, an generalized ADF is derived to refine the separation performance. Experimental results are presented in Section 5 and a conclusion is made in Section 6.

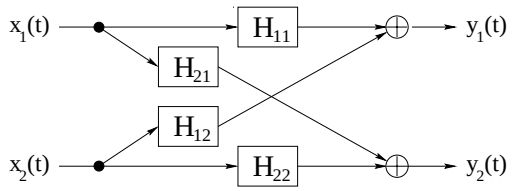

Figure 1. Block diagram of the co-channel system

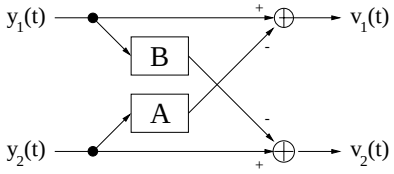

Figure 2. Block diagram of the signal separation system in Eq. (2)

2. CO-CHANNEL SYSTEM AND SIGNAL SEPARATION

This section starts with the description of the co-channel speech separation system. The background of the ADF algorithm is then given briefly. A modification on the ADF algorithm to reduce the computational complexity is proposed at the end of the section.

2.1. Co-Channel Speech Acquisition System

In a co-channel speech acquisition system, each microphone acquires not only its target signal, but also the interference signals from the other source. 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 H_{ij}, i, j =1 or 2, to model the acoustic paths from the source j to the microphone i, the co-channel system can be described in the frequency domain as

$$\begin{aligned} Y_1(f) &= H_{11}(f)X_1(f) + H_{12}(f)X_2(f) \\ Y_2(f) &= H_{22}(f)X_2(f) + H_{21}(f)X_1(f) \end{aligned} \quad (1)$$

The block diagram of the co-channel system is illustrated in Fig. 1.

2.2. Signal Separation by Adaptive Decorrelation Filtering

If the distortion and delay introduced by the acoustic paths between each microphone and its targeted source is neglectable, i.e. $H_{11}(f)$=$H_{22}(f)$=1, it was shown in [4] that

the signals from sources 1 and 2 can be separated out from $y_1(t)$ and $y_2(t)$ into $v_1(t)$ and $v_2(t)$ as

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

with $A(f) = H_{12}(f)$ and $B(f) = H_{21}(f)$. The block diagram of this system is illustrated in Fig. 2. It was also shown that the filters H_{12} and H_{21} can be estimated by the ADF algorithm:

$$\begin{aligned} \underline{a}^{(t)} &= \underline{a}^{(t-1)} + \mu(t)\underline{v}_2^{(t-1)}(t)v_1^{(t-1)}(t) \\ \underline{b}^{(t)} &= \underline{b}^{(t-1)} + \mu(t)\underline{v}_1^{(t-1)}(t)v_2^{(t-1)}(t) \end{aligned} \tag{3}$$

where the vectors $\underline{a}^{(t)}$ of length N_a and $\underline{b}^{(t)}$ of length N_b are the estimates of the coefficients of filters A and B at time t, with $\underline{a}^{(t)} = [a^{(t)}(0), \cdots, a^{(t)}(N_a-1)]^T$ and $\underline{b}^{(t)} = [b^{(t)}(0), \cdots, b^{(t)}(N_b-1)]^T$; $\mu(t)$ is a chosen adaptation gain, 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 $\underline{b}^{(t-1)}$:

$$\begin{aligned} v_1^{(t-1)}(\tau) &= y_1(\tau) - \underline{y}_2(\tau)^T\underline{a}^{(t-1)} \\ v_2^{(t-1)}(\tau) &= y_2(\tau) - \underline{y}_1(\tau)^T\underline{b}^{(t-1)} \end{aligned} \tag{4}$$

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

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

When the effects of H_{11} and H_{22} become significant, Eq. (1) can be rewritten as

$$\begin{aligned} Y_1(f) &= X_1'(f) + H_{22}^{-1}(f)H_{12}(f)X_2'(f) \\ Y_2(f) &= X_2'(f) + H_{11}^{-1}(f)H_{21}(f)X_1'(f) \end{aligned}$$

where $X_1'(f) = H_{11}(f)X_1(f)$ and $X_2'(f) = H_{22}(f)X_2(f)$, and hence the signals from the two sources can still be separated using the system described by Eq. (2) with $A(f) = H_{22}^{-1}(f)H_{12}(f)$ and $B(f) = H_{11}^{-1}(f)H_{21}(f)$. As a result, when applying the ADF in this case, $H_{22}^{-1}H_{12}$ and $H_{11}^{-1}H_{21}$ are estimated instead of H_{12} and H_{21}.

2.3. Simplified Adaptive Decorrelation Filtering

From the discussion above, the estimation of filters $H_{22}^{-1}H_{12}$ and $H_{11}^{-1}H_{21}$ are usually required for speech separation. Even if all the H_{ij}'s are short FIR filters, $H_{22}^{-1}H_{12}$ and $H_{11}^{-1}H_{21}$ become IIR filters. As a result, large N_a and N_b are usually necessary to achieve satisfactory separation results. From Eqs. (3) and (4), the computational complexity of ADF is $O(N_a^2 + N_b^2)$ per sample. Therefore, the required computations increase significantly as N_a and N_b increase.

To simplify the algorithm, Eqs. (3) and (4) can be modified as

$$\begin{aligned} \underline{a}^{(t)} &= \underline{a}^{(t-1)} + \mu(t)\underline{v}_2(t)v_1(t) \\ \underline{b}^{(t)} &= \underline{b}^{(t-1)} + \mu(t)\underline{v}_1(t)v_2(t) \end{aligned} \tag{5}$$

and

$$\begin{aligned} v_1(t) &= y_1(t) - \underline{y}_2(t)^T\underline{a}^{(t-1)} \\ v_2(t) &= y_2(t) - \underline{y}_1(t)^T\underline{b}^{(t-1)} \end{aligned} \tag{6}$$

where in each adaptation step t, instead of recomputing the entire $\underline{v}_1(t)$ and $\underline{v}_2(t)$ vectors according to the filter estimates $\underline{a}^{(t-1)}$ and $\underline{b}^{(t-1)}$, only $v_1(t)$ and $v_2(t)$ are computed. This modification reduces the computational complexity from $O(N_a^2 + N_b^2)$ to $O(N_a + N_b)$, and the modified algorithm is referred to as simplified ADF (SADF). Compared to the original ADF, SADF is more stable in convergence due to its slower propagation of errors than in the original ADF. The only drawback of SADF is the slower convergence rate, and it becomes obvious when the filters are very long.

3. TRANSFORM-DOMAIN ADAPTIVE DECORRELATION FILTERING

As discussed in Section 2.2, the ADF-based separation of co-channel speech signals in general requires the estimation of long filters. As the lengths of filters increase, the adaptation gains must be smaller in order to ensure system stability [3], which slows down the convergence of filter estimates. In addition, the increased interaction between filter coefficients can further slow down the convergence. However, fast convergence is key to the performance of the co-channel speech separation system when the channels are time-varying. In previous works, a transform-domain LMS algorithm [6][7][8] was shown capable of improving the efficiency of LMS algorithm by signal transformation. Assuming $N_a = N_b = N$ and applying the same idea on the SADF, a transform-domain ADF (TDADF) algorithm can be derived from the ADF algorithm by transforming the vectors from the time domain to the Fourier-transform domain as

$$\begin{aligned} \underline{A}^{(t)} &= \underline{A}^{(t-1)} + \mu\Lambda^{-1}(t)\underline{V}_2(t)v_1(t) \\ \underline{B}^{(t)} &= \underline{B}^{(t-1)} + \mu\Lambda^{-1}(t)\underline{V}_1(t)v_2(t) \end{aligned} \tag{7}$$

where the vectors $\underline{V}_1(t)$ and $\underline{V}_2(t)$ are the DFTs of $\underline{v}_1(t)$ and $\underline{v}_2(t)$ in Eq. (5):

$$\begin{aligned} \underline{V}_1(t) &= F\underline{v}_1(t) \\ \underline{V}_2(t) &= F\underline{v}_2(t) \end{aligned} \tag{8}$$

The matrix F is the unitary DFT matrix of size (NxN). In addition, μ is a chosen adaptation gain, and $\Lambda(t)$ is a diagonal normalization matrix defined as

$$\Lambda(t) = diag\left[\lambda_0(t), \cdots, \lambda_{N-1}(t)\right]$$

with the diagonal elements of the matrix equal to the sum of the estimated variances of the DFT coefficients of $\underline{v}_1(t)$ and $\underline{v}_2(t)$:

$$\lambda_k(t) = \alpha\lambda_k(t-1) + (1-\alpha)\left[|V_{1,k}(t)|^2 + |V_{2,k}(t)|^2\right]$$

where $V_{i,k}(t)$, $i = 1, 2$ is the k-th entry of the vector $\underline{V}_i(t)$, and α is a forgetting factor satisfying $0 < \alpha < 1$. With this normalization, the convergence rate of the filter coefficients can be improved significantly.

The estimated vectors $\underline{A}^{(t)}$ and $\underline{B}^{(t)}$ in Eq. (7) are equivalent to the DFTs of the vectors $\underline{a}^{(t)}$ and $\underline{b}^{(t)}$. However, by defining the vectors

$$\begin{aligned} \underline{Y}_1(t) &= F\underline{y}_1(t) \\ \underline{Y}_2(t) &= F\underline{y}_2(t) \end{aligned} \tag{9}$$

the $v_1(t)$ and $v_2(t)$ defined in Eq. (6) can be calculated equivalently by

$$\begin{aligned} v_1(t) &= y_1(t) - \underline{Y}_2(t)^H\underline{A}^{(t-1)} \\ v_2(t) &= y_2(t) - \underline{Y}_1(t)^H\underline{B}^{(t-1)} \end{aligned} \tag{10}$$

and hence the computation of inverse DFTs of $\underline{A}^{(t)}$ and $\underline{B}^{(t)}$ can be avoided. The computational complexity of length-N DFT is $O(NlogN)$ by FFT. However, the entries of the

DFT vectors $\underline{Y}_1(t)$, $\underline{Y}_2(t)$, $\underline{V}_1(t)$, and $\underline{V}_2(t)$ can be updated efficiently by

$$\begin{aligned} Y_{1,k}(t) &= W_N^k Y_{1,k}(t-1) + [y_1(t) - y_1(t-N_b)]/\sqrt{N} \\ Y_{2,k}(t) &= W_N^k Y_{2,k}(t-1) + [y_2(t) - y_2(t-N_a)]/\sqrt{N} \\ V_{1,k}(t) &= W_N^k V_{1,k}(t-1) + [v_1(t) - v_1(t-N_b)]/\sqrt{N} \\ V_{2,k}(t) &= W_N^k V_{2,k}(t-1) + [v_2(t) - v_2(t-N_a)]/\sqrt{N} \end{aligned}$$

where $W_N^k = exp\left(-j\frac{2\pi k}{N}\right)$. This reduces the complexity of computing the length-N DFT vectors to $O(N)$. Therefore, the computational complexity of TDADF stays at $O(N)$.

4. GENERALIZED ADAPTIVE DECORRELATION FILTERING

From the discussion in Section 2.2, estimating IIR filters $H_{22}^{-1}H_{12}$ and $H_{11}^{-1}H_{21}$ is often needed for signal separation. While increasing N_a and N_b enables the FIR filters A and B to approximate $H_{22}^{-1}H_{12}$ and $H_{11}^{-1}H_{21}$ better, longer filters alone cannot achieve satisfactory estimation results when the noncausal parts of $H_{22}^{-1}H_{12}$ and $H_{11}^{-1}H_{21}$ become significant since the ADF algorithm is designed for the estimation of causal filters. In this section, an alternative adaptive filtering configuration is proposed to improve the accuracy of acoustic channel estimation under this situation.

In certain applications, the relative locations between the microphones and their respective targeted sources are fixed, and therefore the acoustic channels H_{11} and H_{22} can be known. In this case, the two signals from the two sources can be separated by

$$\begin{aligned} V_1(f) &= H_{22}(f)Y_1(f) - A(f)Y_2(f) \\ V_2(f) &= H_{11}(f)Y_2(f) - B(f)Y_1(f) \end{aligned} \tag{11}$$

with $A(f) = H_{12}(f)$ and $B(f) = H_{21}(f)$. Since the filters $H_{12}(f)$ and $H_{21}(f)$ are causal, they can be estimated more accurately. By modifying Eq. (6) as

$$\begin{aligned} v_1(t) &= z_1(t) - \underline{y}_2(t)^T \underline{a}^{(t-1)} \\ v_2(t) &= z_2(t) - \underline{y}_1(t)^T \underline{b}^{(t-1)} \end{aligned} \tag{12}$$

where $z_1(t) = H_{22}\{y_1(t)\}$ and $z_2(t) = H_{11}\{y_2(t)\}$, the channel filters H_{12} and H_{21} can be estimated by Eqs. (5) and (12).

By generalization, even if H_{11} and H_{22} are unknown, the signals from different sources can be separated by introducing two time-delay filters, D_1 and D_2, as

$$\begin{aligned} V_1(f) &= D_1(f)Y_1(f) - A(f)Y_2(f) \\ V_2(f) &= D_2(f)Y_2(f) - B(f)Y_1(f) \end{aligned} \tag{13}$$

with $A(f) = D_1(f)H_{22}^{-1}(f)H_{12}(f)$ and $B(f) = D_2(f)H_{11}^{-1}(f)H_{21}(f)$. The filters D_1 and D_2 shift the impulse responses of $H_{22}^{-1}H_{12}$ and $H_{11}^{-1}H_{21}$ to the right so that the filters to be estimated have less noncausal components. It should be noted that in this case the $z_1(t)$ and $z_2(t)$ in Eq. (12) need to be modified as $z_1(t) = D_1\{y_1(t)\}$ and $z_2(t) = D_2\{y_2(t)\}$. The block diagram of this generalized separation system is illustrated in Fig. 3.

The generalized ADF (GADF) algorithm discussed above is sensitive to the initial values of the filter estimates $\underline{a}^{(0)}$ and $\underline{b}^{(0)}$. Therefore, it requires a good initial condition. Practically, the SADF can be used to obtain a preliminary estimates $A_p(f)$ and $B_p(f)$, then Eqs. (5) and (12) can be used to improve separation using $A(f) = D_1(f)A_p(f)$ and $B(f) = D_2(f)B_p(f)$ as the initial condition.

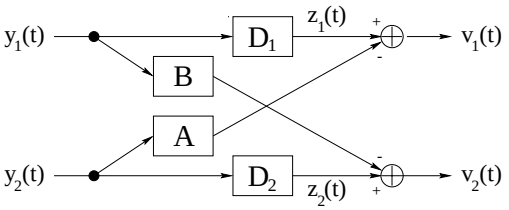

Figure 3. Block diagram of the signal separation system in Eq. (13)

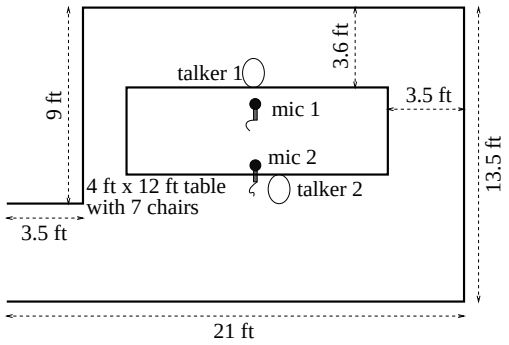

Figure 4. Room acoustic environment used in measuring the acoustic paths.

5. EXPERIMENTS

In this section, several experiments are presented to demonstrate the algorithm improvements discussed in the previous sections. In the experiments, the speech signals from TIMIT database were used as the source signals $x_1(t)$ and $x_2(t)$. The acoustic environment simulated in the experiments is first described. An experiment comparing the convergence rates of ADF, SADF and TDADF then follows. Finally, the performance of SADF and TDADF with various filter lengths are compared and the performance of the GADF algorithm is also evaluated.

5.1. Simulation of Acoustic Environments

The acoustic paths from the talker j (j=1 or 2) to the microphone i (i=1 or 2) were measured in the room environment described in Fig. 4 at the House Ear Institute, Los Angeles, and were represented by FIR filters H_{ij} of length 200. The sampling rate was 10.67 kHz. The four filters were used to generate the co-channel convolutive mixture signals from the source signals. The impulse responses of $H_{22}^{-1}H_{12}$ and $H_{11}^{-1}H_{21}$ are noncausal and have infinite lengths, as shown in Fig. 5.

5.2. Convergence Rate

To compare the convergence rate of the ADF, SADF and TDADF, a pair of co-channel signals were processed by all three algorithm, respectively. The normalized estimation errors (NEE) were defined as

$$E(t) = \frac{\left|\underline{a}^{(t)} - \underline{a}^*\right|^2}{\left|\underline{a}^*\right|^2} + \frac{\left|\underline{b}^{(t)} - \underline{b}^*\right|^2}{\left|\underline{b}^*\right|^2}$$

and were recorded for each step of adaptation. The filter lengths were 200 and the initial estimates were set to zeros for all three algorithms. The NEE's for the three algorithms are plotted in Fig. 6. It can be observed that the difference between ADF and SADF is insignificant, and the NEEs decreased much faster with TDADF. This phenomenon is typical in the processing of all co-channel signals in the experiment.

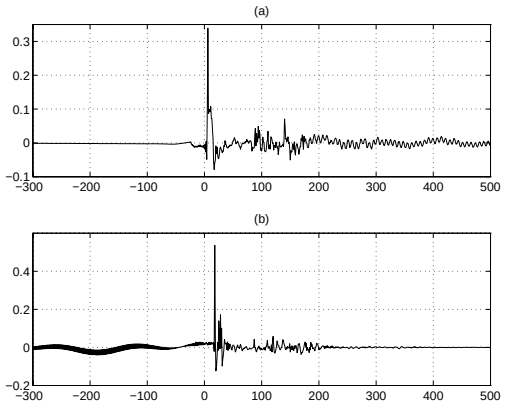

Figure 5. Impulse responses of (a) $H_{22}^{-1}H_{12}$ and (b) $H_{11}^{-1}H_{21}$ measured in the acoustic environment in Fig. 4.

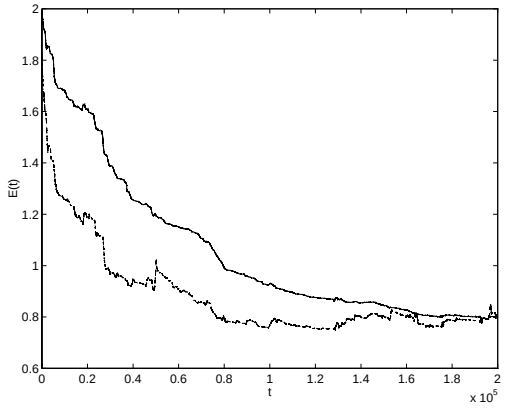

Figure 6. NEEs of ADF (solid curve), SADF (dotted curve just above the solid curve) and TDADF (dashed curve)

5.3. Separation Performance

In this section, a subset of TIMIT database was chosen to form a set of source signals of 156 sentence-pairs. The co-channel mixed signals were generated from the source signals using the filters H_{ij}'s described in Section 5.1 as in Eq. (1). The average SIRs in $y_1(t)$ and $y_2(t)$ were 6.15 dB and 5.38 dB, respectively. The SIRs after separation (in $v_1(t)$ and $v_2(t)$) using different filter lengths and algorithms (SADF and TDADF) are summarized in Table 1. It can be seen that the performances of SADF and TDADF were about the same since the channel was stationary. The SIRs improved as the filter lengths increased from 100 to 400. However, using filter lengths longer than 400 did not further improve the SIRs. By using GADF derived in Section 4, if D_1 and D_2 delayed the signals by 100 samples, the SIRs in $v_1(t)$ and $v_2(t)$ were 18.71 dB and 20.14 dB with filter lengths equal to 500. The SIRs improved to 19.01 dB and 23.87 dB when the delay became 300 samples and the filter lengths became 700. When the known H_{22} and H_{11} were used in place of D_1 and D_2, the SIRs reached 24.55 dB and 19.88 dB.

6. CONCLUSION

In this paper, several significant improvements are made to the ADF algorithm in terms of reducing computational complexity, improving convergence rate, and handling noncausal IIR filters. A simplified ADF is first proposed to reduce the computational complexity of ADF from $O(N_a^2 + N_b^2)$ to $O(N_a + N_b)$ without significant sacrifice in performance. A transform-domain ADF is next proposed to improve the estimation convergence rate while keeping the computational complexity low. Finally, a generalized ADF is proposed to improve the system's ability in handling the noncausal problem often encountered in co-channel speech separation. The experimental results verified the three methods and showed significant improvement in terms of SIRs in the speech signals after processing. The separation performance in terms of speech recognition accuracy will be evaluated in the future work.

Table 1. The signal-to-interference ratio after signal separation using SADF and TDADF with different filter lengths

	SADF	TDADF
N_a, N_b	SIR in v_1 / v_2	SIR in v_1 / v_2
100	14.61 / 14.31 dB	13.75 / 14.13 dB
200	15.83 / 16.40 dB	15.92 / 15.91 dB
300	17.18 / 17.60 dB	16.89 / 17.30 dB
400	18.38 / 18.91 dB	18.02 / 18.11 dB
500	18.41 / 18.49 dB	18.21 / 18.55 dB
600	18.64 / 18.77 dB	17.97 / 18.42 dB
800	18.01 / 18.37 dB	18.02 / 18.07 dB

ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation under Grant No. IRI-95-02074, and by a grant from the Whitaker Foundation. The measurement of room-acoustics provided by Dr. Sig Soli of House Ear Institute, Los Angeles, CA is also acknowledged.

REFERENCES

[1] R. Cole et al, "The Challenge of Spoken Language Systems: Research Directions for the Nineties," *IEEE Trans. on Speech and Audio Processing*, Vol. 3, pp. 1-21, Jan. 1995.

[2] K. Yen and Y. Zhao, "Robust Automatic Speech Recognition Using a Multi-Channel Signal Separation Front-End," *Proc. ICSLP*, Vol. 3, pp. 1337-1340, Oct. 1996.

[3] K. Yen and Y. Zhao, "Co-Channel Speech Separation for Robust Automatic Speech Recognition: Stability and Efficiency," *Proc. ICASSP*, Vol. 2, pp. 859-862, Apr. 1997.

[4] E. Weinstein, M. Feder, and A. V. Oppenheim, "Multi-Channel Signal Separation by Decorrelation," *IEEE Trans. on Speech and Audio Processing*, Vol. 1, No. 4, pp. 405-413, Oct. 1993.

[5] S. Van Gerven and D. Van Compernolle, "Signal Separation by Symmetric Adaptive Decorrelation: Stability, Convergence, and Uniqueness," *IEEE Trans. on Signal Processing*, Vol. 43, No. 7, pp. 1602-1612, Jul. 1995.

[6] S. S. Narayan, A. M. Peterson, and M. J. Narashima, "Transform Domain LMS Algorithm," *IEEE Trans. on Acoust. Speech Signal Processing*, Vol. ASSP-31, pp. 609-615, 1983.

[7] S. Haykin, *Adaptive Filter Theory*, 3rd Edition, Prentice-Hall, 1996.

[8] W. K. Jenkins, A. W. Hull, J. C. Strait, B. A. Schnaufer, and X. Li, *Advanced Concepts in Adaptive Signal Processing*, Kluwer Academic Publishers, 1996.