ADAPTIVE DECORRELATION FILTERING FOR SEPARATION OF CO-CHANNEL SPEECH SIGNALS FROM M > 2 SOURCES

Kuan-Chieh Yen[1] *Yunxin Zhao*[2]

[1,2]Beckman Institute and Dept. of ECE, University of Illinois, Urbana, IL 61801, USA
[2]Department of CECS, University of Missouri, Columbia, MO 65211, USA
[1]*yen@ifp.uiuc.edu* [2]*zhao@cecs.missouri.edu*

ABSTRACT

The ADF algorithm for separating two signal sources by Weinstein, Feder, and Oppenheim is generalized for separation of co-channel speech signals from more than two sources. The system configuration, its accompanied ADF algorithm, and the choice of adaptation gain are derived. The applicability and limitation of the derived algorithm are also discussed. Experiments were conducted for separation of three speech sources with the acoustic paths measured from an office environment, and the algorithm was shown to improve the average target-to-interference ratio for the three sources by approximately 15 dB.

1. INTRODUCTION

The problem of separating co-channel speech from their convolutive mixtures has gained increasing attention recently. A number of co-channel speech separation algorithms utilizing multi-microphone acquisition and second-order statistics have been proposed in the literature [1][2]. These algorithms do not guarantee unique solutions [1][3][4] in general. However, because they are simpler than the algorithms based on higher-order statistics, and the empirical estimates of second-order statistics are usually more reliable than their higher-order counterparts, they are favorable for certain applications.

In our previous work [5][6], we showed that the adaptive decorrelation filtering (ADF) algorithm by Weinstein, Feder, and Oppenheim [1] is effective in separating two speech signals from their convolutive mixtures. We also developed a number of techniques to improve the efficiency and stability of ADF. Although it was stated in [1] that the ADF algorithm can be extended to separate co-channel speech signals from more than two sources, the generalization is not straightforward and has not been evaluated experimentally. In the current work, we generalize the previous two-source ADF algorithm for separation of more than two speech sources. We start on the three-source case as the first step of the generalization, where details of three-source co-channel speech separation, including a derivation of system configuration and its ADF algorithm, are provided, and limitation of the generalized ADF is also discussed in order to provide insights for its applications. The three-source separation algorithm is then further generalized for cases involving $M > 3$ speech sources.

This paper is organized into six sections. The mathematical model of the three-source co-channel speech environment and the objective of co-channel speech separation are defined in Section 2. In Section 3, the configuration of the separation system and its ADF algorithm are derived; the applicability and limitation of the derived algorithm are discussed. In Section 4, the three-source algorithm is extended into the general case of $M > 3$ sources. Experimental results are presented in Section 5 and a conclusion is made in Section 6.

2. FUNDAMENTALS

2.1. The Three-Source Co-Channel Model

Denoting the source signal from talker j by $x_j(t)$, $j = 1, 2, 3$, and the signals acquired at microphone i by $y_i(t)$, $i = 1, 2, 3$, a three-source co-channel speech environment can be defined, in the frequency domain, as

$$\underline{Y}(f) = \mathbf{H}(f)\underline{X}(f) \tag{1}$$

where $\underline{Y} = [Y_1 \quad Y_2 \quad Y_3]^T$, $\underline{X} = [X_1 \quad X_2 \quad X_3]^T$, with ^{T} denoting vector transpose, and $\mathbf{H}$ is a 3-by-3 matrix with the (i, j)-th entry as H_{ij}. Each transfer function H_{ij} models the acoustic path between the talker j and the microphone i. It is assumed that these acoustic paths are unknown and may be time-varying. A block diagram of this model is shown in Fig. 1.

Assuming that the target source of microphone i is talker i, each acquired signal can then be decomposed into two additive components as

$$y_i(t) = H_{ii}\{x_i(t)\} + \sum_{j=1, j\neq i}^{3} H_{ij}\{x_j(t)\} \tag{2}$$

where the first term is referred to as the target signal component and the second term is referred to as the interfering component. It is reasonable to assume that the source speech signals are zero-mean and independent to each other.

2.2. Objective of Co-Channel Speech Separation

The objective of co-channel speech separation is to eliminate the interfering component in each acquired signal and hence separate the signals from different sources from their convolutive mixtures. Denoting the output signals of the separation system by $v_i(t)$, $i = 1, 2, 3$, the frequency-domain input-output relation of a separation system is

$$\underline{V}(f) = \mathbf{F}(f)\underline{Y}(f) \tag{3}$$

where $\underline{V} = [V_1 \quad V_2 \quad V_3]^T$, and $\mathbf{F}$ is a 3-by-3 matrix representing the separation system. If $\mathbf{FH}$ is a diagonal matrix, from Eqs. (1) and (3), $v_i(t)$ will contain nothing more than $x_i(t)$, albeit linearly distorted. Therefore, signal separation is achieved. This separation criterion is used in designing the separation system $\mathbf{F}$ in the next section.

3. CO-CHANNEL SPEECH SEPARATION FOR M = 3 SOURCES

In this section, the configuration of the separation system is first determined according to the separation criterion discussed in the previous section. Then the ADF algorithm corresponding to this configuration is derived. The applicability and limitation of the derived algorithm are discussed at the end of the section.

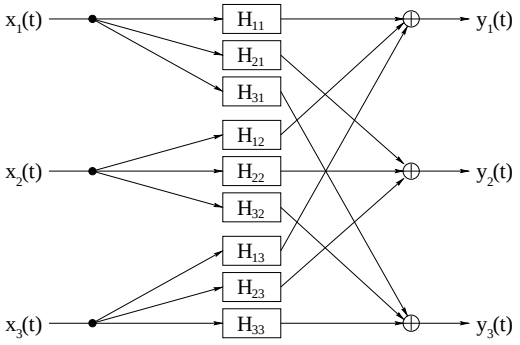

Figure 1. Block diagram of the three-source co-channel speech model

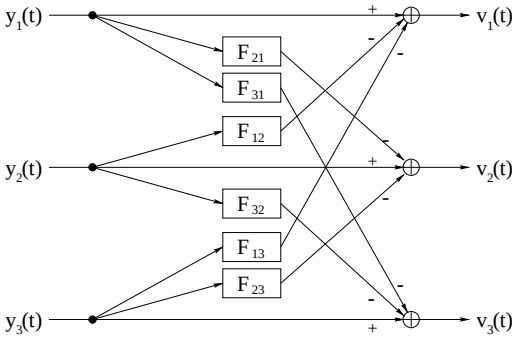

Figure 2. Block diagram of the three-source co-channel speech separation system

3.1. A Modified Three-Source Co-Channel Model

Because the number of acquired signals is equal to the number of source signals and the distributions of the source signals are unknown, it is impossible to identify the absolute acoustic paths (e.g. H_{11} and H_{21}) from the acquired signals. Instead, only the relative acoustic paths (e.g. H_{21}/H_{11}) are identifiable. Taking into consideration of this limitation, Eq. (1) can be modified as

$$\underline{Y}(f) \quad = \quad \tilde{\mathbf{H}}(f) \quad \underline{\tilde{X}}(f) \tag{4}$$

where $\underline{\tilde{X}} = \begin{bmatrix} \tilde{X}_1 & \tilde{X}_2 & \tilde{X}_3 \end{bmatrix}^T$ with $\tilde{X}_j = H_{jj}X_j$, and $\tilde{\mathbf{H}}$ is a 3-by-3 matrix with the (i,j)-th entry as $\tilde{H}_{ij} = H_{ij}/H_{jj}$. It is straightforward to show that $\mathbf{FH}$ is diagonal if and only if $\mathbf{F}\tilde{\mathbf{H}}$ is diagonal.

3.2. System Configuration

From the above discussion, if the relative acoustic paths $\tilde{H}_{ij}$'s can be identified, an intuitive choice for $\mathbf{F}$ is $\tilde{\mathbf{H}}^{-1}$, which yields $v_i(t) = \tilde{x}_i(t)$. However, the quadratic terms and $det\tilde{\mathbf{H}}$ involved in calculating $\tilde{\mathbf{H}}^{-1}$ from $\tilde{H}_{ij}$'s make it difficult to implement such a system. In addition, it does not provide a constructive method for estimating the relative acoustic paths.

Alternatively, $\mathbf{F}$ can be chosen as

$$\mathbf{F}(f) = \begin{bmatrix} 1 & -F_{12}(f) & -F_{13}(f) \\ -F_{21}(f) & 1 & -F_{23}(f) \\ -F_{31}(f) & -F_{32}(f) & 1 \end{bmatrix} \tag{5}$$

with $F_{ij} = \left\{1 - \tilde{H}_{jk}\tilde{H}_{kj}\right\}^{-1}\left\{\tilde{H}_{ij} - \tilde{H}_{ik}\tilde{H}_{kj}\right\}$, $k \in \{1,2,3\}$ and $k \neq i,j$. Instead of estimating the relative acoustic paths, F_{ij}'s are estimated and used for separating the signals. This gives a simpler system configuration. The block diagram of the separation system is given in Fig. 2.

3.3. Algorithm Derivation

As in [1], decorrelation between $v_i(t)$'s is used as the criterion in estimating the filters F_{ij}'s, i.e., $E\left\{V_i(f)V_j^*(f)\right\} = 0$, $i \neq j$, where * denotes the complex conjugate. By combining Eqs. (3) and (5), $E\{V_1(f)V_2^*(f)\} = 0$ can be expanded by substituting V_1 with $Y_1 - F_{12}Y_2 - F_{13}Y_3$, to become $E\{Y_1V_2^*\} = F_{12}E\{Y_2V_2^*\} + F_{13}E\{Y_3V_2^*\}$. Its time-domain equivalent can be written as

$$r_{y_1v_2}(\tau) = f_{12}(\tau) \otimes r_{y_2v_2}(\tau) + f_{13}(\tau) \otimes r_{y_3v_2}(\tau) \tag{6}$$

where $r_{y_iv_j}(\tau) = E\{y_i(t)v_j(t-\tau)\}$ is the cross-correlation between $y_i(t)$ and $v_j(t)$, $f_{ij}(t)$ is the impulse response of filter F_{ij}, and $\otimes$ denotes convolution.

If the filters F_{ij}'s are chosen to be N-tap FIR filters, the following vectors can be defined accordingly:

$$\underline{f}_{ij} = [f_{ij}(0) \quad \cdots \quad f_{ij}(N-1)]^T \tag{7}$$

$$\underline{y}_i(t) = [y_i(t) \quad \cdots \quad y_i(t-N+1)]^T \tag{8}$$

$$\underline{v}_i(t) = [v_i(t) \quad \cdots \quad v_i(t-N+1)]^T \tag{9}$$

and Eq. (6) can be converted into its vector form as

$$E\{\underline{v}_2(t)y_1(t)\} = E\left\{\underline{v}_2(t)\underline{y}_2^T(t)\right\}\underline{f}_{12} + E\left\{\underline{v}_2(t)\underline{y}_3^T(t)\right\}\underline{f}_{13} \tag{10}$$

By manipulating $E\{V_1(f)V_3^*(f)\} = 0$ in the same way, another linear equation of $\left\{\underline{f}_{12}, \underline{f}_{13}\right\}$ can be formulated as

$$E\{\underline{v}_3(t)y_1(t)\} = E\left\{\underline{v}_3(t)\underline{y}_2^T(t)\right\}\underline{f}_{12} + E\left\{\underline{v}_3(t)\underline{y}_3^T(t)\right\}\underline{f}_{13} \tag{11}$$

Following the same procedure, similar linear equation pairs as Eqs. (10) and (11) can also be formulated for $\left\{\underline{f}_{23}, \underline{f}_{21}\right\}$ and $\left\{\underline{f}_{31}, \underline{f}_{32}\right\}$, respectively. Putting together the six linear equations leads to the equation

$$\mathbf{R}_{vy_c}\underline{f}_c = \underline{r}_{vy_c} \tag{12}$$

where

$$\mathbf{R}_{vy_c} = diag\{\mathbf{R}_{vy_c,23}, \mathbf{R}_{vy_c,31}, \mathbf{R}_{vy_c,12}\} \tag{13}$$

$$\underline{f}_c = \begin{bmatrix} \underline{f}_{12}^T & \underline{f}_{13}^T & \underline{f}_{23}^T & \underline{f}_{21}^T & \underline{f}_{31}^T & \underline{f}_{32}^T \end{bmatrix}^T \tag{14}$$

$$\underline{r}_{vy_c} = E\left\{\underline{c}_{vy_c}(t)\right\} \tag{15}$$

with

$$\mathbf{R}_{vy_c,ij} = E\left\{ \begin{bmatrix} \underline{v}_i(t) \\ \underline{v}_j(t) \end{bmatrix} \begin{bmatrix} \underline{y}_i^T(t) & \underline{y}_j^T(t) \end{bmatrix} \right\} \tag{16}$$

$$\underline{c}_{vy_c}(t) = \begin{bmatrix} \underline{v}_2(t)y_1(t) \\ \underline{v}_3(t)y_1(t) \\ \underline{v}_3(t)y_2(t) \\ \underline{v}_1(t)y_2(t) \\ \underline{v}_1(t)y_3(t) \\ \underline{v}_2(t)y_3(t) \end{bmatrix} \tag{17}$$

If all the real parts of eigenvalues of $\mathbf{R}_{vy_c}$ maintain positive when $\underline{f}_c$ is varied during adaptive estimation, the following adaptation equation based on the stochastic approximation method by Robbins and Monro [7] can be applied:

$$\underline{f}_c^{(t+1)} = \underline{f}_c^{(t)} + \mu(t)\underline{c}_{vy_c}(t) \tag{18}$$

In computing Eq. (18), the output equations for $v_i(t)$'s are defined as

$$v_i(t) = y_i(t) - \sum_{j=1;j\neq i}^{3} \underline{y}_j^T(t)\underline{f}_{ij}^{(t)} \tag{19}$$

Based on Eqs. (14) and (17), Eq. (18) can be split into six adaptation equations for $\underline{f}_{ij}$'s, i.e.,

$$\underline{f}_{ij}^{(t+1)} = \underline{f}_{ij}^{(t)} + \mu(t)\underline{v}_j(t)v_i(t) \tag{20}$$

As in the two-source separation case, the adaptation gain $\mu(t)$ controls the convergence of $\underline{f}_{ij}$'s [5]. Following a similar analysis as in [5], the adaptation gain can be chosen as

$$\mu(t) = 2\gamma \left\{ (M-1)N \sum_{i=1}^{M} \hat{\sigma}_{y_i}^2(t) \right\}^{-1} \tag{21}$$

where $\hat{\sigma}_{y_i}^2(t)$ is the current estimate of the variance of y_i using its L ($L \gg N$) most recent samples, and γ is a constant satisfying $0 < \gamma < 1$ and it can be chosen according to the time-varying nature of the acoustic environment. To allow margins for errors in the estimation of the variances, it was determined through experiments that $\gamma = 0.01$ to be a favorable choice.

Eqs. (19), (20), and (21) form the ADF algorithm for three-source co-channel speech separation.

3.4. Applicability and Limitation

As mentioned above, in order for Eq. (18) of stochastic approximation to lead to converged estimation of $\underline{f}_c$, the eigenvalues of $\mathbf{R}_{vy_c}$ need to have positive real parts for $\underline{f}_c$ within the region of operation. If the adaptation starts with $\underline{f}_{ij}^{(0)} = \underline{0}$, $\mathbf{R}_{vy_c}$ will be positive-definite at $t = 0$ since $v_i(t) = y_i(t)$. As $\underline{f}_c$ converges to its ideal solution, it can be shown that the products of the relative acoustic paths between each pair of signal sources, i.e., $\tilde{H}_{ij}\tilde{H}_{ji}$, $i \neq j$, play dominating roles in determining the locations of the eigenvalues. The degree of cross-source interference between sources i and j at frequency f can be quantified as the cross-interference level (CIL)

$$CIL_{ij}(f) = \left| \tilde{H}_{ij}(f)\tilde{H}_{ji}(f) \right| \tag{22}$$

It can be stated that if $CIL_{ij}(f) \ll 1$ for all $i \neq j$ and f, all the eigenvalues will lie in the right-hand side of the complex plane. In practice, this condition is satisfied if each microphone is placed relatively closer to its target source than to the interfering sources. Details of these analysis will be addressed in a future publication.

Furthermore, from Eq. (20), the adjustments for $f_{ij}(0)$ and $f_{ji}(0)$ are made by the same term $\mu(t)v_i(t)v_j(t)$, and hence $f_{ij}^{(t)}(0) - f_{ij}^{(0)}(0)$ is always equal to $f_{ji}^{(t)}(0) - f_{ji}^{(0)}(0)$. Therefore, if $f_{ij,ideal}(0) - f_{ij}^{(0)}(0) \neq f_{ji,ideal}(0) - f_{ji}^{(0)}(0)$, $f_{ij}(0)$ and $f_{ji}(0)$ will never reach their ideal values at the same time. This limitation will have significant impact when any of the $f_{ij}(0)$'s is one of the significant weights in $\underline{f}_{ij}$. However, this seldom happens if each talker is closer to its targeting microphone than to the other microphones.

4. THE ADF ALGORITHM FOR $M > 3$ SOURCES

The three-source separation algorithm derived in the previous section can be further generalized to the cases involving $M > 3$ speech sources. M microphones are used to acquire the mixed signals, $y_j(t)$, $j = 1, 2, \cdots, M$. The output equations for the separated signals $v_i(t)$, $i = 1, 2, \cdots, M$ can be obtained by replacing 3 with M in Eq. (19). The $M(M-1)$ required filters, F_{ij}'s, can be estimated by Eq. (20), with the adaptation gain $\mu(t)$ determined by Eq. (21).

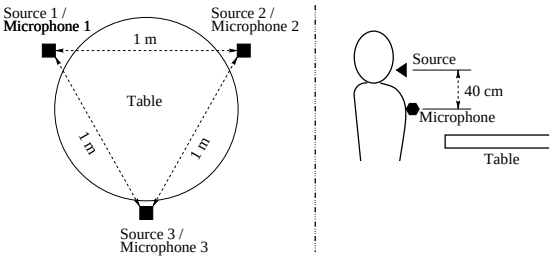

Figure 3. The talker-microphone configuration used in measuring the acoustic paths

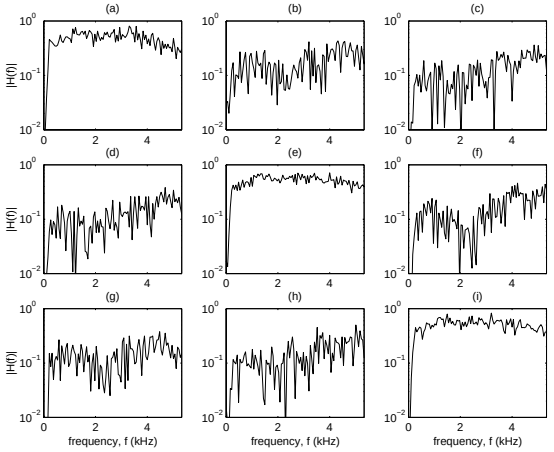

Figure 4. The frequency responses of the measured filters: (a) $\left|\bar{H}_{11}\right|$, (b) $\left|\bar{H}_{12}\right|$, (c) $\left|\bar{H}_{13}\right|$, (d) $\left|\bar{H}_{21}\right|$, (e) $\left|\bar{H}_{22}\right|$, (f) $\left|\bar{H}_{23}\right|$, (g) $\left|\bar{H}_{31}\right|$, (h) $\left|\bar{H}_{32}\right|$, and (i) $\left|\bar{H}_{33}\right|$

5. EXPERIMENTS

In this section, the experimental conditions of the source signals and the acoustic environment are first described. Comparisons are made on the source separation performance under various cases of source energy levels (SELs) and CILs.

5.1. Source Signals

A set of speech signals were chosen from the TIMIT database and were down-sampled from 16 kHz to 10.67 kHz to become the source signals $x_j(t)$'s in Eq. (1).

5.2. Acoustic Environment

The acoustic paths from each talker to each microphone were measured in an office according to the configuration shown in Fig. 3. As shown by the top-down view of Fig. 3, the "talkers" were spaced evenly around a round table with the distance between each pair of adjacent "talkers" at about 1 m. The microphones were installed about 40 cm below their respective targeted sources, as shown by the illustration on the right side of Fig. 3. The measured filter that models the acoustic path from the "talker" j to the microphone i is referred to as $\bar{H}_{ij}$, for all (i, j). For each filter, the first 200 samples of the impulse response were used, which covered a time span of 18.75 msec at the sampling rate of 10.67 kHz. The frequency responses of $\bar{H}_{ij}$'s ($\left|\bar{H}_{ij}(f)\right|$) are given in Fig. 4.

5.3. System Performance under Various SELs

In this experiment, the source signals were first scaled to generate various SELs, and they were then convolved by the measured filters and mixed according to Eq. (1), with

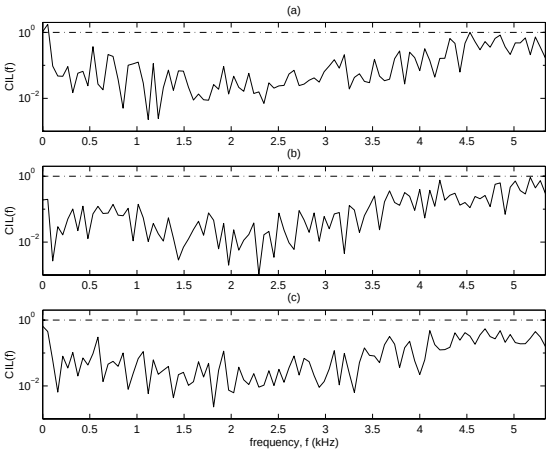

Figure 5. The cross-interference levels of the measured filters: (a) CIL_{12}, (b) CIL_{23}, and (c) CIL_{31}

Table 1. The TIRs (in dB) before and after processing under various SELs

Source	1	2	3	Sum
All three sources having the same energy				
BeforeProcessing	7.73	8.43	7.97	24.12
After processing	22.28	22.39	22.76	67.43
Improvement	14.56	13.96	14.79	43.31
Source 2 being 20 dB weaker than the others				
BeforeProcessing	13.33	-11.57	10.07	11.83
AfterProcessing	23.90	7.89	25.82	57.61
Improvement	10.57	19.46	15.75	45.78
Source 3 being 20 dB stronger than the others				
BeforeProcessing	-6.67	-9.98	27.97	11.32
AfterProcessing	14.83	15.50	26.16	56.49
Improvement	21.50	25.48	-1.81	45.17

$H_{ij}(f) = \bar{H}_{ij}(f)$. The filter length ($N$) was set to 400. The target-to-interfering ratios (TIRs) before and after processing are summarized in Table 1. It can be observed that the system performance was fairly consistent under a wide range of SELs. For all cases, the sums of the TIR improvements on the three mixed signals were between 40 to 50 dB, where greater improvements were obtained on the mixed signals that had lower TIRs. When the filter coefficients were initialized as 0's, it took several minutes of adaptation to reach the performance of Table 1. However, 60 to 80 % of the final TIR improvements were reached in a few seconds. The CILs for the described acoustic environment are plotted in Fig. 5, where the CILs were seen to be close to 1 only at a few very low or high frequency bands. Since the energy of speech signals was relatively low in these frequency bands, the impact of these high CIL values on separation of speech sources was insignificant.

5.4. System Performance under Various CILs

In this experiment, the impulse responses of the cross-channel acoustic paths ($\bar{H}_{ij}$, $i \neq j$) were multiplied by a constant K, while the impulse responses of the intra-channel acoustic paths ($\bar{H}_{ii}$'s) remained unchanged. This resulted in a magnification of CILs by K^2. The source signals were of the same energy level. The TIRs before and after processing are summarized in Table 2. As K increased, the sums of the improved TIRs decreased, indicating that when the CILs become too large, the ADF algorithm would fail.

Table 2. The TIRs (in dB) before and after processing under various CILs

Source	1	2	3	Sum
$K = 1$: The first case in Table 1.				
$K = 2$:				
BeforeProcessing	1.70	2.40	1.95	6.06
AfterProcessing	14.65	13.27	13.59	41.50
Improvement	12.94	10.86	11.64	35.44
$K = 4$:				
BeforeProcessing	-4.32	-3.62	-4.07	-12.00
AfterProcessing	2.61	1.01	2.28	5.90
Improvement	6.93	4.62	6.35	17.90

6. CONCLUSION

In this paper, the ADF algorithm for two-source speech separation by Weinstein, Feder, and Oppenheim is generalized for $M > 2$ speech source separation. A method for determining the adaptation gain is proposed to better balance between system stability and efficiency. The applicability and limitation of the proposed algorithm is discussed. The experimental results show that the algorithm can effectively improve the TIRs of the co-channel speech signals, provided that the talkers are not too closely spaced compared to the distances between each microphone and its target talker. The evaluation and improvement of the proposed technique under additional background noises will follow in a future work.

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 acknowledged.

REFERENCES

[1] 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.

[2] 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.

[3] S. Shamsunder and G. B. Giannakis, " Multichannel Blind Signal Separation and Reconstruction," *IEEE Trans. on Speech and Audio Processing*, Vol. 5, pp. 515-528, Nov. 1997.

[4] D. Yellin and E. Weinstein, "Multichannel Signal Separation: Methods and Analysis," *IEEE Trans. on Signal Processing*, Vol. 44, pp. 106-118, 1996.

[5] 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.

[6] K. Yen and Y. Zhao, "Improvements on Co-Channel Speech Separation Using ADF: Low Complexity, Fast Convergence, and Generalization," *Proc. ICASSP*, Vol. 2, pp. 1025-1028, May 1998.

[7] H. Robbins and S. Monro, "A Stochastic Approximation Method," *Annals of Mathematical Statistics*, Vol, 22, pp. 400-407, 1951.