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

Kuan-Chieh Yen*  
Yunxin Zhao†

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 sepa-
ration 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 sepa-
ration of three speech sources with the acoustic paths mea-
sured 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 re-
cently. A number of co-channel speech separation algo-
rithms utilizing multi-microphone acquisition and second-
order statistics have been proposed in the literature [1][2].
These algorithms do not guarantee unique solutions [3][4] in
general. However, because they are simpler than the algo-
rithms based on higher-order statistics, and the empirical
descriptions of second-order statistics are usually more reliable
than their higher-order counterparts, they are favorable for
certain applications.

In our previous work [3][6], we showed that the adap-
tive 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 generaliza-
tion is not straightforward and has not been evaluated ex-
perimentally. In the current work, we generalize the previ-
ous 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 the derivation of sys-
tem 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 mathe-
matical model of the three-source co-channel speech envi-
enronment 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_i(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 de-

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

where the first term is referred to as the target signal com-
ponent and the second term is referred to as the interfer-
ing 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 elimi-
nate 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 \( \hat{v}_i(t) \), \( i = 1, 2, 3 \), the frequency-domain
input-output relation of a separation system is

\[ \hat{V}(f) = \mathbf{F}(f) \hat{Y}(f) \]

where \( \hat{V} = [V_1 \ V_2 \ V_3]^T \), and \( \mathbf{F} \) is a 3-by-3 matrix rep-
representing the separation system. If \( \mathbf{F} \) 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 dis-
cussed in the previous section. Then the ADF algorithm
operating corresponding to this configuration is derived. The applica-
bility and limitation of the derived algorithm are discussed at
the end of the section.
diagram of the separation system is given in Fig. 1, which yields $H$ and $\frac{1}{3}$-by-$\frac{1}{3}$ matrix with the $ij$-th entry in $F$. From the above discussion, if the relative acoustic paths $H_{ij}$ are straightforward to show that $F'H$ is diagonal if and only if $F'H$ is diagonal.

### 3.2. System Configuration

From the above discussion, if the relative acoustic paths $H_{ij}$'s can be identified, an intuitive choice for $F$ is $H^T$, which yields $v_i(t) = x_i(t)$. However, the quadratic terms and $detH$ involved in calculating $H^{-1}$ from $H_{ij}$'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, $F$ can be chosen as

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

(5)

with $F_{ij} = \{1 - \bar{H}_{ji}\bar{H}_{ki}\}^{-1}\{\bar{H}_{ji} - \bar{H}_{ik}\bar{H}_{kj}\}$, $k \in \{1, 2, 3\}$ 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), $E\left\{V_i(f)V_j^*(f)\right\} = 0$ can be expanded by substituting $V_i$ with $Y_i - F_{ij}Y_j = F_{ij}V_j$, to become $E\left\{V_iV_j^*\right\} = F_{ij}E\left\{V_jV_j^*\right\} + F_{ij}E\left\{V_iV_j^*\right\}$. Its time-domain equivalent can be written as

$$r_{y_iy_j}(\tau) = f_{ij}(\tau) \otimes r_{y_ry_j}(\tau) + f_{i3}(\tau) \otimes r_{y_ry_3}(\tau)$$

(6)

where $r_{y_ry_j}(\tau) = E\left\{y_i(t)v_j(t + \tau)\right\}$ is the cross-correlation between $y_i(t)$ and $v_j(t)$, $f_{ij}(\tau)$ 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:

$$\begin{align*}
\underline{f}_{ij} &= [f_{ij}(0) \cdots f_{ij}(N-1)]^T \\
\underline{y}_i(t) &= [y_i(t) \cdots y_i(t-N+1)]^T \\
\underline{v}_i(t) &= [v_i(t) \cdots v_i(t-N+1)]^T
\end{align*}$$

(7)-(9)

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

$$E\left\{\underline{w}_i(t)y_i(t)\right\} = E\left\{\underline{w}_i(t)\underline{y}_i(t)^T\right\}\underline{f}_{ij} + E\left\{\underline{w}_i(t)\underline{v}_i(t)^T\right\}\underline{f}_{i3}$$

(10)

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

$$E\left\{\underline{w}_i(t)y_i(t)\right\} = E\left\{\underline{w}_i(t)\underline{y}_i(t)^T\right\}\underline{f}_{ij} + E\left\{\underline{w}_i(t)\underline{v}_i(t)^T\right\}\underline{f}_{i3}$$

(11)

Following the same procedure, similar linear equation pairs as Eqs. (10) and (11) can also be formulated for $\underline{f}_{i3}$ and $\underline{f}_{ij}$, respectively. Putting together the six linear equations leads to the equation

$$R_{y_iy_j}\underline{f}_{ij} = \underline{w}_{y_i}$$

(12)

where

$$R_{y_iy_j} = diag\{R_{y_ry_2}, R_{y_ry_3}, R_{y_ry_{i3}}\}$$

(13)

$$\underline{f}_{ij} = \begin{bmatrix} \underline{f}_{ij}^{T} \\ \underline{f}_{i3}^{T} \end{bmatrix}$$

(14)

$$\underline{w}_{y_i} = E\left\{\underline{w}_{y_i}(t)^T\right\}$$

(15)

with

$$R_{y_iy_j,ij} = E\left\{\begin{bmatrix} \underline{w}_i(t) \\ \underline{w}_j(t) \end{bmatrix} \begin{bmatrix} \underline{y}_i(t)^T \\ \underline{y}_j(t)^T \end{bmatrix}\right\}$$

(16)

$$\underline{w}_{y_i}(t) = \begin{bmatrix} \underline{w}_i(t) \\ \underline{w}_j(t) \end{bmatrix}$$

(17)

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

$$\underline{f}_{ij}^{(t+1)} = \underline{f}_{ij}^{(t)} + \mu(t)\underline{w}_{y_i}(t)$$

(18)

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

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

(19)
Based on Eqs. (14) and (17), Eq. (18) can be split into six adaptation equations for $L_j$, i.e.,

$$f_j^{t+1} = f_j^t + \mu(t) y_j(t) v_i(t)$$

(20)

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

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

(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 $\gamma$ 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 $f_j$, the eigenvalues of $R_{y,y}$ need to have positive real parts for $f_j$ within the region of operation. If the adaptation starts with $f_j(0) = 0$. $R_{y,y}$ will be positive-definite at $t = 0$ since $v_i(t) = y_i(t)$. As $f_j$ 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., $H_{ij}H_{ji}$, $i \neq j$, 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)

$$CIL_{ij} (f) = |H_{ij}(f)H_{ji}(f)|$$

(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 $f_{ij}(0) - f_{ij}(0) = f_{ji}(0) - f_{ji}(0)$ is always equal to $f_{ij}(0) - f_{ij}(0)$. Therefore, if $f_{ij}(0) - f_{ij}(0) \neq f_{ij}(0) - f_{ij}(0)$, $f_{ij}(0)$ 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 $f_j$. However, this seldom happens if eachtalker 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, $y_i(t), j = 1, 2, \ldots, M$. The output equations for the separated signals $v_i(t), i = 1, 2, \ldots, M$ can be obtained by replacing $3$ with $M$ in Eq. (19). The $M(M-1)$ required filters, $F_{ji}$'s, can be estimated by Eq. (20), with the adaptation gain $\mu(t)$ determined by Eq. (21).

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 $x_i(t)$'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 $\hat{H}_{ij}$, 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 $\hat{H}_{ij}$'s ($|H_{ij}(f)|$) 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
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


