# FAST CONVERGENCE SPEECH SOURCE SEPARATION IN REVERBERANT ACOUSTIC ENVIRONMENT

*Yunxin Zhao and Rong Hu*
Department of Computer Science
University of Missouri, Columbia, MO 65211, USA
zhaoy@missouri.edu rhq2c@mizzou.edu

**ABSTRACT**

Three significant enhancements to time-domain adaptive decorrelation filtering (ADF) are proposed for effective separation and recognition of simultaneous speech sources in reverberant room conditions. The methods include whitening filtering on cochannel speech prior to ADF to improve condition of adaptive estimation, a novel block-iterative implementation of ADF to speed up convergence rate, and an integration of multiple ADF outputs through optimal post filtering. Experimental data were generated by convolving TIMIT speech with acoustic path impulse responses measured in real acoustic environment, with a 2m microphone-source distance and an initial target-to-interference ratio of about 0 dB. The proposed methods are shown to have speeded up the convergence rate of ADF to a level feasible for online applications, and they have significantly improved target-to-interference ratio and accuracy of phone recognition.

## 1. INTRODUCTION

Blind source separation of simultaneous speech signals has been an area of active research in recent years. Among many approaches, time-domain adaptive decorrelation filtering (ADF) [1,2,3] and frequency-domain independent component analysis (ICA) [4,5,6] are heavily studied. The performance of speech source separation is known to depend on conditions of room reverberation, locations of sources and microphones, the properties of the sources, etc. *

The time-domain ADF method, when evaluated under favorable conditions, was able to deliver large gains in signal-to-interference ratio [2,3]. However, under unfavorable conditions of room reverberation and large source-microphone distance, ADF is very slow in convergence and delivers limited gain of target-to-interference ratio (TIR). The difficulty can be attributed to the ADF principle, the spectral characteristics of speech, and room reverberation. In ADF, acoustic paths need to be modeled by finite-impulse response (FIR) filters in order to obtain correct solutions. The increased lengths of FIR filters with long acoustic paths make them less distinguishable from IIR filters. Further, in reverberant rooms, reverberation-induced long tails of impulse responses are inherently difficult to identify. In speech, voiced sounds have strong low frequency components, which cause large spreads of eigenvalues in the correlation matrices of source speech as well as speech mixtures, and hence slow down convergence rate of adaptive filtering algorithms. Furthermore, in ADF no distinction is made between perceptually important and irrelevant spectral components of speech, and therefore the separation processing is not optimized for automatic speech recognition systems.

In the current work, three significant enhancements to the time-domain ADF approach are proposed for effective separation and recognition of simultaneous speech sources in reverberant acoustic conditions. First, whitening filtering is performed on cochannel speech prior to ADF to improve its condition of adaptive estimation, where three whitening filtering methods that are motivated by spectral characteristics of speech are investigated. Second, a novel block-iterative implementation of ADF is made to speed up convergence for online speech source separation. Third, an optimal post filtering method is developed to integrate ADF outputs of multiple microphone pairs to further reduce reverberation noise and interference speech. Evaluation experiments were performed on phone recognition of separated speech by using a hidden Markov model (HMM) based speaker-independent phone recognition system, with the source speech materials taken from the TIMIT database. The proposed techniques are shown to have significantly improved system performance in convergence rate, signal to interference ratio, and recognition accuracy.

## 2. OVERVIEW OF ADF

Assume that two microphones are used to acquire convolutive mixtures of two mutually uncorrelated source signals $s_j(t),\ j=1,2$ and produce outputs $y_i(t)$, $i=1,2$. Denote the acoustic transfer function from the source *j* to the microphone *i* by $H_{ij}(z)$. The cochannel environment is then modeled as

$$\begin{bmatrix} Y_1(z) \\ Y_2(z) \end{bmatrix} = \begin{bmatrix} 1 & F_{12}(z) \\ F_{21}(z) & 1 \end{bmatrix} \begin{bmatrix} H_{11}S_1(z) \\ H_{22}S_2(z) \end{bmatrix} \quad (1)$$

with $F_{ij}(z) = H_{ij}(z)/H_{jj}(z)$. Define $\underline{f}_{ij}^{(t)}$ to be length-*N* FIR filters of $F_{ij}(z)$ at time *t*. The output signals of the ADF system are then generated as

$$\begin{aligned} v_1(t) &= y_1(t) - \underline{y}_2^T(t)\underline{f}_{12}^{(t)} \\ v_2(t) &= y_2(t) - \underline{y}_1^T(t)\underline{f}_{21}^{(t)} \end{aligned} \quad (2)$$

where $\underline{y}_j(t) = \begin{bmatrix} y_j(t) & y_j(t-1) & \cdots & y_j(t-N+1) \end{bmatrix}^T$. By taking decorrelation of system outputs as the separation criterion, the filters can be adaptively estimated as

* This work is supported in part by NSF under the grant NSF EIA 9911095

$$\underline{f}_{12}^{(t+1)} = \underline{f}_{12}^{(t)} + \mu(t)\underline{v}_2(t)v_1(t)$$
$$\underline{f}_{21}^{(t+1)} = \underline{f}_{21}^{(t)} + \mu(t)\underline{v}_1(t)v_2(t) \quad (3)$$

To ensure system stability, the adaptation gain is determined in [2] as

$$\mu(t) = \frac{2\gamma}{N(\sigma_{y_1}^2(t) + \sigma_{y_2}^2(t))} \quad (4)$$

where $0 < \gamma < 1$ , and $\sigma_{y_1}^2(t)$ and $\sigma_{y_2}^2(t)$ are short-time energy estimates of inputs. The outputs $v_i(t),\ \ i = 1,2$ converge to linearly transformed source signals $s_i(t),\ \ i = 1,2$ [1,2,3]. A block diagram of the source mixing and separation systems is shown in Fig. 1.

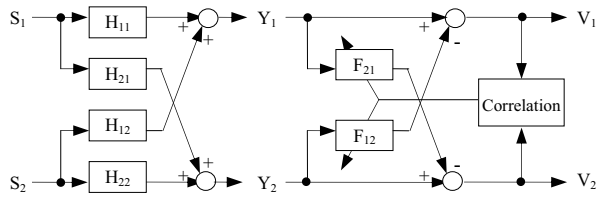


*Figure 1 Block diagram of source mixing and separation systems*

## 3. THE EHANCEMENT METHODS

### 3.1. Whitening Filtering

The proposed whitening filtering operations include preemphasis, prewhitening, and inverse filtering based on long-term LPC analysis. Preemphasis is a first-order high-pass filter in the form $P(z) = 1 - \mu z^{-1}$, with $\mu \approx 1$, and it is commonly used in linear predictive coding of speech to compensate for the 6-dB per octave spectral tilt in voiced speech [7]. In prewhitening, long-term power spectral density of speech is measured and its inverse filter is designed to "whiten" speech spectral distribution. In the current work, an inverse filter is designed by an FIR filter of order 5 based on the long-term speech power spectrum detailed in [8]. In long-term LPC analysis, short-time autocorrelation coefficients are averaged from offline training speech data and a P[th] order LPC analysis is carried out. The inverse filter of LPC is then used as the whitening filter.

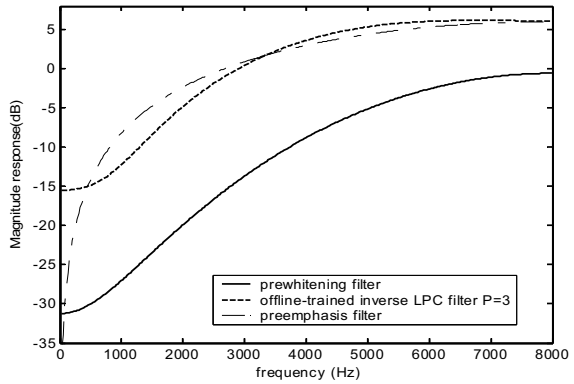


*Figure 2 Frequency responses of whitening filters*

In Fig. 2, frequency responses are shown for the whitening filters of preemphasis ( $\mu = 1$ ), prewhitening, and a 3[rd] order LPC inverse filter. The three filters have similar 6 dB per octave high-pass characteristics in the range of 1 KHz to 5 KHz, while the degrees of low-frequency attenuation are different.

### 3.2. Block-Iterative ADF

The adaptive filter estimation and source separation as defined in Eqs. (3) and (2) for ADF are performed sample by sample. In the block iterative implementation, input speech data are divided into size-$B$ blocks, i.e., $Y_{i,n} = [y_{i,nB+t}, t = 0,1,\cdots,B-1],\ \ i = 1,2$ and $n = 0,1,\cdots$ . Within each block $n$, adaptive estimation and separation are iteratively performed by Eqs. (2) and (3), where the filter estimates obtained at the end of the data block in the $r^{\text{th}}$ iteration, $\underline{f}_{12}^{(B-1)}(n,r)$ and $\underline{f}_{21}^{(B-1)}(n,r)$, are used as the initial filter estimates at the $r+1^{\text{th}}$ iteration, $\underline{f}_{12}^{(0)}(n,r+1)$ and $\underline{f}_{21}^{(0)}(n,r+1)$. The relative change of filter estimates between two successive iterations is computed as

$$C_{n,r+1} = \frac{1}{2}\left( \frac{\left\| \underline{f}_{12}^{(B-1)}(n,r+1) - \underline{f}_{12}^{(B-1)}(n,r) \right\|}{\left\| \underline{f}_{12}^{(B-1)}(n,r+1) \right\|} + \frac{\left\| \underline{f}_{21}^{(B-1)}(n,r+1) - \underline{f}_{21}^{(B-1)}(n,r) \right\|}{\left\| \underline{f}_{21}^{(B-1)}(n,r+1) \right\|} \right) \quad (5)$$

If $C_{n,r+1} < \varepsilon$ , then the estimation terminates for the block and ADF is moved on to next block $n$+1. The initial filter estimates for the block $n$+1 is set as the final filter estimates of the block $n$.

The block length $B$ and the termination threshold $\varepsilon$ are important implementation parameters. A block should be sufficiently long such that different phonetic sounds are included within each block, since second-order statistic methods of blind source separation such like ADF do not guarantee correct solutions for stationary signal sources [4]. Obviously, block iterative estimation also induces a buffering delay that is determined by the block length. Therefore, the choice of $B$ is a tradeoff between accuracy and delay. In the extreme case, block length is set as data sequence length, and batch iterative estimation is resulted. The threshold $\varepsilon$ is a tradeoff between convergence rate and computation load. A small $\varepsilon$ calls for more iterations that speeds up convergence but incurs more computation, and a large $\varepsilon$ calls for fewer iterations and hence slower convergence and less computation. It was found experimentally that running too many iterations within a block could cause instability. In this work, the threshold $\varepsilon$ was chosen as 0.0005, and the iteration number $r$ was also hard limited to be between 3 and 8.

### 3.3. Post Filtering

The proposed optimal post filtering is based on minimum mean squared error estimation [9,10]. Assume in a $K$ microphone-pair system $K$ pairs of ADF outputs are available, i.e., $\{v_1^{(k)}(t), v_2^{(k)}(t)\}$, $k = 1,2,\cdots,K$ . Further assume that source 1 is the target and $v_1^{(1)}(t)$ is the reference. Then $v_1^{(k)}(t)$ 's are filtered to match $v_1^{(1)}(t)$ by $\overline{V}_1^{(k)}(f) = H_1^{(k)}(f)V_1^{(k)}(f)$ , with $H_1^{(k)}(f) = P_{v_1^{(k)}v_1^{(1)}}(f) \big/ P_{v_1^{(k)}v_1^{(k)}}(f)$ , where the numerator and the denominator are cross and auto power spectral density (psd), respectively. The enhanced target signal is taken as $\hat{V}_1(f) = \frac{1}{K}\left( V_1^{(1)}(f) + \sum_{k=2}^{K} \overline{V}_1^{(k)}(f) \right)$.

For online applications, the filters are recursively estimated. Specifically, the psds and filters are estimated by utilizing two forgetting factors $\alpha_1, \alpha_2$ as

$$P_{v_1^{(k)} v_1^{(k')}}(f, mT) = \alpha_1 P_{v_1^{(k)} v_1^{(k')}}(f, (m-1)T) + (1-\alpha_1) V_1^{(k)}(f, mT) V_1^{(k')}(f, mT))^* \quad (6)$$

$$H_1^{(k)}(f, mT) = \alpha_2 H_1^{(k)}(f, (m-1)T) + (1-\alpha_2) P_{v_1^{(k)} v_1^{(1)}}(f, mT) \Big/ P_{v_1^{(k)} v_1^{(k)}}(f, mT) \quad (7)$$

where $T$ is the data window length for FFT analysis, $k'=1$ or $k$, and * denotes complex conjugation. The time-domain target signal is obtained by the standard overlap and add method. The forgetting factors $\alpha_1, \alpha_2$ were experimental chosen as 0.999 and 0.95 respectively, and T was chosen to have 2048 samples with 2048 zeros padded before FFT.

### 3.4. Integrated system for speech source separation

The speech source separation system that integrates the three enhancement techniques is shown in Fig. 3 Three pairs of microphones provide inputs to three ADF modules, and the inputs are each filtered by a whitening filter. In online application or nonstationary environment, ADF should be implemented in the block iterative mode, otherwise a batch iterative mode can be used to achieve a higher TIR gain (see section 4.4). The ADF outputs are dewhitened and the post-filtering module combines the target signals to generate an enhanced target signal which is then recognized by the ASR system.

## 4. EXPERIMENTS

### 4.1. Cochannel Condition and Data

Cochannel speech data were generated based on acoustic paths measured in real acoustic environment (RWCP) [11], and the source speech materials were taken from the TIMIT database. In RWCP, a circular microphone array with a radius of 15 cm was used to capture speech signals of two sources. The speaker-to-microphone distances were approximately 2 meters. Three pairs of microphones on the circular array, as shown in Fig. 3, were used in the experiment reported here: 14 and 4, 15 and 3, 16 and 2. The pair 15 and 3 was also used for the condition of single microphone pair and the distance between the two microphones was 21 cm. The recording room had a reverberation time of $T_{[60]} = 0.3$ sec . In the target speaker location, speech data of four speakers (faks0, felc0, mdab0, mreb0) were taken from the TIMIT database, each had ten sentences. In the jammer speaker location, speech data were randomly taken from the entire set of TIMIT sentences excluding those of the target speakers. Speech data were sampled at 16 KHz, and the ADF filter lengths were fixed as $N = 400$ .

Assume that the microphones at the locations 15 and 3 acquires speech mixture signals $y_1, y_2$ , respectively. The input target-to-interference ratio in $y_i$ , $TIR_{y_i}$ , is defined as the energy ratio (dB) of the target component $s_i$ in $y_i$ to the interference component

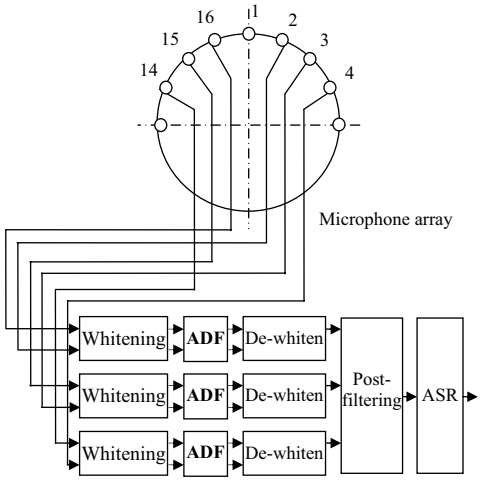


*Figure 3 Proposed system for speech source separation*

$s_j, j \neq i$ in $y_i$ . The ADF output $TIR_{v_i}$ are defined accordingly.

The initial conditions were $TIR_{y_1}$ = 3.01 dB and $TIR_{y_2}$ = -2.18 dB.

### 4.2. Evaluation of ADF Convergence Rate

The effects of prewhitening and block-iterative implementation on ADF convergence rate were evaluated by normalized filter estimation error on $\underline{f}_{12}^{(t)}$ and $\underline{f}_{21}^{(t)}$ . Fig. 4 hows three cases: baseline batch ADF, batch ADF with prewhitening, and block-iterative ADF with prewhitening. It is observed that prewhitening significantly speeded up convergence, and the block iterative implementation produced yet another significant improvement.

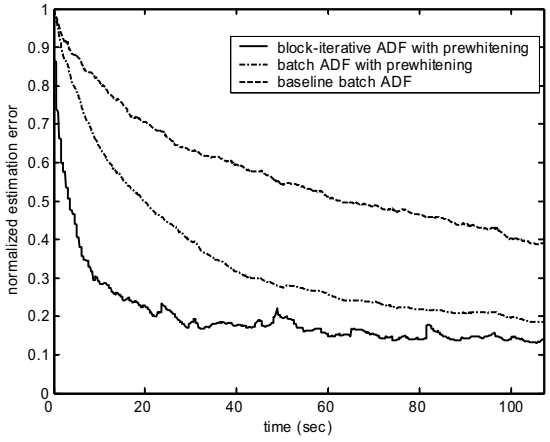


*Figure 4 Comparison of convergence rate*

### 4.3. Target-to-Interference Ratio

A comparison of TIRs with and without prewhitening processing is shown in Table 1, where to enable meaningful comparisons both the input and output speech were filtered by the prewhitening filters in calculating the TIRs. It is observed that the prewhitening filtering produced significantly faster improvement to output TIRs. The whitening weighted TIR data also better correlate with speech intelligibility since otherwise low frequency components that are

quality rather than intelligibility indicators of speech would dominate the TIR values.

*Table 1. Comparison of target-to-interference ratios (dB)*

| Estimation Passes | Baseline | | Prewhitening | |
|---|---|---|---|---|
| | $TIR_{v1}$ | $TIR_{v2}$ | $TIR_{v1}$ | $TIR_{v2}$ |
| 1 | 7.4 | 0.7 | 12.3 | 3.7 |
| 2 | 10.2 | 5.1 | 16.3 | 10.1 |
| 3 | 11.7 | 7.3 | 16.7 | 9.1 |
| 4 | 12.7 | 8.4 | 17.7 | 11.5 |
| 5 | 13.4 | 9.4 | 17.8 | 11.6 |
| 6 | 14.0 | 10.0 | 17.9 | 11.7 |
| 7 | 14.3 | 10.2 | 17.9 | 11.7 |

### 4.4. Phone Recognition Accuracy

Speech feature representation included 13 cepstral coefficients, and their first and second-order time derivatives. There were 39 context-independent phone units, with each unit modeled by three emission states of HMM, and each state had a size-8 Gaussian mixture density. Phone bigram was used as "language model." Cepstral mean subtraction was applied to training and test data. Phone recognition accuracies of clean TIMIT target speech and the mixed speech were 68.9% and 29.1%, respectively.

In Fig. 5 a comparison on phone recognition accuracy is made among the cases of (1) baseline batch ADF (2) batch ADF with preemphasis (3) batch ADF with prewhitening (4) block-iterative ADF with prewhitening, and (5) post filtering using three pairs of microphone data for each case of (1) through (4). Only one iteration pass was used for the batch ADF methods in order to compare with the block iterative method. It is observed that whitening improved recognition accuracy over baseline where prewhitening is superior to preemphasis, block-iterative is superior to batch, and post-filtering yielded significant gains over the single-microphone-pair cases of (1) through (4).

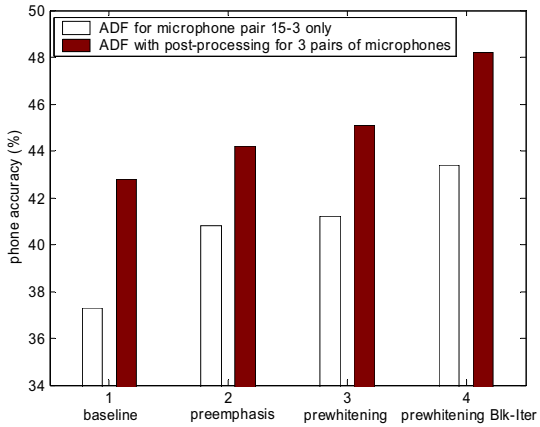


*Figure 5 Phone recognition accuracies with one pass of ADF*

In Fig. 6 phone recognition accuracy results are shown for batch ADFs with one to ten iteration passes. The cases are baseline, preemphasis, prewhitening, and with and without post filtering for each case. Again, whitening filtering improved baseline, post filtering improved the case of single-microphone-pair. Through long iterations the batch methods achieved a higher level of recognition accuracy than the block-iterative method.

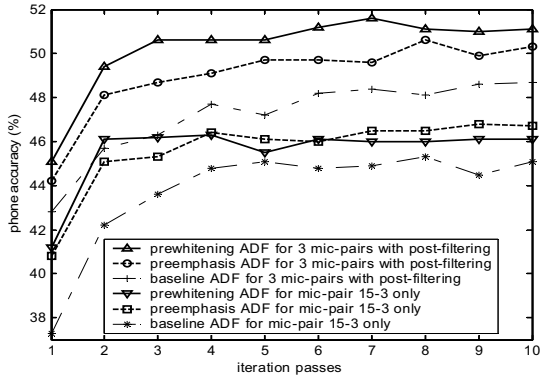


*Figure 6 Phone recognition accuracies with iterative batch ADF.*

## 5. CONCLUSION

The proposed techniques of whitening filtering, block-iterative implementation of ADF, and post filtering are shown to be simple and yet very effective for online speech source separation. Phone recognition accuracy has been significantly improved as compared with the baseline ADF system, and the fast convergence behavior of the proposed system allows tracking of time-varying sound source locations in reverberant room conditions.

## REFERENCES


[1]. E. Weinstein, M. Feder, and A. V. Oppenheim, "Multi-channel signal separation by decorrelation", *IEEE Trans. on SAP*, Vol. 1, pp. 405-413, Oct. 1993.

[2]. K. Yen Y. and Y. Zhao, "Adaptive co-channel speech separation and recognition,' *IEEE Trans. on SAP*, Vol. 7, No. 2, pp. 138-151, 1999.

[3]. K. Yen. and Y. Zhao, ``Adaptive decorrelation filtering for separation of co-channel speech signals from M > 2 sources," *Proc. ICASSP*, pp. 801-804, Phonex AZ, 1999.

[4]. L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources*," IEEE Trans. on SAP*, Vol. 8, No. 3, pp. 320--327, May 2000.

[5]. M. Z. Ikram and D. R. Morgan, "Exploring permutation inconsistency in blind separation of speech signals in a reverberant environment," *Proc. ICASSP*, pp. 1041--1044, Istanbul, Turkey, 2000.

[6]. R. Mukai, S. Araki, S. Makino, "Separation and dereverberation performance of frequency-domain blind source separation for speech in a reverberant environment," *Proc. EuroSpeech*, pp. 2599--2602, Alborg, Denmark, 2001.

[7]. J. Makhoul, "Linear prediction: a tutorial review," *Proc. IEEE*, vol. 63, pp. 561--580, Apr. 1975.

[8]. L. Rabiner and R. W. Schafer, **Digital Processing of Speech**, Prentice Hall, 1978.

[9] G.C. Carter, "Coherence and time delay estimation," *Proc. IEEE*, Vol.75, No.2, pp.236--255, 1987.

[10] R.L. Bouquin and R. Faucon, "Using the coherence function for noise reduction," *IEE Proceedings-I*, Vol.139, No.3, pp276--280, June 1992.

[11]. RWCP Sound Scene Database in Real Acoustic Environments, ATR Spoken Language Translation Research Laboratory, Japan 2001.