WHITENING PROCESSING FOR BLIND SEPARATION OF SPEECH SIGNALS

Yunxin Zhao, Rong Hu, and Satoshi Nakamura†

Department of CECS, University of Missouri, Columbia, MO 65211, USA
ATR Spoken Language Translation Research Labs, Kyoto 619–0288, Japan†
zhaoy@missouri.edu rhq2c@mizzou.edu satoshi.nakamura@atr.co.jp†

ABSTRACT

Whitening processing methods are proposed to improve the effectiveness of blind separation of speech sources based on ADF. The proposed methods include preemphasis, prewhitening, and joint linear prediction of common component of speech sources. The effect of ADF filter lengths on source separation performance was also investigated. Experimental data were generated by convolving TIMIT speech with acoustic path impulse responses measured in real acoustic environment, where microphone-source distances were approximately 2 m and initial target-to-interference ratio was 0 dB. The proposed methods significantly speeded up convergence rate, increased target-to-interference ratio in separated speech, and improved accuracy of automatic phone recognition on target speech. The preemphasis and prewhitening methods alone produced large impact on system performance, and combined preemphasis with joint prediction yielded the highest phone recognition accuracy.

1. INTRODUCTION

Cochannel speech separation, or blind-source separation on simultaneous speech signals, has become an active area of research in recent years. Among various approaches, time-domain adaptive decorrelation filtering (ADF) [1,2,3] and frequency-domain independent component analysis (ICA) [4,5,6] have been heavily studied. In addition, frequency-domain informax was proposed to be combined with time-delayed decorrelation method [7] to speed up blind separation of speech mixtures, and natural gradient algorithm was extended [8] for separation of speech mixtures while preserving temporal characteristics of source signals.

In our previous work, ADF has been successfully extended and applied to cochannel speech recognition and assistive listening [2,9]. In these studies, multi-microphone configurations were arranged such that cross-coupled acoustic paths exerted heavier attenuation on source speech than did the direct paths, and the microphone-source distances of direct paths were not far.

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

In our recent investigation of a more challenging acoustic condition, where the microphone-source distances of direct paths were far and the attenuation levels of cross-coupled acoustic paths and direct paths were comparable, the performance of ADF was found to be significantly deteriorated. The difficulty can be attributed to the limitation of the ADF principle and the spectral characteristics of speech in the following three aspects. First, in ADF, acoustic paths need to be modeled by finite-impulse response (FIR) filters in order to reach correct solution. The increased length of FIR filters with long acoustic paths makes them less distinguishable from IIR filters. Second, ADF assumes the source signals to be uncorrelated. When the source signals are speech, even though the long-term cross correlations of sources are low, strong cross correlation among sources may occur for non-negligible time durations due to spectral similarities of speech sounds. Third, voiced speech have strong low frequency components. There is therefore a large spread of eigenvalues in the correlation matrices of source speech as well as those of speech mixtures. The spread of eigenvalues is known to slow down convergence rate of adaptive filtering, in general. Furthermore, it is well known that not all frequency components of speech are of importance to human perception or machine recognition, and separation processings that place more emphasis on perceptually important spectral regions are therefore of interests.

In the current work, whitening processing that is motivated by known spectral characteristics of speech is proposed to integrate with ADF to improve convergence rate and estimation condition for cochannel speech separation and recognition. The investigated techniques include preemphasis that is commonly used in linear predictive coding [10], prewhitening that is based on long-term speech spectral density [11], and joint linear prediction of common components of source signals that is developed in the current work. In addition, the effect of FIR filter length of estimated acoustic paths on ADF performance is also studied. Evaluation experiments were performed on phone recognition of separated speech by using a hidden Markov model based speaker-independent automatic speech recognition system, with the source speech materials taken from the TIMIT database. The proposed techniques significantly

improved system performance.

The rest of the paper is organized as following. In section 2, the ADF algorithm is briefly reviewed. In section 3, the whitening processing techniques are discussed. In section 4, experimental condition is described and results are provided, and in section 5 a conclusion is made.

2. OVERVIEW OF ADF

2.1. Cochannel Model

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

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

with $F_{ij}(z) = H_{ij}(z)/H_{jj}(z)$. The task of ADF is to estimate $F_{12}(z)$ and $F_{21}(z)$ so as to separate the source signals that are mixed in the acquired signals.

2.2. Adaptive Decorrelation Filtering

Based on the assumed source properties of zero mean and zero mutual correlation, perfect outputs of the separation system should also be mutually uncorrelated. Define $\underline{f}_{12}^{(t)}$ and $\underline{f}_{21}^{(t)}$ to be length-*N* FIR filters that correspond to $F_{12}(z)$ and $F_{21}(z)$ and are estimated at time t . The ADF algorithm generates output signals $v_i(t), i = 1, 2$, according to the equation

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

where $\underline{y}_j(t) = [y_j(t-\tau)]_{0\le\tau\le N-1}^T$. Taking decorrelation of system outputs as the separation criterion, i.e., $E[v_i(t)v_j(t-\tau)] = 0,\ i \neq j,\ \forall\tau$, the cross-coupled filters can be adaptively estimated as

$$\begin{aligned} \underline{f}_{12}^{(t+1)} &= \underline{f}_{12}^{(t)} + \mu(t)\underline{v}_2(t)v_1(t) \\ \underline{f}_{21}^{(t+1)} &= \underline{f}_{21}^{(t)} + \mu(t)\underline{v}_1(t)v_2(t) \end{aligned} \tag{3}$$

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

$$\mu(t) = \frac{2\gamma}{N\left(\hat{\sigma}_{y_1}^2(t) + \hat{\sigma}_{y_2}^2(t)\right)} \tag{4}$$

where $0 < \gamma < 1$, and $\hat{\sigma}_{y_1}^2(t), \hat{\sigma}_{y_2}^2(t)$ are short-time energy estimates of input signals $y_i(t),\ i = 1, 2$. When the filter estimates converge to true solution, the output signal $v_i(t)$ becomes a linearly transformed source signal $s_i(t), i = 1, 2$. Details of the ADF algorithm can be found in [1,2,3].

3. WHITENING PROCESSING METHODS

3.1. Preemphasis

Preemphasis is a first-order high-pass filter in the form $P(z) = 1 - \mu z^{-1}$, with $\mu \approx 1$. It's frequency response is shown in Fig. 1. It is commonly used as a preprocessing step in linear predictive coding of speech. In general, voiced speech has a 6–dB per octave spectral tilt with strong low frequency energy. This wide dynamic range causes ill conditioning in autocorrelation matrix of speech and hence difficulty in estimation of LPC parameters. Preemphasis improves the condition number of autocorrelation matrix and therefore makes high-order LPC parameters better estimated [10].

For ADF, preemphasis is performed on mixed speech $y_i(t),\ i = 1, 2$. Through this processing, the spectral tilt of source speech signals as well as their mixtures are compensated, thereby improving the convergence rate in adaptive estimation of cross-coupled acoustic path filters.

3.2. Prewhitening

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 based on the long-term speech power spectrum provided in [11]. The frequency response of the inverse filter, called whitening filter, is also shown in Fig. 1. It is observed that the whitening filter has a 6 dB per octave high-pass characteristics in the frequency range of 1 KHz to 5 KHz, and its low-frequency attenuation is less as steep as the preemphasis filter.

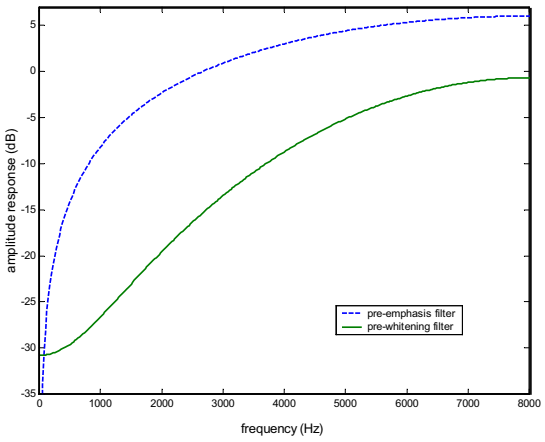

Figure 1 Frequency responses of preemphasis and prewhitening filters.

3.3. Joint Prediction of Source Signals

Formulation

Joint linear prediction as formulated here aims at dynamically whitening slow-varying common components of source signals so as to improve the input condition of ADF. In [12], joint prediction was used to whiten a reference signal component in a mixed signal with a lattice-ladder formulation. For estimation of a common-component prediction filter $\underline{a} = (a_1, a_2, \cdots, a_P)^T$, it is desired that the filter makes the prediction error of the source $s_i(t)$ be uncorrelated with the source $s_j(t),\ i \neq j$, i.e.,

$$E\left[\left(s_1(t) - \sum_{k=1}^{P} a_k\, s_1(t-k)\right) s_2(t-j)\right] = 0$$
$$E\left[\left(s_2(t) - \sum_{k=1}^{P} a_k\, s_2(t-k)\right) s_1(t-j)\right] = 0$$

for $j = 1, 2, \cdots, P$.

Define $r_{ij}(\tau) = r_{ji}(-\tau) = E[s_i(t)s_j(t-\tau)]$, and $\tilde{r}_{12}(j) = r_{12}(j) + r_{12}(-j)$. The system equation for solving the prediction parameters can be written as

$$\tilde{r}_{12}(j) = \sum_{k=1}^{P} a_k \tilde{r}_{12}(k-j)$$

Further define the cross-correlation matrix R_{12} to be a symmetric Toeplitz matrix with the diagonal elements being $\tilde{r}_{12}(0)$ and the jth subdiagonal (or superdiagonal) elements being $\tilde{r}_{12}(j),\ j = 1, 2, \cdots, P-1$, and the cross-correlation vector $\underline{r}_{12}$ to be $(\tilde{r}_{12}(1), \tilde{r}_{12}(2), \cdots, \tilde{r}_{12}(P))^T$. Then the matrix equation solution for $\underline{a}$ is $\underline{a} = R_{12}^{-1}\underline{r}_{12}$.

Since R_{12} is not always positive definite, solving $\underline{a}$ encounters difficulty when R_{12} becomes singular. This problem usually occurs when the cross correlations between two sources are low, such as in fricative segments of speech. To enable inversion of R_{12}, a positive constant δ is introduced to the determinant of R_{12} as: $\Delta = \det(R_{12}) + \delta \times \mathrm{sign}(\det(R_{12}))$. An obvious alternative is to simply turn off the prediction when R_{12} is found to be ill conditioned.

Implementation

In blind source separation, the prediction parameters need to be estimated from the separation outputs $v_i(t),\ i = 1, 2$ in an iterative fashion. Currently, the prediction parameters used in the k^{th} iteration to perform filtering on $y_i(t)$'s is computed from $v_i(t)$'s of the $(k-1)^{th}$ iteration. Within each iteration, cross-correlation statistics are computed from data blocks as $\tilde{r}_{12}^{(t)}(j),\ j = 1, 2, \cdots, P$, with t indexing the blocks. Averaged statistics are computed from a longer window as

$$\bar{r}_{12}^{(t)}(j) = \sum_{u=t-L}^{t} \rho^{(t-u)} \tilde{r}_{12}^{(u)}(j),\ j = 1, 2, \cdots, P$$

where ρ is a forgetting factor with value close to one. The prediction parameters $\underline{a}^{(t)}$ are computed from $\bar{r}_{12}^{(t)}(j)$'s and the mixed signals $y_i(t),\ i = 1, 2$ in the block t are filtered and used as inputs for ADF.

4. EXPERIMENTS

4.1. Cochannel Condition and Data

Cochannel speech data were generated based on acoustic path impulse responses measured in real acoustic environment [13], and the source speech materials were taken from the TIMIT database. The microphone-speaker configuration is shown in Fig. 2. At locations 3 and 15 were two microphones, and S_1 and S_2 denote target and jammer speakers, respectively. The speaker-to-microphone distances were approximately 2 meters, and the distance between the two microphones was 21 cm. The recording room had a reverberation time of $T_{[60]} = 0.3\ sec$. There were four target speakers (faks0, felc0, mdab0, mreb0), each spoke ten TIMIT sentences. Jammer speech were randomly taken from the entire set of TIMIT sentences excluding those of the target speakers. Speech data were sampled at 16 KHz.

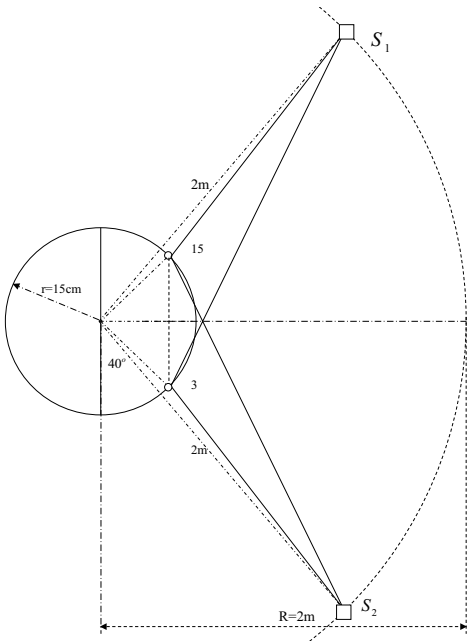

Figure 2 Microphone-speaker configuration of the acoustic environment.

Assume that the microphones at the locations 15 and 3 target respectively the speakers S_1 and S_2 with the acquired speech mixtures y_1 and y_2. The initial target-to-interference ratio in y_1, TIR_{y_1}, is defined as the energy ratio of the s_1 component in y_1 to the s_2 component in y_1, measured in dB. Similarly, the initial target-to-interference ratio in y_2, TIR_{y_2}, is defined as the energy ratio of the s_2 component in y_2 to the s_1 component in y_2. The ADF output TIRs in v_1 and v_2 are defined accordingly. Averaging over the test data of 40 TIMIT sentences, the initial conditions were $\mathrm{TIR}_{y_1} = 0.5327$ dB and $\mathrm{TIR}_{y_2} = -0.5576$ dB.

4.2. Whitening Effect on ADF Convergence

First, the effects of preemphasis and prewhitening on convergence rate of ADF were evaluated by normalized filter estimation error on $\underline{f}_{12}^{(t)}$ and $\underline{f}_{21}^{(t)}$. The results from the case of filter length $N = 400$ and $\gamma = 0.005$ are shown in Figure 3. Compared with the baseline condition of without whitening processing, preemphasis and prewhitening both significantly improved the convergence rate of ADF, with prewhitening having a larger effect.

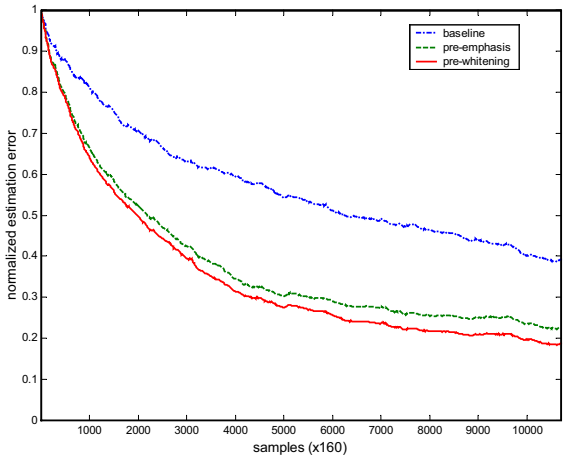

Figure 3 Convergence behavior of ADF without and with whitening processing.

The improved convergence rate of ADF can be attributed to the fact that whitening processing improved condition numbers of autocorrelation matrices of source signals, and in addition, it reduced cross correlation between source signals. In Fig. 4, the cross-correlation coefficients between two speech sources are shown for baseline and prewhitening, where prewhitening reduced cross correlation significantly. Experiments also showed similar effect from preemphasis and joint prediction.

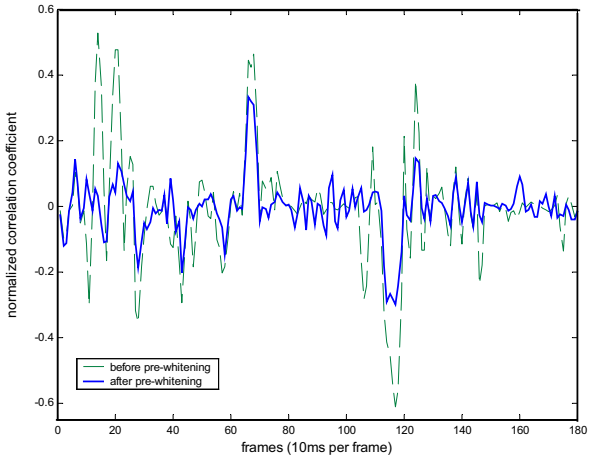

Figure 4 Cross-correlation coefficients between two speech sources without and with whitening processing.

4.3. Target-to-Interference Ratio

To enable meaningful comparison of TIRs with and without whitening processing, the output signals of the baseline system were filtered by the respective whitening filters in calculating the TIRs. In addition, initial TIRs in y_1 and y_2 were also recomputed by taking into account of the whitening effect of preemphasis or prewhitening, yeilding somewhat different initial TIRs in each case. ADF processing was performed using the same filter length N and the step size γ as in Section 4.2. Seven passes of ADF were performed over the test data, with the filter estimate obtained at the end of the current pass used as the initial estimate for the next pass. For each pass, average TIR_{v_i} was computed over the separation output v_i. In Table 1, a comparison is made between baseline and preemaphasis, with the recomputed initial TIRs of $\mathrm{TIR}_{y_1} = 3.01$ dB and $\mathrm{TIR}_{y_2} = -2.18$ dB.

Table 1 Comparison of output target-to-interference ratios (dB) between preemphasis and baseline.

Estimation Passes	Baseline		Preemphasis	
	TIR_{v_1}	TIR_{v_2}	TIR_{v_1}	TIR_{v_2}
1	7.88	3.12	10.17	4.55
2	11.07	7.64	12.25	8.22
3	12.20	8.90	13.68	9.30
4	12.69	9.25	13.90	9.48
5	13.04	9.66	14.00	9.60
6	13.33	9.88	14.09	9.67
7	13.43	9.86	14.13	9.72

In Table 2, a comparison is made between baseline and prewhitening, with the recomputed initial TIRs of $\mathrm{TIR}_{y_1} = 4.22$ dB and $\mathrm{TIR}_{y_2} = -3.01$ dB.

Table 2 Comparison of output target-to-interference ratios (dB) between prewhitening and baseline.

Estimation Passes	Baseline		Prewhitening	
	TIR_{v_1}	TIR_{v_2}	TIR_{v_1}	TIR_{v_2}
1	7.44	0.71	12.30	3.79
2	10.25	5.17	16.38	10.11
3	11.78	7.36	16.72	9.18
4	12.72	8.48	17.73	11.53
5	13.44	9.40	17.87	11.69
6	14.03	10.00	17.94	11.72
7	14.39	10.26	17.90	11.75

It is observed that the whitening processing produced significantly faster improvement to TIRs in both outputs v_1 and v_2 as compared with the baseline method. Although these TIR data were weighted by the whitening curves, they better correlate with intelligibility of separated speech since otherwise low frequency components that are quality rather than intelligibility indicators of speech would dominate the TIR values.

4.4. Phone Recognition Accuracy

For phone recognition, the ADF output of target speech was subjected to cepstral analysis and then recognized by the HMM-based speaker-independent phone recognition system. Feature vector size was 39, including 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 an observation pdf of size-8 Gaussian mixture density. Phone bigram was used as "language model." Cepstral mean subtraction was applied to training and test data.

With this setup, the phone recognition accuracies of clean TIMIT target speech, the target speech after passing through the direct channel H_{11}, and the mixed speech $y_1(t)$ were found to be 68.9%, 57.5%, and 29.1%, respectively.

Effects of ADF filter lengths

Although the impulse responses of measured acoustic path H_{ij}'s were on the order of 2000 samples, in order to enable ADF to converge to correct solution of $\underline{f}_{12}$ and $\underline{f}_{21}$, various FIR filter lengths were first evaluated for the baseline condition of performing ADF without whitening processing. The results are summarized in Table 3 for the case of $\gamma = 0.005$. It is observed that intermediate filter lengths of 400 to 600 taps yielded best results. With long filters, divergence occurred within a few iterations of ADF estimation (shown as X's).

Table 3 Phone accuracy (%) vs. ADF filter length for ADF without whitening processing

Passe1	1	2	3	4	5	6
N=200	38.7	43.3	43.7	44.0	44.4	44.3
N=400	37.3	42.2	43.6	44.8	45.1	44.8
N=600	37.1	40.0	42.5	42.4	42.5	43.7
N=800	36.4	39.9	X	X	X	X
N=1000	36.8	39.0	39.6	X	X	X
N=1200	35.6	38.1	X	X	X	X

Effects of whitening processing

The proposed whitening processing methods were used to process the mixed speech inputs, and ADF was then performed with filter length of $N = 400$ and adaptation step size of $\gamma = 0.005$. The separation output of the target speech $v_1(t)$ was recognized by the phone recognition system. In Fig. 5, recognition results vs. ADF iteration passes are shown for the following cases:

a. baseline — ADF without whitening processing
b. joint prediction of P = 2
c. preemphasis
d. prewhitening
e. preemphasis combined with joint prediction of P = 2
f. preemphasis combined with joint prediction of P = 3

As a reference, the filters $\underline{f}_{12}$ and $\underline{f}_{21}$ were also computed from the measured H_{ij}'s and then truncated to 400 taps, and the approximated FIR filters were then used for speech separation according to Eq.(2). In such a case, phone recognition accuracy on target speech was 53.1%. This performance figure set an upper limit to the achievable accuracy by the ADF separation system.

It is observed that the proposed whitening processing methods of cases b through e all improved the baseline results. Preemphasis and prewhitening alone produced large impact, and the combination of preemphasis with joint prediction of source signals yielded the best results. The performance of joint prediction alone was inferior to those of preemphasis and prewhitening, indicating that the colored speech spectrum is a dominating factor in slowing down ADF estimation, and the cross-correlation between speech sources is a secondary factor. In addition, initially $v_i(t)$'s were unavailable, and hence the joint prediction was not performed. In case b, the joint prediction performance was limited by the reliability of $v_i(t)$'s produced in the first iteration.

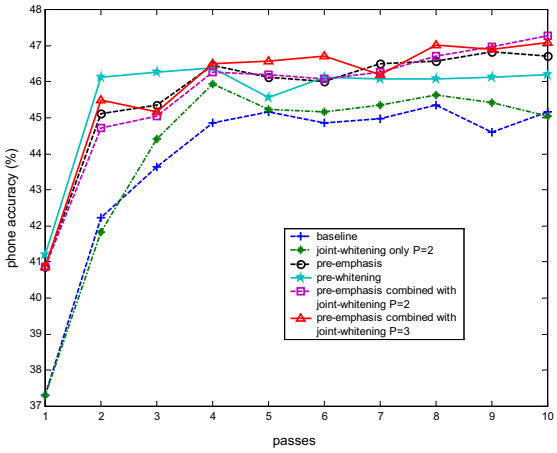

Figure 5 Phone recognition accuracy with various whitening processing methods.

It is worth mentioning that the convergence rate of ADF is adjustable by the adaptation step size parameter γ. Both preemphasis and prewhitening were observed to be able to survive γ values up to 0.015, with accompanied faster initial convergence rate. As a contrast, ADF without whitening processing would diverge quickly when using such larger γ values.

5. CONCLUSION

In the current work, whitening processing methods are proposed for integration with ADF-based blind separation of source speech signals. It is shown that under difficult cochannel acoustic conditions, directly processing speech inputs by ADF suffers from a poor convergence performance. Preemphasis and prewhitening are not only simple and effective methods for improving condition number of autocorrelation matrix of source speech, and they also reduce cross correlation between speech sources. As the result, their integration with ADF led to significant speed up of convergence rate. In addition, the deemphasis on low-frequency components of speech allow better source separation of spectral regions with perceptual importance and thereby increased phone recognition accuracy on separated speech. The joint prediction method is shown useful when combined with preemphasis, as it further reduced cross correlation between source speech signals. The implementation of joint prediction needs to be modified for online application, and alternative estimation criteria such as ICA or higher-order statistics might be formulated to avoid difficulty of inverting cross correlation matrix R_{12}. Further work is under way to improve convergence rate of the speech separation system and accuracy of the speech recognition system for online applications.

ACKNOWLEDGMENT

The authors would like to thank Xiaodong He and Xiaolong Li of CECS Department, University of Missouri for their help with the phone recognition experiments.

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. of EuroSpeech*, pp. 2599–2602, Aalborg, Denmark, 2001.

[7]. T-W. Lee, A. Ziehe, R. Orglmeister and T. J. Sejnowski, "Combining time-delayed decorrelation and ICA: towards solving the cocktail party problem," *Proc. of ICASSP*, pp. 1249–1252, Seattle, WA, 1998.

[8]. S. C. Douglas and X. Sun, "A natural gradient convolutive blind source separation algorithm for speech mixtures," *Proc. 3rd IEEE Int. Workshop on ICASS*, pp. 59–64, San Diego, CA, 2001.

[9]. Y. Zhao, K. Yen, S. Soli, S. Gao, and A. Vermiglio, "On application of adaptive decorrelation filtering to assistive listening," *J. Acoustic. Soc. Amer.*, Vol. 111, No. 2, pp. 1077–1085, Feb. 2002.

[10]. J. Makhoul, "Linear prediction: a tutorial review," *Proceedings of the IEEE*, vol. 63, pp. 561–580, Apr. 1975.

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

[12]. K. Yen and Y. Zhao, "Lattice-ladder structured adaptive decorrelation filtering for cochannel speech separation,"*Proceedings of ICASSP*, pp. 388–391, Istanbul, Turkey, June 2000.

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