CO-CHANNEL SPEECH SEPARATION FOR ROBUST AUTOMATIC SPEECH RECOGNITION: STABILITY AND EFFICIENCY

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

A signal-separation front-end based on adaptive decorrelation filtering (ADF) was integrated with an HMM based speaker independent continuous speech recognition system for co-channel speech recognition. The ADF is improved by addressing the adaptation gain for system stability and efficiency: an upper bound of adaptation rate is derived for system stability, and an accelerated sequence of adaptation gain is introduced for system efficiency. The system was evaluated under simulated room acoustic conditions with both time-invariant and time-varying channels. It is shown that the system significantly improved the signal-to-interference ratio and the recognition word accuracy, and that the combination of the derived upper bound for adaptation rate with the accelerated adaptation gain sequence achieved the best performance for system stability and efficiency.

1. INTRODUCTION

The state-of-the-art techniques in automatic speech recognition (ASR) are still vulnerable in the presence of interferences [1]. Although many research efforts are currently focused on broadband noise sources, a more difficult problem 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], a newly appeared signal processing technique of adaptive decorrelation filtering (ADF) [3] [4] was used as a signal-separation front-end for improving the signal-to-interference ratio (SIR) of the desired 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. The acquired signals are first ADF-processed to separate out the co-channel speech signals; the separated signals are then analyzed by a cross-spectra based source detection algorithm [2] to determine the active region of each source. The separated speech signals in their respective active regions are recognized by an HMM-based speaker-independent coutinuous speech recognition (SICSR) system [5]. Our experiments in [2] showed that under a simulated acoustic environment which is nonreverberant and time-invariant, when the average SIR in both channels was around 8 dB, the integrated system achieved a recognition accuracy very close to that of the interference-free condition.

In the current work, we extend the co-channel speech recognition technique into handling time-varying coupling channels that simulate certain room-acoustic environments. Specifically, a stability condition of adaptation rate is derived for the ADF algorithm; the effects of the adaptation rate on the estimation accuracy of time-varying channels and co-channel speech recognition are evaluated; and a strategy of adjusting adaptation gain for acceleration of convergence is also described. This paper is organized into six sections. In sections 2 and 3, the co-channel system and the ADF algorithm are briefly described. In section 4, an upper bound of adaptation rate for the ADF is derived, and the acceleration strategy for adaptation is proposed. Experimental results are presented in section 5 and a conclusion is made in section 6.

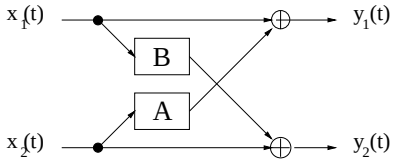

Figure 1. *The block diagram of the co-channel system*

2. CO-CHANNEL SYSTEM

Assume that the coexistent signal sources are independent to each other. For simplicity, our discussion is limited to the two-source two-microphone case. Let $x_1(t)$ and $x_2(t)$ be the signals generated by the independent sources 1 and 2, and let $y_1(t)$ and $y_2(t)$ be the signals acquired by the microphones that target the sources 1 and 2, respectively. Letting the linear filters A and B model the channel coupling effects, and assuming that there is no distortion between each microphone and its target source, the co-channel system can be described in the frequency domain as:

$$\begin{aligned} Y_1(f) &= X_1(f) + A(f)X_2(f) \\ Y_2(f) &= X_2(f) + B(f)X_1(f) \end{aligned} \tag{1}$$

This co-channel system is illustrated in Fig. 1.

3. SIGNAL SEPARATION BY ADAPTIVE DECORRELATION FILTERING

Define the filter $C(f) = 1 - A(f)B(f)$, and define the Fourier transforms of the signals $v_1(t)$ and $v_2(t)$ as:

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

It is easy to verify that $V_i(f) = C(f)X_i(f)$, $i = 1, 2$. Therefore, if the filters A and B are known, the signals from the sources 1 and 2 can be separated from the acquired signals by Eq. (2). Furthermore, if $C(f)$ is invertible, the source signals $x_i(t)$ can be perfectly reconstructed by $\hat{X}_i(f) = C(f)^{-1}V_i(f)$, $i = 1, 2$. The complete separation system is illustrated by Fig. 2.

Since in most applications, the filters A and B are time-varying and unknown, reducing the cross-channel interference requires reliable estimates of these filters. It was shown in [3] that if the source signals are zero-mean and uncorrelated and if the filters A and B are approximated by the

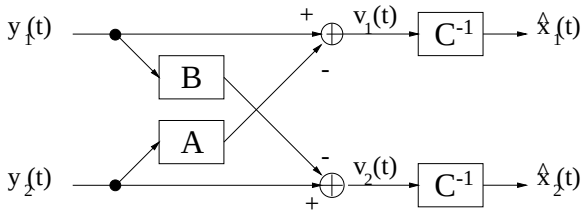

Figure 2. *The block diagram of the source separation system*

FIR filters $\underline{a} = [a_0, \cdots, a_{N_a-1}]^T$ and $\underline{b} = [b_0, \cdots, b_{N_b-1}]^T$, then the filter coefficients can be estimated interatively by the following equations, with the superscript (t) denoting the estimates at time t, and T denoting vector transpose:

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

where

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

with

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

The adaptation gain $\mu(t)$ is defined as γ/t.

4. ADAPTATION GAIN

In implementing the ADF algorithm of Eq. (3), the adaptation rate γ in the gain $\mu(t)$ is proven to be an important factor that determines the stability and efficiency of the system. Generally speaking, if the coupling-channel is changing fast, using a larger γ enables the system to follow the time variation better. On the other hand, if the coupling channel is changing slowly or remains steady, using a smaller γ enables the system to obtain more stable estimates of the channel. Furthermore, instability could be resulted if γ exceeds certain value.

In this section, a conservative upper bound for γ is first derived for ensuring system stability; then an accelerated adaptation gain sequence other than $\mu(t) = \gamma/t$ is discussed to better accommodate time-varying coupling channels.

4.1. An Upper Bound for γ

By expanding Eq. (3) and ignoring the quadratic terms of $\underline{a}$ and $\underline{b}$, the following equation is derived for updating the filter coefficients:

$$\underline{w}^{(t)} = \underline{w}^{(t-1)} + \frac{\gamma}{t}\underline{h}(t) - \frac{\gamma}{t}R(t)\underline{w}^{(t-1)} \tag{4}$$

where

$$\underline{w}^{(t)} = \begin{bmatrix} \underline{a}^{(t)} \\ \underline{b}^{(t)} \end{bmatrix}; \qquad \underline{h}(t) = \begin{bmatrix} \underline{y}_2(t)y_1(t) \\ \underline{y}_1(t)y_2(t) \end{bmatrix}$$

$$R(t) = \begin{bmatrix} \underline{y}_2(t)\underline{y}_2(t)^T & y_1(t)C_1(t) \\ y_2(t)C_2(t) & \underline{y}_1(t)\underline{y}_1(t)^T \end{bmatrix}$$

with

$$\begin{aligned} C_1(t) &= [\underline{y}_1(t), \cdots, \underline{y}_1(t - N_a + 1)]^T \\ C_2(t) &= [\underline{y}_2(t), \cdots, \underline{y}_2(t - N_b + 1)]^T \end{aligned}$$

Defining $\Delta\underline{w}(t)$ as the estimation error of $\underline{w}^{(t)}$, it can be derived that

$$E\{\Delta\underline{w}(t)\} = \left(I - \frac{\gamma}{t}E\{R(t)\}\right)E\{\Delta\underline{w}(t-1)\}$$

In order to maintain stability, it is necessary that $E\{\Delta\underline{w}(t)\}$ be reduced toward zero for sufficiently large t. To satisfy this condition, γ should be limited under $2/\lambda_{max}$, where λ_{max} is the largest eigenvalue of $E\{R(t)\}$. Since $trace(E\{R(t)\}) = N_a var\{y_2(t)\} + N_b var\{y_1(t)\} \geq \lambda_{max}$, the following bound can be used for γ to avoid the calculation of the eigenvalues:

$$0 < \gamma < \frac{2}{N_a var\{y_2(t)\} + N_b var\{y_1(t)\}} = \Gamma \tag{5}$$

Based on our experimental evaluation, the bound Γ works well for the ADF algorithm in most situations.

4.2. Accelerated ADF (AADF)

For efficient adaptation, it is desirable that when the previous estimates differ from the current channel filter coefficients significantly, a larger adaptation gain is applied; on the other hand, when the previous estimates are close to the current channel filter coefficients, a smaller adaptation gain is applied. Therefore, instead of using $\mu(t) = \gamma/t$ as the adaptation gain, we can apply Kesten's procedure of accelerating convergence [6] to modify the estimation equations as:

$$\begin{aligned} a_k^{(t)} &= a_k^{(t-1)} + \frac{\gamma}{i_{a,k}(t)}v_2^{(t-1)}(t-k)v_1^{(t-1)}(t) \\ b_k^{(t)} &= b_k^{(t-1)} + \frac{\gamma}{i_{b,k}(t)}v_1^{(t-1)}(t-k)v_2^{(t-1)}(t) \end{aligned} \tag{6}$$

where

$$i_{a,k}(1) = 1; \qquad k = 0, 1, \cdots, N_a - 1$$

$$i_{a,k}(t+1) = \begin{cases} i_{a,k}(t), & if \quad \tilde{a}_k^{(t)}\tilde{a}_k^{(t-1)} > 0 \\ i_{a,k}(t) + 1, & if \quad \tilde{a}_k^{(t)}\tilde{a}_k^{(t-1)} \leq 0 \end{cases}$$

$$i_{b,k}(1) = 1; \qquad k = 0, 1, \cdots, N_b - 1$$

$$i_{b,k}(t+1) = \begin{cases} i_{b,k}(t), & if \quad \tilde{b}_k^{(t)}\tilde{b}_k^{(t-1)} > 0 \\ i_{b,k}(t) + 1, & if \quad \tilde{b}_k^{(t)}\tilde{b}_k^{(t-1)} \leq 0 \end{cases}$$

and

$$\begin{aligned} \tilde{a}_k^{(t)} &= v_2^{(t-1)}(t-k)v_1^{(t-1)}(t) \\ \tilde{b}_k^{(t)} &= v_1^{(t-1)}(t-k)v_2^{(t-1)}(t) \end{aligned}$$

As defined in Eq. (6), the signs (positive, negative) of the consecutive correlation terms controls the adjustment of the adaptation gain for each filter coefficient, where the gain decreases only when the sign changes. For stability, the adaptation rate γ should also be bounded by Γ as derived in section 4.1.

5. EXPERIMENT

In order to handle time-varying channels, the ADF and AADF were implemented blockwisely (referred to as BADF and BAADF, respectively), i.e., the aquired signals from the two microphones were synchronously blocked into frames; the adaptation rate was decided in each frame; and the time t in both Eqs. (3) and (6) was reset to zero at the beginning of each frame. The estimates of $\underline{a}$ and $\underline{b}$ at the end of the current frame were used as the initial values of the next frame, and the initial values in the first frame were simply set to zero. A subset of TIMIT database was chosen to form 156 sentence pairs as the source signals in the following experiments.

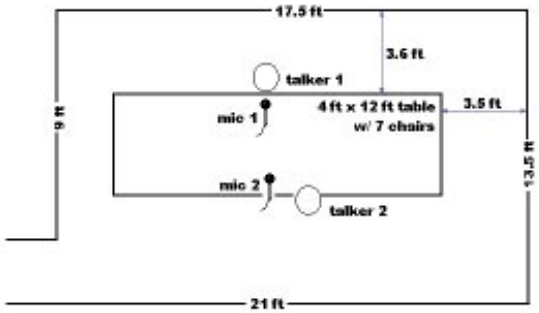

Figure 3. *Room-acoustic environment 1,* (A_1, B_1)

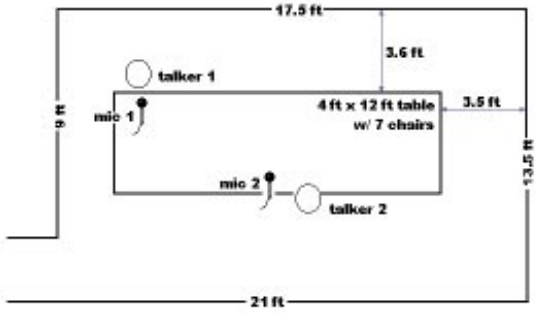

Figure 4. *Room-acoustic environment 2,* (A_2, B_2)

5.1. Stability and Adaptation Rate

In this experiment, one set of co-channel signals was processed by BADF, with each frame containing 200 samples. Three adaptation rates γ were tested:

$$\gamma_1 = \frac{2}{\sqrt{N_a N_b var\{y_1\} var\{y_2\}}}$$

$$\gamma_2 = \frac{2}{max\big(N_a var\{y_2\}, N_b var\{y_1\}\big)}$$

$$\gamma_3 = \frac{2}{N_a var\{y_2\} + N_b var\{y_1\}}$$

The stability of the system was examined after each step of adaptation (i.e., every sample). Once the system became unstable, the filter coefficients were reset to zero and the estimation restarted. In more than 40 million iterations, the system was reset 23 times when using γ_1, 3 times when using γ_2, and 0 times when using γ_3. The bound in Eq. (5) is therefore considered as a safe choice.

5.2. Simulation of Acoustic Paths

Two pairs of FIR filters (A_1, B_1) and (A_2, B_2) were measured to simulate the acoustic paths in the room environments described in Figs. 3 and 4:

A_1 : from talker 2 to microphone 1 in Fig. 3

B_1 : from talker 1 to microphone 2 in Fig. 3

A_2 : from talker 2 to microphone 1 in Fig. 4

B_2 : from talker 1 to microphone 2 in Fig. 4

The distortions from talker 1 to microphone 1 and from talker 2 to microphone 2 in both environments were assumed neglectable. In the experiments described below, the first L samples (L was varied in different experiments) of the impulse response of each filter were included in the FIR filter for generating the co-channel signals from the source signals.

5.3. Estimation Accuracy and Adaptation Gain

In this experiment, the co-channel speech signal pairs were generated by (A_1, B_1) (L=100) from the source signals, and were processed by the following three schemes:

1. the BADF; $\gamma = \Gamma$
2. the BADF; $\gamma = 0.5\Gamma$
3. the BAADF; $\gamma = 0.5\Gamma$

To evaluate the performance of the BADF, the squared estimation error of the filter coefficients, $E(t)$, was defined as:

$$E(t) = \left[\Delta \underline{a}^{(t)}\right]^T \Delta \underline{a}^{(t)} + \left[\Delta \underline{b}^{(t)}\right]^T \Delta \underline{b}^{(t)}$$

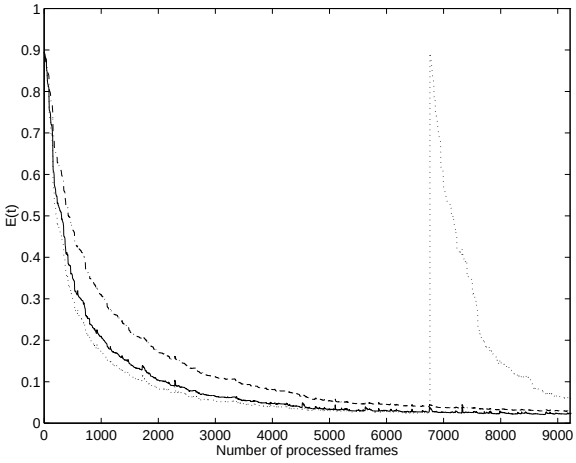

Figure 5. *The squared estimation error versus the number of processed frames: dotted curve: BADF with* $\gamma = \Gamma$*; dashed curve: BADF with* $\gamma = 0.5\Gamma$*; and solid curve: BAADF with* $\gamma = 0.5\Gamma$

with

$$\Delta \underline{a}^{(t)} = \underline{a}^{(t)} - \underline{a}^* \quad and \quad \Delta \underline{b}^{(t)} = \underline{b}^{(t)} - \underline{b}^*$$

where $*$ denotes the true filter coefficients, and $E(t)$ was measured at the end of each frame. The relation between $E(t)$ and the number of processed frames is ploted in Fig. 5 for all three cases. Since the filter coefficients were all initialized as zeros, the beginning part of each curve represents the system behavior for a fast-changing channel, and the ending part represents the system behavior for a time-invariant channel. Comparing the results of the BADF with $\gamma = \Gamma$ and 0.5Γ, $E(t)$ was reduced much faster at the beginning with $\gamma = \Gamma$, but became more stable at the end with $\gamma = 0.5\Gamma$. The results also showed that the BAADF with $\gamma = 0.5\Gamma$ could follow the channel variation as well as BADF with $\gamma = \Gamma$, and was as stable as BADF with $\gamma = 0.5\Gamma$ when the channel became steady.

5.4. Time-Invariant Channel Simulation

In this experiment, the co-channel speech signals were processed by BADF with $\gamma = 0.1\Gamma$, and then recognized by the SICSR system based on the HMM of phone units [5]. The cepstrum coefficients of the PLP analysis (8th order) and log energy were taken as instantaneous features and their first-order 50 msec temporal regression coefficients as dynamic features. The recognition task has vocabulary size of 853 and grammar perplexity of 105. Three sets of co-channel signals were generated, processed, and then recognized:

Table 1. *The SIR and WRA of the simulated time-invariant channel conditions*

	Channel 1	Channel 2
Clear source signals	--	91.2 dB
Filter length = 100		
SIR before processing	10.9 dB	11.7 dB
SIR after processing	25.8 dB	25.2 dB
WRA before processing	41.7 %	38.1 %
WRA after processing	85.5 %	83.8 %
Filter length = 200		
SIR before processing	10.9 dB	11.6 dB
SIR after processing	25.4 dB	25.1 dB
WRA before processing	41.8 %	36.1 %
WRA after processing	84.2 %	82.6 %
Filter length = 100		
SIR before processing	20.9 dB	1.7 dB
SIR after processing	35.2 dB	17.5 dB
WRA before processing	64.6 %	-10.1 %
WRA after processing	90.0 %	71.0 %

1. The coupling filters were A_1 and B_1 with L=100. The source signals in both channels had the same energy level.
2. The coupling filters were A_1 and B_1 with L=200, where the magnitude of the last 100 samples were attenuated by 1/3 to reduce the reverberation effect. The source signals in both channels had the same energy level.
3. The coupling filters were A_1 and B_1 with L=100. The source signals in channel 2 were 10 dB weaker than the source signals in channel 1.

The SIR and word recognition accuracy (WRA) before and after processing for each case are summarized in Table 1. It can be observed that the system improved both SIR and WRA significantly under the simulated conditions.

5.5. Time-Varying Channel Simulation

In this experiment, when producing the co-channel speech signals, the coupling filters were made to change from (A_1, B_1) toward (A_2, B_2) in an N-sample interval using linear interpolation, and then change back at the same rate. This filter-changing process continued until the end of signals. Two sets of co-channel speech signals were produced with $N = 10^4$ and $N = 10^6$, respectively (corresponding to 1 and 100 seconds with a 10 kHz sampling rate). Each set of speech signals were processed by the following three schemes:

1. the BADF, $\gamma = \Gamma$
2. the BADF, $\gamma = 0.1\Gamma$
3. the BAADF, $\gamma = 0.1\Gamma$

The separated signals within their respective active regions were then recognized by the SICSR system described in the previous experiment. The SIR and WRA are summarized in Table 2. It is seen that in the fast-changing environment, the BADF with $\gamma = \Gamma$ performed better than the BADF with $\gamma = 0.1\Gamma$, but not as well in the slowly-changing environment, both in terms of SIR and WRA. The BAADF with $\gamma = 0.1\Gamma$ achieved similar performances as the better one of the two BADFs in both environments, which is consistant with the result in section 5.3.

6. CONCLUSION

The current work shows that the proposed co-channel speech separation front-end is promising for robust speech recognition under the simulated room-acoustic environments. The combination of the derived upper bound for the adaptation rate with the accelarated adaptation gain sequence achieved the best performance for system stability and efficiency. Extended research is currently underway to explore the reverberation effects in the cross-channel interference.

Table 2. *The SIR and WRA of the simulated time-varying channel conditions*

	SIR	WRA
Clear source signals	--	91.2 dB
Fast-changing channel ($N = 10^4 \simeq 1sec$)		
Before processing	14.9 dB	58.4 %
BADF, $\gamma = \Gamma$	19.2 dB	78.0 %
BADF, $\gamma = 0.1\Gamma$	17.1 dB	68.4 %
BAADF, $\gamma = 0.1\Gamma$	18.8 dB	76.4 %
Slowly-changing channel ($N = 10^6 \simeq 100sec$)		
Before processing	15.1 dB	59.4 %
BADF, $\gamma = \Gamma$	21.9 dB	84.8 %
BADF, $\gamma = 0.1\Gamma$	25.7 dB	88.7 %
BAADF, $\gamma = 0.1\Gamma$	24.9 dB	87.6 %

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, and the suggestion by Mr. Shaojun Wang of the Beckman Institute, UIUC, are also acknowledged.

REFERENCES

[1] R. Cole et al, "The Chanllenge 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] 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.

[4] 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.

[5] Y. Zhao, "A Speaker-Independent Continuous Speech Recognition System Using Continuous Mixture Gaussian Density HMM of Phoneme-Sized Units," *IEEE Trans. on Speech and Audio Processing*, Vol. 1, No. 3, pp. 345-361, Jul. 1993.

[6] H. Kesten, "Accelerated Stochastic Approximation," *Ann. Math. Statist.*, No. 29, pp. 41-59, 1958.