Variable Step Size Adaptive Decorrelation Filtering for Competing Speech Separation

Rong Hu and Yunxin Zhao

Department of Computer Science
University of Missouri-Columbia, USA
rhq2c@mizzou.edu zhaoy@missouri.edu

Abstract

Two variable step size (VSS) techniques are proposed for adaptive decorrelation filtering (ADF) to improve the performance of competing speech separation. The first VSS method applies gradient adaptive step-size (GAS) to increase ADF convergence rate. Under some simplifying assumptions, the GAS technique is generalized to allow the combination with additional VSS techniques for ADF algorithm. The second VSS method is based on error analysis of ADF estimates under a simplified signal model to decrease steady state filter error. An integration of both techniques into ADF was tested with TIMIT speech data convolutively mixed by reverberant room impulse responses. Experimental results showed that the proposed algorithm significantly increased ADF convergence rate and improved gain in both target-to-interference ratio (TIR) and phone recognition accuracy of the target speech.

1. Introduction

Multi-channel separation of co-channel speech signals is an important and challenging problem in hands-free automatic speech recognition (ASR) as well as in speech communication. Time-domain adaptive de-correlation filtering (ADF) [1-4] is one of the promising approaches for this task. Although attractive in simplicity of structure and ease of implementation, for practical applications under acoustic conditions of long reverberation time, the enhancement gain and convergence rate of ADF still need to be improved.

In our previous work [3], a pre-whitening technique was proposed to improve working condition and stability of ADF, and a block-iterative scheme was used to accelerate learning rate. The gain-normalized ADF was proposed in [4], where normalization of adaptation gain with respect to input energy produced an effective variable step size (VSS) sequence and the adaptive estimation was proven to converge. Observing the relationship between ADF and LMS, [4] also provided a coherence-function based scheme that switched between these two algorithms depending on whether both speech sources were present or not.

Earlier studies of Thi and Jutten [5] proposed output-normalized VSS. The so-called "non-permanent learning" was used to stop updating certain filters when some criterion on output energy was satisfied. The update-stopping technique was intuitive and may suffer from error of the hard-decisions on speaker-presence.

It is well known from the analysis of the LMS family of adaptive algorithms [6-8] that better trade-off between convergence rate and mean steady-state error (MSE) could be achieved by suitable choices of step-sizes. The gradient-adaptive step-size (GAS) algorithm [6] was proposed as one of the many VSS LMS algorithms for such a purpose. Douglas and Cichocki [9] introduced GAS into a natural gradient based blind source separation (BSS) algorithm. They classified VSS into two types: "adaptive" and "non-adaptive". Adaptive (e.g., GAS) methods are based on on-line measurements of system state and were emphasized for the ability of quickly tracking unknown non-stationary environments, while non-adaptive methods compute step size according to *a priori* knowledge about the system and can achieve higher performances, e.g., lower MSE, from such additional information [9].

In the current work, we propose two VSS techniques to further improve ADF algorithm for effective separation of competing cochannel speech. Contrary to many existing methods that isolated non-adaptive from adaptive VSS, we combine them to achieve faster learning rate with lower MSE. For this purpose, a vector representation of the system is given, based on which the derivation of GAS gain is made. An ADF error analysis provides *a priori* knowledge for the proposed non-adaptive gain factor. The integrated VSS-ADF algorithm is evaluated by speech separation and phone recognition experiments.

2. Fundamentals

2.1. System models

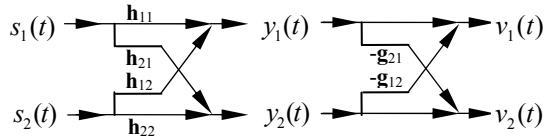

Figure 1. Speech mixing and ADF separation system models.

The two-speaker-two-microphone signal mixing and ADF system models are shown in Fig. 1, in which the cross-coupling filters to be estimated are $\mathbf{g}_{ij}=[g_{ij}(0), \ldots, g_{ij}(N\text{-}1)]^T$, with $(\cdot)^T$ for transpose. In the following derivations, variables in bold capital case are for matrices, bold lower case for vectors, $E\{\cdot\}$ for expectation, the correlation vector of a signal sample $a(t)$ and a signal vector $\mathbf{b}(t)$ is denoted by $\mathbf{r}_{ab} = E\{a(t)\mathbf{b}(t)\}$, and the cross correlation matrix of two signal vectors is denoted by $\mathbf{R}_{\mathbf{ab}} = E\{\mathbf{a}(t)\mathbf{b}^T(t)\}$.

The speech mixing model can be described as:

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

where the ideal cross-coupling filter from the jth speaker to the ith microphone is $G_{ij}^o(z) = H_{ij}(z)/H_{jj}(z)$, $i,j=1,2$ $i \neq j$. Under FIR assumption for the coupling filters and adopting

the vector representation of [10], the time-domain speech mixing model can be described by

$$\widetilde{\mathbf{y}} = \widetilde{\mathbf{H}} \cdot \bar{\mathbf{s}}, \tag{2}$$

where $\widetilde{\mathbf{y}}=\left[\widetilde{\mathbf{y}}_1^T(t)\ \ \widetilde{\mathbf{y}}_2^T(t)\right]^T$ with $\widetilde{\mathbf{y}}_i(t)=[y_i(t),\cdots,y_i(t-N+1),\cdots,y_i(t-2N+2)]^T$, $\bar{\mathbf{s}}=\left[\bar{\mathbf{s}}_1^T(t)\ \ \bar{\mathbf{s}}_2^T(t)\right]^T$ with $\bar{\mathbf{s}}_j=\left(s_j(t),\ \cdots,\ s_j(t-4N+4)\right)^T$, and $\widetilde{\mathbf{H}}=\begin{bmatrix}\widetilde{\mathbf{H}}_{11} & \widetilde{\mathbf{H}}_{12}\\ \widetilde{\mathbf{H}}_{21} & \widetilde{\mathbf{H}}_{22}\end{bmatrix}$ with

$$\widetilde{\mathbf{H}}_{ij}=\begin{bmatrix} h_{ij}(0) & \cdots & h_{ij}(2N-2) & 0 & \cdots & 0\\ 0 & h_{ij}(0) & \cdots & h_{ij}(2N-2) & \cdots & 0\\ \vdots & & & & & \vdots\\ 0 & \cdots & 0 & h_{ij}(0) & \cdots & h_{ij}(2N-2)\end{bmatrix}.$$

The time-domain ADF model has the vector form [10]

$$\mathbf{v} = \mathbf{G} \cdot \widetilde{\mathbf{y}}, \tag{3}$$

where output $\mathbf{v}=\left[\mathbf{v}_1^T(t)\ \ \mathbf{v}_2^T(t)\right]^T$ with $\mathbf{v}_i(t)=[v_i(t),\cdots,v_i(t-N+1)]^T$ and system matrix $\mathbf{G}=\begin{bmatrix}\left[\mathbf{I}_N\ \ \mathbf{0}_{N\times(N-1)}\right] & -\mathbf{G}_{12}\\ -\mathbf{G}_{21} & \left[\mathbf{I}_N\ \ \mathbf{0}_{N\times(N-1)}\right]\end{bmatrix}$, with

$$\mathbf{G}_{ij}=\begin{bmatrix} g_{ij}(0) & \cdots & g_{ij}(N-1) & 0 & \cdots & 0\\ 0 & g_{ij}(0) & \cdots & g_{ij}(N-1) & \cdots & 0\\ \vdots & & & & & \vdots\\ 0 & \cdots & 0 & g_{ij}(0) & \cdots & g_{ij}(N-1)\end{bmatrix},$$

and $\mathbf{I}_N$ denoting the $N\times N$ identity matrix, for $i,j=1,2\ \ i\neq j$.

The ADF system solution is given by [10]:

$$\begin{bmatrix}\mathbf{0} & \mathbf{R}_{\mathbf{y}_1\mathbf{v}_1}^T\\ \mathbf{R}_{\mathbf{y}_2\mathbf{v}_2}^T & \mathbf{0}\end{bmatrix}\cdot\begin{bmatrix}\mathbf{g}_{12}\\ \mathbf{g}_{21}\end{bmatrix}=\begin{bmatrix}\mathbf{r}_{y_2\mathbf{v}_1}\\ \mathbf{r}_{y_1\mathbf{v}_2}\end{bmatrix}. \tag{4}$$

2.2. Basic ADF algorithm

Solving Eq.(4) by minimization of output cross-correlation [10]

$$\mathbf{g}_{ij}^{opt} = \arg\min J_{ij} = \tfrac{1}{2}\left(\mathbf{r}_{v_i\mathbf{v}_j}^T\,\mathbf{r}_{v_i\mathbf{v}_j}\right), \tag{5}$$

the basic ADF algorithm [1, 10] is derived as the following pair of equations:

$$v_i(t) = y_i(t) - \mathbf{g}_{ij}^T(t)\mathbf{y}_j(t) \tag{6}$$

$$\mathbf{g}_{ij}(t+1) = \mathbf{g}_{ij}(t) + \mu(t)v_i(t)\mathbf{v}_j(t), \tag{7}$$

where $i,j=1,2\ \ i\neq j$. The choice of VSS $\mu(t)$ can be input-normalized [4] as

$$\mu(t) = 2\alpha\Big/N\left(\sigma_{y_1}^2(t)+\sigma_{y_2}^2(t)\right), \tag{8}$$

where the adaptation gain α is a constant, or output-normalized as in [5]. Other heuristic methods can also be used to determine the VSS sequence $\mu(t)$ with different convergence performances.

3. Variable step size ADF

3.1. Gradient adaptive gain

To introduce GAS [6] into ADF, we adapt step-size in the negative direction of the instantaneous gradient of ADF output cross-correlation and in the meantime, we introduce additional gain factors into the adaptation equation. The VSS ADF filter update thus becomes

$$\mathbf{g}_{ij}(t+1) = \mathbf{g}_{ij}(t) + \gamma_{ij}(t)\mu_{ij}(t)v_i(t)\mathbf{v}_j(t), \tag{9}$$

where the product of two separate variable gains, $\gamma_{ij}(t)$ and $\mu_{ij}(t)$ forms the overall step-size. The gain factor $\gamma_{ij}(t)$ represents the component of VSS that is adjustable through gradient adaptive procedures, and the gain factor $\mu_{ij}(t)$ absorbs the rest "non-adaptive" procedures (e.g., normalization as in (8)).

The update of the GAS gain-factor γ_{ij} at time t+1 aims at the minimization of the instantaneous criterion function

$$J_{\gamma_{ij}}(t+1) = \tfrac{1}{2}\left(v_i(t+1)\mathbf{v}_j(t+1)\right)^T v_i(t+1)\mathbf{v}_j(t+1), \tag{10}$$

and $\gamma_{ij}(t+1)$ is given by

$$\gamma_{ij}(t+1) = \gamma_{ij}(t) - \varepsilon\frac{\partial}{\partial\gamma_{ij}(t)}J_{\gamma_{ij}}(t+1)\cdot \tag{11}$$

From (6) and (9), $\mathbf{v}_j(t+1)$ is independent of $\gamma_{ij}(t)$. We then obtain

$$\frac{\partial}{\partial\gamma_{ij}(t)}J_{\gamma_{ij}}(t+1) = v_i(t+1)\mathbf{v}_j^T(t+1)\mathbf{v}_j(t+1)\cdot\frac{\partial}{\partial\gamma_{ij}(t)}v_i(t+1), \tag{12}$$

where using (6) for time t+1 and (9),

$$\frac{\partial}{\partial\gamma_{ij}(t)}v_i(t+1) = -\mu_{ij}(t)v_i(t)\mathbf{y}_j^T(t+1)\mathbf{v}_j(t)\cdot \tag{13}$$

The derivation of (13) also uses the simplifying assumption that the non-adaptive gain $\mu_{ij}(t)$ is independent of GAS gain factor $\gamma_{ij}(t)$. Such an assumption is valid when the non-adaptive gain is determined by some short-term statistics of the adaptive system, such as short-term variance of ADF inputs in (8). The reason is that $\mu_{ij}(t)$ measures system statistics of a very short time interval, e.g. 50ms, while GAS gain $\gamma_{ij}(t)$ measures automatically the state of the system and usually reflects convergence trends of longer term, e.g., in seconds, from observations of ADF system up to time t.

Finally, substituting (13) into (12) and then back into (11), the adaptation of GAS is obtained as

$$\gamma_{ij}(t+1)=\gamma_{ij}(t)+\varepsilon\mu_{ij}(t)v_i(t)v_i(t+1)\mathbf{v}_j^T(t+1)\mathbf{v}_j(t+1)\mathbf{v}_j^T(t)\mathbf{y}_j(t+1). \tag{14}$$

3.2. Gain factor derived from error analysis

3.2.1. ADF error analysis

For non-adaptive gain factor $\mu_{ij}(t)$, an effective choice can be made from the error analysis of ADF. Since speech signals are complex and direct analysis based on (2) and (3) is difficult, we consider the simplifying case that sources are white with zero mean, i.e.,

$$\mathbf{R}_{\bar{\mathbf{s}}\bar{\mathbf{s}}} = blkdiag\left(p_1\mathbf{I}_{4N-3}\quad p_2\mathbf{I}_{4N-3}\right), \tag{15}$$

where p_1 and p_2 are variances of sources s_1 and s_2 respectively.

It is obvious from (4) that the accuracy and stability of new ADF estimates $\mathbf{g}_{ij}^{est}$'s, based on current $\mathbf{g}_{ij}$'s, are determined by the following equation,

$$\mathbf{R}_{\mathbf{y}_j\mathbf{v}_j}^T\mathbf{g}_{ij}^{est} = \mathbf{r}_{y_i\mathbf{v}_j}. \tag{16}$$

Following the development in the Appendix under white-source assumption, we have

$$\mathbf{R}_{\mathbf{y}_j\mathbf{v}_j} = p_i\mathbf{T}_i^j + p_j\mathbf{T}_j^j = p_i\left(\mathbf{T}_i^j + p_j/p_i\cdot\mathbf{T}_j^j\right), \tag{17}$$

$$\mathbf{r}_{y_i\mathbf{v}_j} = p_i\boldsymbol{\eta}_i^j + p_j\boldsymbol{\eta}_j^j = p_i\left(\boldsymbol{\eta}_i^j + p_j/p_i\cdot\boldsymbol{\eta}_j^j\right), \tag{18}$$

with

$$\mathbf{T}_i^j = \left[\mathbf{I}_N\ \ \mathbf{0}_{N\times(N-1)}\right]\widetilde{\mathbf{H}}_{ji}\left(\left[\mathbf{I}_N\ \ \mathbf{0}_{N\times(N-1)}\right]\widetilde{\mathbf{H}}_{ji} - \mathbf{G}_{ji}\widetilde{\mathbf{H}}_{ii}\right)^T, \tag{19}$$

$$\mathbf{T}_j^j = \left[\mathbf{I}_N\ \ \mathbf{0}_{N\times(N-1)}\right]\widetilde{\mathbf{H}}_{jj}\left(\left[\mathbf{I}_N\ \ \mathbf{0}_{N\times(N-1)}\right]\widetilde{\mathbf{H}}_{jj} - \mathbf{G}_{ji}\widetilde{\mathbf{H}}_{ij}\right)^T, \tag{20}$$

$$\boldsymbol{\eta}_i^j = \left(\mathbf{H}_{ji} - \mathbf{G}_{ji}\widetilde{\mathbf{H}}_{ii}\begin{bmatrix}\mathbf{I}_{2N-1}\\ \mathbf{0}_{(2N-2)\times(2N-1)}\end{bmatrix}\right)\widetilde{\mathbf{h}}_{ii}, \tag{21}$$

$$\boldsymbol{\eta}_j^i = \left(\mathbf{H}_{jj} - \mathbf{G}_{ji} \widetilde{\mathbf{H}}_{ij} \begin{bmatrix} \mathbf{I}_{2N-1} \\ \mathbf{0}_{(2N-2)\times(2N-1)} \end{bmatrix} \right) \widetilde{\mathbf{h}}_{ij}, \quad (22)$$

where $\mathbf{H}_{ji}$ and $\mathbf{H}_{jj}$ are the $N\times(2N\text{-}1)$ upper-left sub-matrices of $\widetilde{\mathbf{H}}_{ji}$ and $\widetilde{\mathbf{H}}_{jj}$, respectively, and $\widetilde{\mathbf{h}}_{ii}$ and $\widetilde{\mathbf{h}}_{ij}$ the $(2N\text{-}1)\times1$ impulse response vectors.

Both (17) and (18) state that the contributions of the ith and jth sources to the input-output cross correlations are functions of source powers p_i and p_j. Based on (17)-(22), numerical analysis on the condition of $\mathbf{R}_{\mathbf{y}_j\mathbf{v}_j}$ and the error of $\mathbf{g}_{ij}^{est}$ can be performed. As filter length N grows, the condition worsens. To alleviate the negative effect on filter estimates, a robust regularized solution of (16) is used,

$$\mathbf{g}_{ij}^{est} = \left(\mathbf{R}_{\mathbf{y}_j\mathbf{v}_j}^T\right)^{-1} \mathbf{r}_{y_i\mathbf{v}_j} = \boldsymbol{\Phi}^T \boldsymbol{\Theta} \boldsymbol{\Psi} \mathbf{r}_{y_i\mathbf{v}_j}, \quad (23)$$

where $\mathbf{R}_{\mathbf{y}_j\mathbf{v}_j}^T = \boldsymbol{\Psi}^T \boldsymbol{\Sigma} \boldsymbol{\Phi}$ is singular value decomposition (SVD) of $\mathbf{R}_{\mathbf{y}_j\mathbf{v}_j}^T$, $\boldsymbol{\Sigma}=diag(\sigma_1,\ldots,\sigma_N)$, and $\boldsymbol{\Theta}=diag(\theta_1,\ldots,\theta_N)$, with

$$\theta_n = \begin{cases} 1/\left(\sigma_n + \delta_0 \sigma_{\max}\right) & if \quad \sigma_n > \delta_1 \sigma_{\max} \\ 0 & if \quad \sigma_n \le \delta_1 \sigma_{\max} \end{cases}. \quad (24)$$

Numerical analysis are based on the impulse response data measured in a room with reverberation time T_{60}=0.3sec [11]. The error of new ADF estimate $\mathbf{g}_{12}^{est}$ based on current value $\mathbf{g}_{12}$, as function of power ratio p_2/p_1, for N=400 (25ms), is shown in Fig.2, where ADF errors are normalized as $e_{ij}^2 = \left\|\mathbf{g}_{ij}^{est} - \mathbf{g}_{ij}^o\right\|^2 \Big/ \left\|\mathbf{g}_{ij}^o\right\|^2$. The threshold in (24) is chosen to be δ_1=0.05 and δ_0=0.01. The values of current filters are set to be 0.6 times the ideal values, i.e., $\mathbf{g}_{ij} = 0.6\mathbf{g}_{ij}^o$'s.

3.2.2. Non-adaptive gain factor

The above error analysis under white-source assumption shows that the lower the power of the *jth* source, the higher the error will be for the estimate of filter $\mathbf{g}_{ij}$. Based on this *a priori* knowledge, a heuristic VSS gain factor is proposed to discount the step-size for filter $\mathbf{g}_{ij}$ when source j is weak. Since source powers are unavailable, ADF output powers are used as approximations. The non-adaptive gain factor can now be modified as a discount of (8) and be given more conservatively as

$$\mu_{ij}(t) = \sigma_{v_j}^2(t) \Big/ \left(\sigma_{y_1}^2(t) + \sigma_{y_2}^2(t)\right) \cdot \mu(t). \quad (25)$$

3.3. Implementation of VSS-ADF

The result of 3.2 can be carried over to speech signals in an approximate sense. In practice, pre-whitening [3] is performed, which makes ADF source signals closer to the white-assumption.

The update of GAS in (14) is a function of filter length N. To eliminate this effect, (14) is normalized by N^2. As in LMS, which can be improved by introducing forgetting factor into step size update [12], the same technique is also incorporated in VSS-ADF to improve performance. Therefore, GAS update implemented with forgetting factor ρ is

$$\gamma_{ij}(t+1) = \rho\gamma_{ij}(t) + \varepsilon\mu_{ij}(t)v_i(t)v_i(t+1)\mathbf{v}_j^T(t+1)\mathbf{v}_j(t+1)\mathbf{v}_j^T(t)\mathbf{y}_j(t+1)/N^2. \quad (26)$$

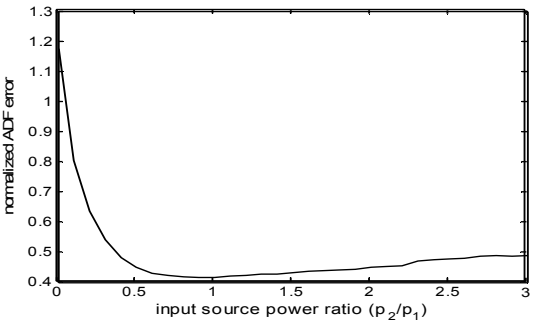

Figure 2. ADF error as a function of (p_2/p_1).

As a result, (6), (8), (9), (25), and (26) define the implementation of the VSS-ADF algorithm that integrates both adaptive and heuristic gains. Since both (26) and (25) can be updated at low cost, the increase in time complexity over baseline ADF is slight.

As widely applied in many adaptive algorithms, precautions could be taken to further prevent divergence by setting upper bounds for step-size adaptations. However, no such measures are taken for VSS-ADF described in this paper.

4. Experiments

4.1. Speech data and acoustic environment

Speech mixtures were simulated using the TIMIT speech data and the acoustic impulse responses measured in real environment [11] approximately 2 meters away from sources. Beginning and ending silences of TIMIT sentences were cut off before they were concatenated to form target $s_1(t)$ and jammer $s_2(t)$. Target speech contained 40 sentences of 4 speakers (faks0, felc0, mdab0, mreb0) and jammer speech contained sentences randomly chosen from the rest of the speakers. Speaker locations were the same as described in [10].

4.2. Convergence performance

The performance of the integrated VSS-ADF algorithm was evaluated by normalized filter errors, shown in Fig.3, and compared with block-iterative ADF, and baseline ADF with and without prewhitening. The VSS algorithm with only GAS gain γ_{ij} and with only non-adaptive gain $\gamma_c\mu_{ij}$ were also tested, where γ_c=2.4 was the average value of the GAS gain γ_{ij} shown in Fig. 4. The following conditions were used in all experiments: filter lengths N=400, α=0.005 (0.0035 for GAS-only case), prewhitening processing was the same as in [3]. For VSS-ADF, the initial GAS gain was $\gamma_{ij}(0)$=0, the gain for step-size update was $\varepsilon=4\times10^{-4}$, the forgetting factor was ρ=0.999994. The convergence speed of VSS-ADF algorithm was significantly improved over baseline ADFs and comparable to that of block-iterative ADF [3]. VSS-ADF also had lower MSE when heuristic gain μ_{ij} was applied, whether GAS gain was used or not.

4.3. TIR and phone accuracy

The initial TIR of speech mixtures were 0.53dB and -0.55dB respectively. Phone recognition experiments were conducted for both mixed and separated speech signals. Speech feature vector had 13 cepstral coefficients and their first and second-order time derivatives. Acoustic model contained 39 context-independent phone units. Each unit had 3 emission states of HMM, and each state had a size-8 Gaussian mixture density.

Phone bigram was used as "language model." Both training and test data were processed by cepstral mean subtraction. Prewhitening was used in all ADF tests. TIR and phone accuracy results listed in Table 1 show enhanced separation effects of VSS-ADF.

5. Conclusions

GAS technique is introduced into ADF for competing speech separation and a heuristic adaptation gain is obtained based on a simplified analysis of ADF errors. The proposed VSS-ADF algorithm incorporates both gradient-adaptive and heuristic gain factors effectively. Experimental results show that convergence rate was improved significantly and filter MSE was reduced. The enhancement in separation effects were demonstrated in both TIR and phone accuracy.

6. Appendix

From (2), the correlation of mixing system output is

$$\mathbf{R}_{\tilde{\mathbf{y}}\tilde{\mathbf{y}}} = \begin{bmatrix} \mathbf{R}_{\tilde{\mathbf{y}}_1\tilde{\mathbf{y}}_1} & \mathbf{R}_{\tilde{\mathbf{y}}_1\tilde{\mathbf{y}}_2} \\ \mathbf{R}_{\tilde{\mathbf{y}}_2\tilde{\mathbf{y}}_1} & \mathbf{R}_{\tilde{\mathbf{y}}_2\tilde{\mathbf{y}}_2} \end{bmatrix} = \tilde{\mathbf{H}} \cdot \mathbf{R}_{\mathbf{ss}} \cdot \tilde{\mathbf{H}}^T, \quad (27)$$

where the auto- and cross-correlations are reduced by (15) as

$$\mathbf{R}_{\tilde{\mathbf{y}}_j\tilde{\mathbf{y}}_j} = p_i \tilde{\mathbf{H}}_{ji} \tilde{\mathbf{H}}_{ji}^T + p_j \tilde{\mathbf{H}}_{jj} \tilde{\mathbf{H}}_{jj}^T, \quad (28)$$

$$\mathbf{R}_{\tilde{\mathbf{y}}_j\tilde{\mathbf{y}}_i} = p_j \tilde{\mathbf{H}}_{jj} \tilde{\mathbf{H}}_{ij}^T + p_i \tilde{\mathbf{H}}_{ji} \tilde{\mathbf{H}}_{ii}^T, \quad (29)$$

The correlation analysis of ADF based on (3) shows that

$$\mathbf{R}_{\mathbf{y}_j\mathbf{v}_j} = \mathbf{R}_{\mathbf{y}_j\mathbf{y}_j} - \mathbf{R}_{\mathbf{y}_j\tilde{\mathbf{y}}_i} \mathbf{G}_{ji}^T, \quad (30)$$

where $\mathbf{y}_i(t) = [y_i(t), \cdots, y_i(t-N+1)]^T$ is $N \times 1$ vector, and

$$\mathbf{R}_{\mathbf{y}_j\mathbf{y}_j} = \begin{bmatrix} \mathbf{I}_N & \mathbf{0}_{N\times(N-1)} \end{bmatrix} \mathbf{R}_{\tilde{\mathbf{y}}_j\tilde{\mathbf{y}}_j} \begin{bmatrix} \mathbf{I}_N & \mathbf{0}_{N\times(N-1)} \end{bmatrix}^T, \quad (31)$$

$$\mathbf{R}_{\mathbf{y}_j\tilde{\mathbf{y}}_i} = \begin{bmatrix} \mathbf{I}_N & \mathbf{0}_{N\times(N-1)} \end{bmatrix} \mathbf{R}_{\tilde{\mathbf{y}}_j\tilde{\mathbf{y}}_i}. \quad (32)$$

Substituting (28) and (29) into (31) and (32) respectively, and then both into (30), we obtain (17) together with (19) and (20).

Similarly, correlation analysis for (3) also has

$$\mathbf{r}_{y_i\mathbf{v}_j} = \mathbf{r}_{y_i\mathbf{y}_j} - \mathbf{G}_{ji} \mathbf{r}_{y_i\tilde{\mathbf{y}}_i}, \quad (33)$$

where, under the assumption of white uncorrelated sources,

$$\mathbf{r}_{y_i\mathbf{y}_j} = p_i \mathbf{H}_{ji} \tilde{\mathbf{h}}_{ii} + p_j \mathbf{H}_{jj} \tilde{\mathbf{h}}_{ij}, \quad (34)$$

$$\mathbf{r}_{y_i\tilde{\mathbf{y}}_i} = p_i \tilde{\mathbf{H}}_{ii} \begin{bmatrix} \mathbf{I}_{2N-1} \\ \mathbf{0}_{(2N-2)\times(2N-1)} \end{bmatrix} \tilde{\mathbf{h}}_{ii} + p_j \tilde{\mathbf{H}}_{ij} \begin{bmatrix} \mathbf{I}_{2N-1} \\ \mathbf{0}_{(2N-2)\times(2N-1)} \end{bmatrix} \tilde{\mathbf{h}}_{ij}. \quad (35)$$

The relations of (34) and (35) reduce (33) to (18) together with (21) and (22).

7. References

[1] E. Weinstein, M. Feder, and A. V. Oppenheim, "Multi-channel signal separation by decorrelation," *IEEE Trans. SAP*, Vol. 1, No. 4, pp. 405-413, Oct., 1993.

[2] S. V. Gerven and D. V. Compernolle, "Signal separation by symmetric adaptive decorrelation: stability, convergence, and uniqueness," *IEEE Trans. SP*, Vol. 43, No. 7, pp. 1602-1612, July 1995.

[3] Y. Zhao and R. Hu, "Fast convergence speech source separation in reverberant acoustic environment," ICASSP-04, Vol.3, pp897-900, April, 2004.

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

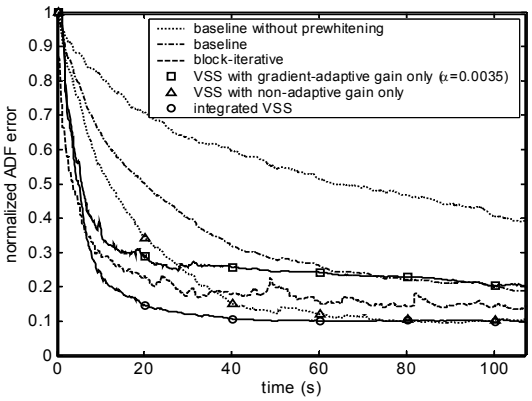

Figure 3. Normalized ADF filter errors.

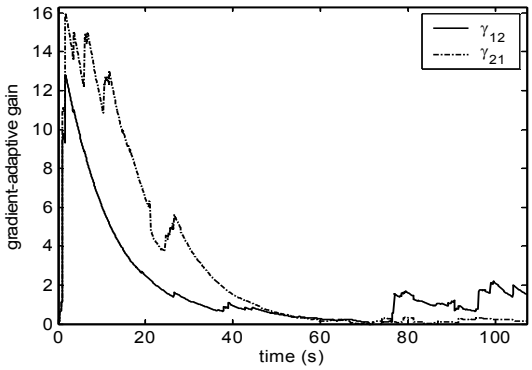

Figure 4. Gradient-adaptive gain factors.

Table 1: TIR and phone accuracy of target speech

	Mixture	Baseline	Blk-Iterative	VSS-ADF
TIR	0.53 dB	7.29 dB	8.57 dB	12.66 dB
Phone accuracy	29.1%	41.1%	43.5%	47.8%

[5] H. N. Thi and C. Jutten, "Blind source separation for convolutive mixtures," *Signal Processing*, Vol. 45, pp.209-229, 1995.

[6] V. J. Mathews and Z. Xie, "A stochastic gradient adaptive filter with gradient adaptive step size," *IEEE Trans. on SP*, Vol.41, No.6, pp.2075-2087, June 1993.

[7] S. C. Douglas, "Generalized gradient adaptive step sizes for stochastic gradient adaptive filters," *ICASSP-95*, Vol. 2, pp. 1396-1399, 9-12 May 1995.

[8] W. P. Ang and B. F. Boroujeny, "A new class of gradient adaptive step-size LMS algorithms," *IEEE Trans. on SP*, Vol. 49, No.4, Apr. 2001.

[9] S.C. Douglas and A. Cichocki, "Adaptive step size techniques for decorrelation and blind source separation," *Conference Record of the 32nd Asilomar Conf. on Signals, Sys & Comp*, Vol.2, Nov. 1998, pp. 1191 – 1195.

[10] R. Hu and Y. Zhao, "Adaptive decorrelation filtering algorithm for speech source separation in uncorrelated noises," *Proc. of ICASSP,* Vol. I, pp. 1113-1116, Philadelphia, USA, 2005.

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

[12] H. C. Woo, "Improved stochastic gradient adaptive filter with gradient adaptive step size," *Electronics Letters*, Vol.34, No.13, pp. 1300-1301, June 1998.