A Simultaneous Multi-Speaker Identification System using Blind Source Separation

Samuel Kim
Department of Electrical and Electronic Engineering, Yonsei University
worshipersam@mcsp.yonsei.ac.kr

Abstract

This paper presents a new speaker identification method using blind source separation schemes, so that a conventional speaker identification system could work several speakers simultaneously. Not only it reduces the noise, but also it separates the signal from competing speakers, which were considered as one of the most difficult situation in speaker recognition fields, and authenticates each speaker. There are improvements by 12.62dB in signal to interference ratio (SIR) point of view and by 1.39 in log likelihood ratio (LLR) point of view for overlapped signals. For the special case, there is improvement by 9.48 in LLR for non-overlapped signals.

1. Introduction

The state of the art speech and speaker recognition systems score fairly good performance in clean situation. It is still vulnerable, however, in the presence of interfering signals such as noise and competing speakers. Several techniques based on multi-microphone processing such as speech enhancement has been explored by many researchers. Normally, however, those algorithms focus on restoring the primary signal only. In this paper, we propose a new speaker recognition system using blind source separation (BSS) which enable the

conventional identification system to be capable of authenticating multiple talkers simultaneously as well as reducing noise.

Several approaches for BSS are introduced in [1]. Our approach for BBS here consists of reconstructing the input signals by assuming that they are statistically uncorrelated and imposing this constraint on the signal estimates. In a co-channel speech acquisition system, each microphone acquires not only its target signal, but also the interfering signals from the other sources. For simplicity, we set the number of channel and source to be 2 [2][3]. After the reconstructing procedure, we perform the speaker identification experiment using a conventional method of speaker recognition system [4][5][6].

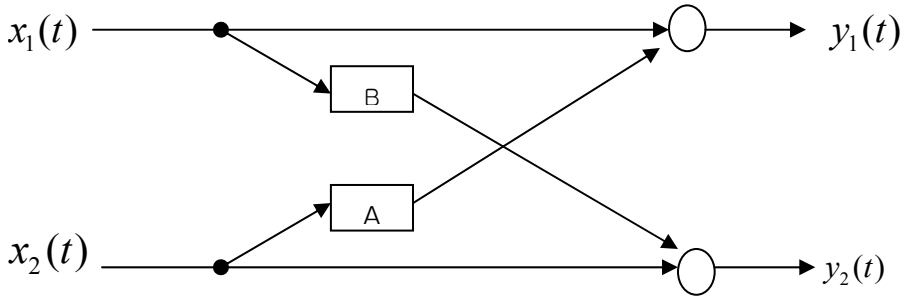


Fig. 1 Block diagram of the co-channel system

In Fig. 1, let $x_1(t)$ and $x_2(t)$ be the signals generated by sources 1 and 2, respectively, which are independent of each other. The signal acquired by the microphone that targets the source 1 and 2 are denoted by $y_1(t)$ and $y_2(t)$, respectively. A and B represent the coupling channel filters, which cause the interference in this co-channel system.

$$Y_1(\omega) = X_1(\omega) + A(\omega) X_2(\omega)$$
$$Y_2(\omega) = X_2(\omega) + B(\omega) X_1(\omega)$$

This can be interpreted as a 2 X 2 LTI system

$$H(\omega) = \begin{bmatrix} 1 & A(\omega) \\ B(\omega) & 1 \end{bmatrix}$$

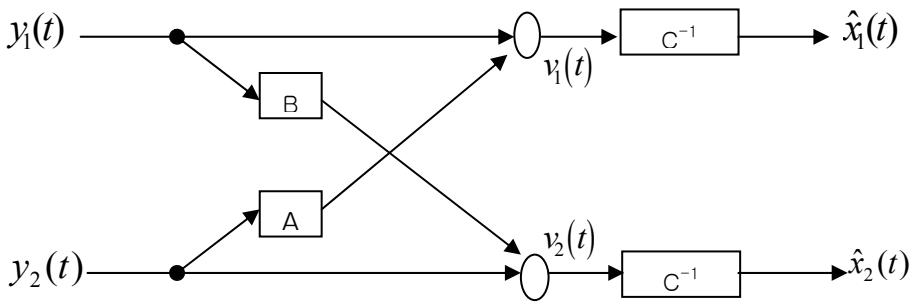


Fig. 2 Block diagram of the reconstruction system

In inverting process, assuming that we estimate $A(\omega)$, $B(\omega)$ successively to get $\hat{A}(\omega)$, $\hat{B}(\omega)$, inverting process is easy to implement by matrix inversion.

$$H(\omega)^{-1} = \frac{1}{1-\hat{A}(\omega)\hat{B}(\omega)}\begin{bmatrix} 1 & -\hat{A}(\omega) \\ -\hat{B}(\omega) & 1 \end{bmatrix}$$

This method can be shown in the Fig. 2, where

$$C(\omega)^{-1} = \frac{1}{1-\hat{A}(\omega)\hat{B}(\omega)}$$

The pre-requisition of this system is that

$$1-\hat{A}(\omega)\hat{B}(\omega) \neq 0 \quad ,\forall\omega$$

The estimation of $\hat{A}(\omega)$ and $\hat{B}(\omega)$ to the real value is the main issue of the algorithm development in the next section.

We will discuss the algorithm development in section 2, implement issues in section 3, and simulation results in section 4.

2. Algorithm development

Based on the discussion in previous section, the relationship of the power spectra of the reconstruction system will be

$$\begin{bmatrix} P_{\hat{x}_1\hat{x}_1}(\omega) & P_{\hat{x}_1\hat{x}_2}(\omega) \\ P_{\hat{x}_2\hat{x}_1}(\omega) & P_{\hat{x}_2\hat{x}_2}(\omega) \end{bmatrix} = \frac{1}{\left|1-\hat{A}(\omega)\hat{B}(\omega)\right|^2} \begin{bmatrix} 1 & -\hat{A}(\omega) \\ -\hat{B}(\omega) & 1 \end{bmatrix} \begin{bmatrix} P_{y_1y_1}(\omega) & P_{y_1y_2}(\omega) \\ P_{y_2y_1}(\omega) & P_{y_2y_2}(\omega) \end{bmatrix} \begin{bmatrix} 1 & -\hat{B}^*(\omega) \\ -\hat{A}^*(\omega) & 1 \end{bmatrix}$$

Assuming that $\hat{x}_1$ and $\hat{x}_2$ are independent, i.e., $P_{\hat{x}_1\hat{x}_2}(\omega) = 0 \ , \forall \omega$

$$P_{y_1y_2}(\omega) - \hat{A}(\omega) P_{y_2y_2}(\omega) - \hat{B}^*(\omega) P_{y_1y_1}(\omega) + \hat{A}(\omega)\hat{B}^*(\omega) P_{y_2y_1}(\omega) = 0$$

Since we can measure the $P_{y_iy_j}(\omega)$ in practice from the observation, we can derive

$$\hat{A}(\omega) = \frac{P_{y_1y_2}(\omega) - \hat{B}^*(\omega) P_{y_1y_1}(\omega)}{P_{y_2y_2}(\omega) - \hat{B}^*(\omega) P_{y_2y_1}(\omega)}$$

or

$$\hat{B}(\omega) = \frac{P_{y_2y_1}(\omega) - \hat{A}^*(\omega) P_{y_2y_2}(\omega)}{P_{y_1y_1}(\omega) - \hat{A}^*(\omega) P_{y_1y_2}(\omega)}$$

From the relationship of the power spectra of the co-channel system,

$$\begin{bmatrix} P_{y_1y_1}(\omega) & P_{y_1y_2}(\omega) \\ P_{y_2y_1}(\omega) & P_{y_2y_2}(\omega) \end{bmatrix} = \begin{bmatrix} 1 & A(\omega) \\ B(\omega) & 1 \end{bmatrix} \begin{bmatrix} P_{x_1x_1}(\omega) & 0 \\ 0 & P_{x_2x_2}(\omega) \end{bmatrix} \begin{bmatrix} 1 & B^*(\omega) \\ A^*(\omega) & 1 \end{bmatrix}$$

Substituting this into above equation of the reconstruction,

$$P_{x_1x_1}(\omega)\left[1-\hat{A}(\omega)B(\omega)\right]\left[B(\omega)-\hat{B}(\omega)\right]^* + P_{x_2x_2}(\omega)\left[1-\hat{B}(\omega)A(\omega)\right]\left[A(\omega)-\hat{A}(\omega)\right]^* = 0$$

This implicates that if one of the coupling filters is known, then the other can be found easily. There are practical situations in which one of the coupling systems is know a priori or can be measured independently. For the topic of the speech separation in this paper, it can be done by identifying a quiet period for one of the speakers, the acoustic transfer function with respect to the other speaker can be estimated separately and then used to identify the unknown transfer function when both speaker are active.

Assuming the coupling filters are FIR of the form

$$A(\omega) = \sum_{k=0}^{q_1} a_k e^{-j\omega k}$$

$$B(\omega) = \sum_{k=0}^{q_2} b_k e^{-j\omega k}$$

where q_1 and q_2 are the pre-specified filter orders.

Then,

$$v_1(t) = y_1(t) - \sum_{k=0}^{q_1} a_k y_2(t-k)$$

$$v_2(t) = y_2(t) - \sum_{k=0}^{q_2} b_k y_1(t-k)$$

where $\hat{x}_1(t)$ and $\hat{x}_2(t)$ are generated from $v_1(t)$ and $v_2(t)$ by

$$\sum_{k=0}^{q_1+q_2} c_k \hat{x}(t-k) = v_i(t)$$

where

$$c_k = \delta_k - \sum_{l=0}^{k} a_l b_{k-l} \quad k = 0,1,\cdots,(q_1+q_2)$$

Using the power spectra relationship again,

$$\begin{bmatrix} P_{y_1v_1}(\omega) & P_{y_1v_2}(\omega) \\ P_{y_2v_1}(\omega) & P_{y_2v_2}(\omega) \end{bmatrix} = \begin{bmatrix} P_{y_1y_1}(\omega) & P_{y_1y_2}(\omega) \\ P_{y_2y_1}(\omega) & P_{y_2y_2}(\omega) \end{bmatrix} \begin{bmatrix} 1 & -\hat{B}^*(\omega) \\ -\hat{A}^*(\omega) & 1 \end{bmatrix}$$

where $P_{y_1v_1}(\omega)$ denotes the cross-spectrum between $y_i(t)$ and $v_i(t)$. Substituting this into previous equations,

$$P_{y_2v_2}(\omega)\hat{B}(\omega) = P_{y_1v_2}(\omega)$$

$$P_{y_1v_1}(\omega)\hat{A}(\omega) = P_{y_2v_1}(\omega)$$

These can be represented in time domain by inverse Fourier Transform as followings.

$$\sum_{k=0}^{q_1} a_k c_{y_1v_2}(\tau-k) = c_{y_1v_2}(\tau)$$

$$\sum_{k=0}^{q_2} b_k c_{y_2v_1}(\tau-k) = c_{y_2v_1}(\tau)$$

where $c_{y_i v_j}(\tau)$ is the cross-correlation like

$$c_{y_i v_j}(\tau) = E\left\{ y_i(t) v_j^*(t-\tau) \right\}$$

These can be expressed in matrix form

$$\mathbf{C}_{\mathbf{y_2 v_2}} \mathbf{a} = \mathbf{c}_{\mathbf{y_1 v_2}}$$

$$\mathbf{C}_{\mathbf{y_1 v_1}} \mathbf{b} = \mathbf{c}_{\mathbf{y_2 v_1}}$$

where

$$\mathbf{C}_{\mathbf{y}_i \mathbf{v}_j} = E\left\{ \mathbf{v}_j^*(t) \mathbf{y}_i^T(t) \right\}$$

$$\mathbf{a} = [a_0\, a_1\, \cdots\, a_{q_1}]^T$$

$$\mathbf{b} = [a_0\, a_1\, \cdots\, a_{q_1}]^T$$

$$\mathbf{v}_i^*(t) = [v_i^*(t)\ v_i^*(t-1)\ \cdots\ v_1^*(t-q_2)]^T$$
$$\mathbf{y}_i^*(t) = [y_i^*(t)\ \mathrm{y}_i^*(t-1)\ \cdots\ \mathrm{y}_1^*(t-q_2)]^T$$

This equations which is equivalent to the previous equations in the frequency domain can be solved by

$$\mathbf{a} = \mathbf{C}_{\mathbf{y_2 v_2}}{}^{-1} \mathbf{c}_{\mathbf{y_1 v_2}}$$

$$\mathbf{b} = \mathbf{C}_{\mathbf{y_1 v_1}}{}^{-1} \mathbf{c}_{\mathbf{y_2 v_1}}$$

By alternating between these two equations, we obtain an iterative algorithm for adjusting both filter coefficients. We can easily notice that this is nothing but the extended version of the LMS or RMS algorithm. Since the correlation functions are unknown, they are approximately estimated by their samples;

$$\mathbf{C}_{\mathbf{y_2 v_2}} \approx \sum_{t=1}^{N} \beta_1^{N-t} \mathbf{v}_2^*(t) \mathbf{y}_2^T(t)$$

$$\mathbf{c}_{\mathbf{y}_1\mathbf{v}_2} \approx \sum_{t=1}^{N} \beta_1^{N-t} \mathbf{v}_2^{*}(t) y_1(t)$$

$$\mathbf{C}_{\mathbf{y}_1\mathbf{v}_1} \approx \sum_{t=1}^{N} \beta_2^{N-t} \mathbf{v}_1^{*}(t) \mathbf{y}_1^{T}(t)$$

$$\mathbf{c}_{\mathbf{y}_2\mathbf{v}_1} \approx \sum_{t=1}^{N} \beta_2^{N-t} \mathbf{v}_1^{*}(t) y_2(t)$$

where β_1 and β_2 are real number between 0 and 1.

Therefore, following adaptive algorithm can be derived.

$$\mathbf{a}(t) = \mathbf{a}(t-1) + \mu_1(t) \mathbf{v}_2^{*}(t) v_1(t)$$

$$\mathbf{b}(t) = \mathbf{b}(t-1) + \mu_2(t) \mathbf{v}_1^{*}(t) v_2(t)$$

where $\mu_1(t)$ and $\mu_2(t)$ are the step size can decide the type of the adaptive filters [2]. We will discuss how to decide the step size $\mu_1(t)$ and $\mu_2(t)$ in the next section.

3. Implement Issues

3.1. Blockwise Implementation

In the blockwise implementation, the co-channel speech signals are acquired simultaneously by two sequences of frames, where the frames are synchronized between the two channels. Each frame has *N* samples, and the shift between successive frames is *M* samples. To make sure that an adaptive filtering system is stable, I initialize the coefficients in each frame with the last coefficients of previous frame. In this paper, I let *N* and *M* be 200 samples, which means there is no overlapped analysis. If $N > M$, there are overlaps between successive frame which allow better restoration at the cost of more computation [3].

3.2. Adaptation step size

It can be shown that the following bound can be used for γ to maintain stability [3], that is

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

where $\operatorname{var}\{y_1(t)\}$ and $\operatorname{var}\{y_2(t)\}$ are the variance of $y_1(t)$ and $y_2(t)$, respectively and they can be computed in each frame. In this paper, the adaptation step size is set to be

$$\mu(t) = \frac{\alpha\Gamma}{t}, \quad 0<\alpha<1$$

where α is a positive constant chosen according to the expected time variation rate of the acoustic environment.

3.3. Performance evaluation

In this paper, I evaluate 3 different types of performance. Firstly, to evaluate the performance of adaptive filter, the squared estimation error of filter coefficients, $E(t)$, is defined as

$$E(t) = [\Delta\mathbf{a}^{(t)}]^T \Delta\mathbf{a}^{(t)} + [\Delta\mathbf{b}^{(t)}]^T \Delta\mathbf{b}^{(t)}$$
$$\Delta\mathbf{a}^{(t)} = \mathbf{a}^{(t)} - \mathbf{a}^*$$
$$\Delta\mathbf{b}^{(t)} = \mathbf{b}^{(t)} - \mathbf{b}^*$$

where * denotes the true filter coefficients.

Secondly, signal to interference ratio (SIR) is computed with following criterion.

$$SIR = 10\log_{10}\left(\frac{E\{s^2(t)\}}{E\{i^2(t)\}}\right)$$

where $s(t)$ and $i(t)$ represent desired signal and interference signal respectively.

Lastly, speaker identification experiments are performed. The identification experiments use the Mel Frequency Cepstral Coefficient (MFCC) for speaker dependent feature vectors and Gaussian Mixture Model (GMM) for speaker models [5] [6]. Speaker can be decided using this criterion

$$\text{Speaker} = \arg\max_{i} \log_e\left(f\left(\Theta_i \mid \mathbf{X}\right)\right)$$

where Θ_i and $\mathbf{X}$ represent the speaker model of i th speaker and feature vectors respectively, and also

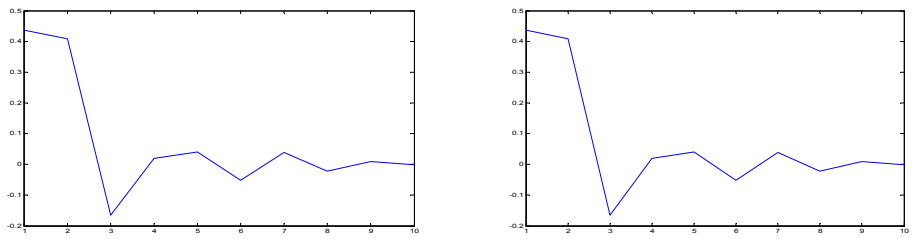
$$f_{X|\Theta}(\mathbf{X} \mid \Theta) = \sum_{i=1}^{M} \rho_i f_{X|\Theta}(\mathbf{X} \mid \Theta)$$

$$f_{\mathbf{X}|\theta_i}(\mathbf{X} \mid \theta_i) = f_{\mathbf{X}|\mu_i,\mathbf{C}_i}(\mathbf{X} \mid \mu_i, \mathbf{C}_i) = \frac{1}{(2\pi)^{d/2} \mid \mathbf{C}_i \mid^{1/2}} e^{-\frac{1}{2}(\mathbf{X}-\mu_i)^T \mathbf{C}_i^{-1}(\mathbf{X}-\mu_i)}$$

where $\mathbf{C}$ is the covariance matrix of $\mathbf{X}$ which is assumed a diagonal matrix in this paper [6].

4. Simulations

The basic scenario for this paper is that two microphones pick up the sounds from two independent sources assuming the delay difference between two acquisition microphones are negligibly small so that we can model those impulse response with no delay. Fig. 3 shows the impulse response of the cross-interference. Two types of database are used for this simulation. One is for Soongsil database and the other is TIMIT database. Both are recorded with 16kHz sampling rate and 16bit per sample. While Soongsil database contains the same sentence, saying "Open, sesame", of two different Korean speakers, TIMIT contains sets of free sentences in English for each speaker. The former is for speech separation and speaker recognition, the later is for speech separation only.



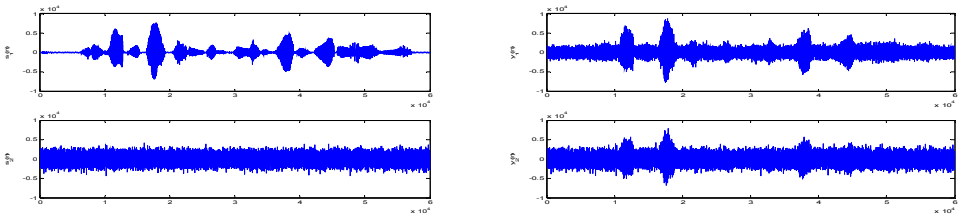
(a) (b)

Fig. 3 (a) (b) impulse response of co-channel interference A and B respectively

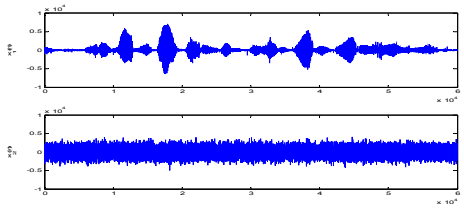
4.1. Noise Reduction

Before starting the real speech separation, here I perform the noise reduction simulation using the same environments except we only need a primary signal without noise. Fig. 4 (a) depicts that the input signals and Fig. 4 (b) and Fig. 4 (c) represent mixed and restored signals respectively.

Table 1 shows the SIR of the noise reduction processing for the reconstructed signal. The best performance, 10.98dB improvement, is accomplished when $\alpha = 1$. Fig. 5 shows that the squared estimation error of filter coefficients, $E(t)$ with different value of α. It is easily seen that larger α is, faster the convergence is accomplished.



(a) original signals (b) interfered signals



(c) recovered signals

Fig. 4 Signals in noise reduction experiments

Table 1 SIR of the noise reduction processing depends on the value of α

[dB]	Recovered Signal	Mixed Signal
$\alpha = 0.1$	10.98	2.25
$\alpha = 0.5$	12.48	
$\alpha = 1$	13.23	

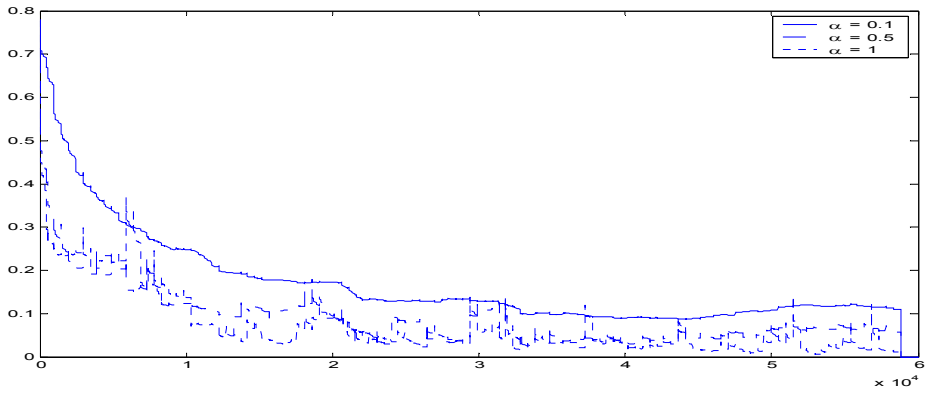
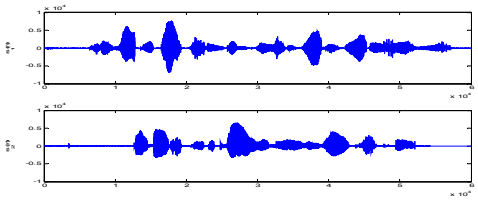


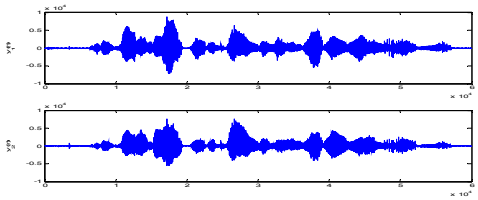
Fig. 5 $E(t)$ in noise reduction experiments

4.2. Speech Separation

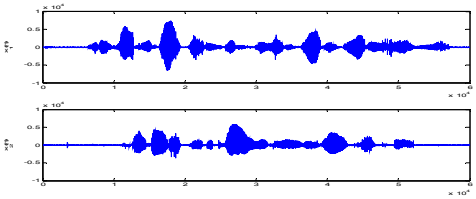
Retaining the environments, I changed the input signal with independent speech signal instead of noise. Now, we need to separate these two signals. Fig. 6 (a) depicts that the input signals and Fig. 6 (b) and Fig. 6 (c) represent mixed and restored signals respectively.



(a) original signals



(b) interfered signals

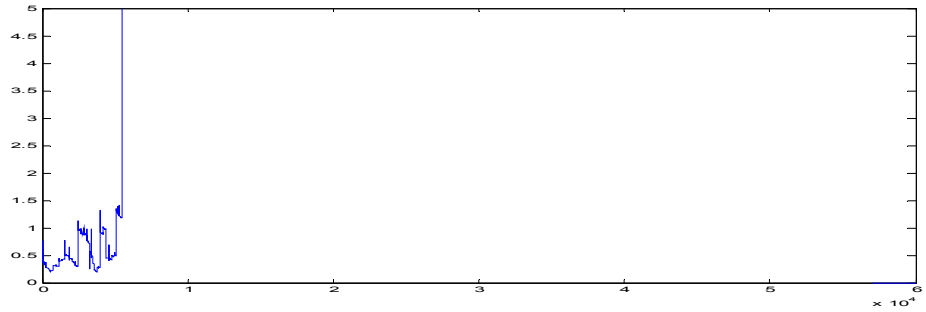


(c) recovered signals

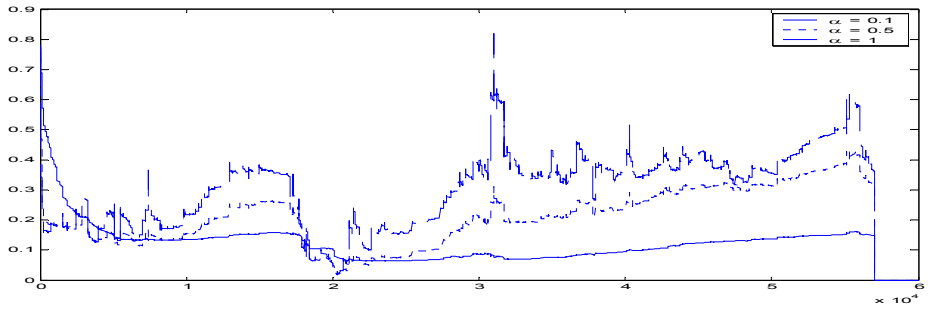
Fig. 6 Signals in speech separation experiments

Fig. 7 (a) illustrates the example of the squared estimation error of filter coefficients fall unstable when $\alpha = 5$. Fig. 7 (b) shows that the squared estimation error of filter coefficients, $E(t)$, with different values of α. To emphasis the speed of convergence, I rescale to the Fig.

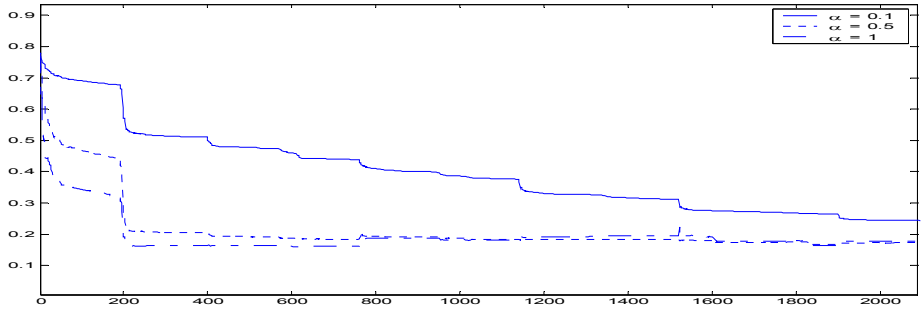
7 (c). We can easily observe that the speed of convergence is faster with larger value of α, while the average error is larger in the steady state.



(a) $\alpha = 5$



(b) $\alpha = 0.1, 0.5, 1.0$



(c) Rescaled version of (b)

Fig. 7 $E(t)$ in speech separation experiments

Table 2 shows that SIR of recovered signals with different value of α. The values of α which enables the highest SIR in recovered signal are 0.1 and 0.5 by 12.60dB and 12.64dB for

channel 1 and 2 respectively, compared it was 1 in noise reduction experiments. It confirms that the value of α is different from the each task.

Table 2. Table 1 SIR of the speech separation depends on the value of α

[dB]	α	Recovered Signal	Mixed Signal
Channel 1	$\alpha = 0.1$	17.04	4.44
	$\alpha = 0.5$	15.93	
	$\alpha = 1$	14.60	
Channel 2	$\alpha = 0.1$	13.53	1.54
	$\alpha = 0.5$	14.18	
	$\alpha = 1$	13.13	

4.3. Speaker Identification

From the above experimental output, we perform the speaker identification experiments. As we discussed in section 2, 12 orders of MFCC and GMM with 10 mixtures are used for speaker identification experiments and database for this task is Soongsil database [4]. Each speaker model is made up with average 20 seconds speech signal. Speaker identification experiments are performed 6 times for original, co-channeled, and reconstructed signals for both speakers. To evaluate the performance we define log likelihood ratio (LLR) as

$$LLR_{TF} \triangleq \frac{L_T - L_F}{|L_T|} \times 100$$

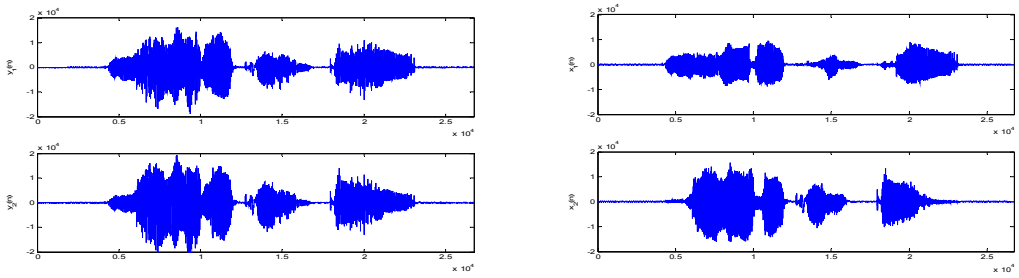
where L_T and L_F represent the value of log likelihood of true speaker and false speaker respectively. If *LLR* is negative, the decision is wrong. Table 3 and Table 4 represent the value of log likelihood against each speaker model and also Fig. 8 and Fig. 9 show the speech signals for overlapped case and non-overlapped case respectively. Table 3 indicates that there is a high possibility of speaker identification with recovered signals for each speaker even in the overlapped signal case. As a special case of non-overlapped signals, represented in Table 4, we can get LLR as high as original signals with endpoint detecting (EDP) using energy criterion [5].

Table 3 Log likelihood against each speaker model of A and B for overlapped speech

Log Likelihood	Speaker A Model	Speaker B Model	LLR_{TF}
Original A	-25.97	-29.47	13.48
Original B	-31.37	-27.63	13.54
Co-channeled A	-29.62	-29.61	- 0.03
Co-channeled B	-29.89	-29.49	1.36
Reconstructed A	-31.30	-31.43	0.42
Reconstructed B	-30.50	-29.96	1.80

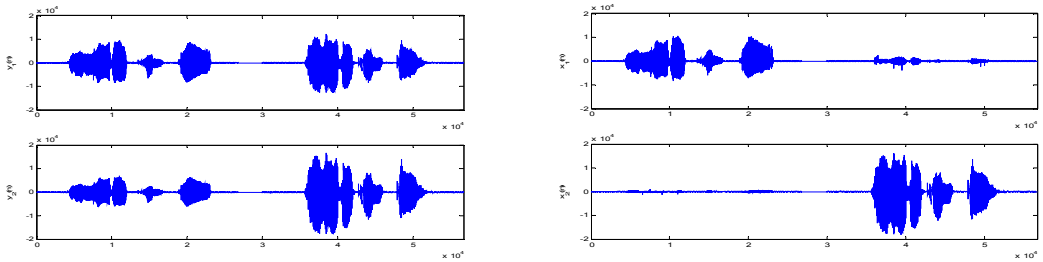
Table 4 Log likelihood against each speaker model of A and B for non-overlapped speech

Log Likelihood	Speaker A Model	Speaker B Model	LLR_{TF}
Co-channeled A	-29.62	-29.61	-0.03
Co-channeled B	-29.90	-29.49	1.39
Reconstructed A	-31.30	-31.43	0.41
Reconstructed B	-30.50	-29.96	1.80
Reconstructed A (EPD)	-26.47	-28.98	9.48
Reconstructed B (EPD)	-30.57	-27.57	10.88



(a) Co-channeled signal (b) Separated signal

Fig. 8 Overlapped signal separation process



(a) Co-channeled signal (b) Separated signal

Fig. 9 Non-overlapped signal separation process

There are improvement in LLR average 1.39 for overlapping case and 9.48 for non-overlapped case with EPD.

5. Conclusion

In this paper, we develop a simultaneous multi-speaker identification system. The rationale behind this scheme is that we can separate each signal using a conventional BSS algorithm and put each separated signals into conventional speaker identification systems. To evaluate the performance, we compute the squared estimation error of filter coefficients, SIR, and LLR. This confirms that there is a high possibility to build a comparable multi-speaker identification system to single speaker identification system. However, since feature vectors are extracted from spectral information while my speech separation algorithm is based on time domain, we need to work on the frequency domain to get higher improvements for speaker identification.

References

[1] J. Benesty and Y. Huang, "*Adaptive signal processing*", Springer, 1993.

[2] E. Weinstein, M. Feder, and A.V. Oppenheim, "Multi-channel signal separation by decorrelation", *IEEE Trans. Speech Audio Processing*, vol.1, pp. 405-143, Oct. 1993.

[3] Kuan-Chieh Yen and Yunxin Zhao, "Adaptive co-channel speech separation and recognition", *IEEE Trans. Speech Audio Processing*, vol.7, pp. 138-151, Oct. 1999

[4] Samuel Kim, Thomas Eriksson, Hong-Goo Kang, and Dae Hee Youn, "A pitch synchronous feature extraction method for speaker recognition", *Proc. Internat. Conf. Acoust. Speech Signal, Submitted, 2004*.

[5] L. R. Rabiner, B.H. Juang, "*Fundamentals of speech recognition*", Prentice Hall, 1993.

[6] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture models", *Speech Communication*, vol. 17, pp. 91-108, 1995.