LIMITATION OF FREQUENCY DOMAIN BLIND SOURCE SEPARATION FOR CONVOLUTIVE MIXTURE OF SPEECH

Shoko Araki [†] *Shoji Makino* [†] *Tsuyoki Nishikawa* [‡] *Hiroshi Saruwatari* [‡]

[†] NTT Communication Science Laboratories
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan
Email: shoko@cslab.kecl.ntt.co.jp
[‡] Nara Institute of Science and Technology
8916-5 Takayama-cho, Ikoma, 630-0101, Japan

ABSTRACT

Despite several recent proposals to achieve Blind Source Separation (BSS) for realistic acoustic signal, separation performance is still not enough. In particular, when the length of impulse response is long, performance is highly limited. In this paper, we show it is useless to be constrained by the condition, $P \ll T$, where T is the frame size of FFT and P is the length of room impulse response. From our experiments, a frame size of 256 or 512 (32 or 64 ms at a sampling frequency of 8 kHz) is best even for the long room reverberation of $T_R = 150$ and 300 ms. We also clarified the reason for poor performance of BSS in long reverberant environment, finding that separation is achieved chiefly for the sound from the direction of jammer because BSS cannot calculate the inverse of the room transfer function both for the target and jammer signals.

1. INTRODUCTION

Blind Source Separation (BSS) is an approach to estimate original source signals $s_i(t)$ using only the information of the mixed signals $x_j(t)$ observed in each input channel. This technique is applicable to the realization of noise robust speech recognition and high-quality hands-free telecommunication systems. It may also become a cue for auditory scene analysis.

To achieve BSS of convolutive mixtures, several methods have been proposed [1, 2]. Some approaches consider the impulse responses of a room h_{ji} as FIR filters, and estimate those filters [3, 4]; other approaches transform the problem into the frequency domain to solve an instantaneous BSS problem for every frequency simultaneously [5, 6].

In this paper, we consider the BSS of convolutive mixtures of speech in the frequency domain, for the sake of mathematical simplicity and reduction of computational complexity. First, we discuss the frame size of FFT used in the frequency domain BSS. It is commonly believed that the frame size T must be $P \ll T$ to estimate the unmixing matrix for the P-point room impulse response [7, 8]. We point out this is not the case for BSS, and show that smaller frame size is much better, even for long room reverberation. Next, we discuss the limitations of frequency domain BSS technique. We clarify why frequency domain BSS cannot be a good solution in a realistic acoustical environment that has a long reverberation time.

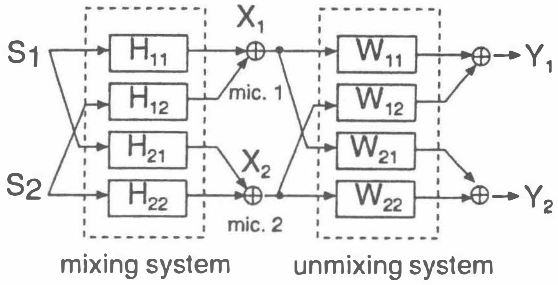

Figure 1: BSS system configuration.

2. FREQUENCY DOMAIN BSS OF CONVOLUTIVE MIXTURES OF SPEECH

The signals recorded by M microphones are given by

$$x_j(n) = \sum_{i=1}^{N}\sum_{p=1}^{P} h_{ji}(p)s_i(n-p+1) \quad (j = 1,\cdots,M), \quad (1)$$

where s_i is the source signal from a source i, x_j is the received signal by a microphone j, and h_{ji} is the P-point impulse response from source i to microphone j. In this paper, we consider a two-input, two-output convolutive BSS problem, *i.e.*, $N = M = 2$ (Fig. 1).

The frequency domain approach to the convolutive mixture is to transform the problem into an instantaneous BSS problem in the frequency domain [5, 6]. Using T-point short time Fourier transformation for (1), we obtain,

$$\boldsymbol{X}(\omega, m) = \boldsymbol{H}(\omega)\boldsymbol{S}(\omega, m). \quad (2)$$

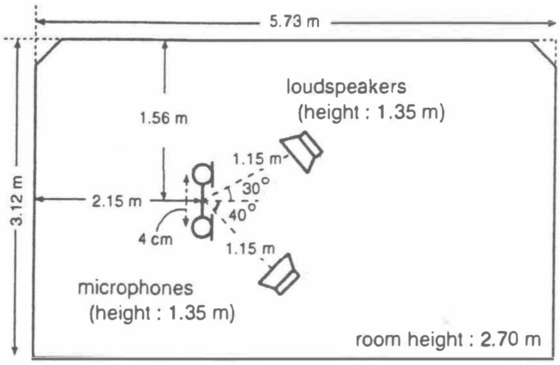

Figure 2: Layout of a room used in experiments.

We assume that the following separation has been completed in a frequency bin ω:

$$\boldsymbol{Y}(\omega,m) = \boldsymbol{W}(\omega)\boldsymbol{X}(\omega,m), \quad (3)$$

where $\boldsymbol{X}(\omega) = [X_1(\omega), X_2(\omega)]$ is the observed signal at the frequency bin ω, $\boldsymbol{Y}(\omega) = [Y_1(\omega), Y_2(\omega)]$ is the estimated source signal, and $\boldsymbol{W}(\omega)$ represents the unmixing matrix. $\boldsymbol{W}(\omega)$ is determined so that $Y_1(\omega,m)$ and $Y_2(\omega,m)$ become mutually independent. The above calculations are carried out in each frequency independently.

As for the calculation of the unmixing matrix, $\boldsymbol{W}$, we use the optimization algorithm based on the minimization of the Kullback-Leibler divergence [5, 9]. The optimal $\boldsymbol{W}$ is obtained by using the following iterative equation:

$$\boldsymbol{W}_{i+1} = \boldsymbol{W}_i + \eta\left[\mathrm{diag}\left(\langle\Phi(\boldsymbol{Y})\boldsymbol{Y}^H\rangle\right) - \langle\Phi(\boldsymbol{Y})\boldsymbol{Y}^H\rangle\right]\boldsymbol{W}_i, \quad (4)$$

where $\langle\cdot\rangle$ denotes the averaging operator, i is used to express the value of the i-th step in the iterations, and η is the step size parameter. Also, we define the nonlinear function $\Phi(\cdot)$ as

$$\Phi(\boldsymbol{Y}) = \frac{1}{1+\exp(-\boldsymbol{Y}^{(R)})} + j\,\frac{1}{1+\exp(-\boldsymbol{Y}^{(I)})}, \quad (5)$$

where $\boldsymbol{Y}^{(R)}$ and $\boldsymbol{Y}^{(I)}$ are the real and the imaginary parts of $\boldsymbol{Y}$, respectively.

3. EXPERIMENTS

3.1. Conditions for experiments

Separation experiments were conducted using the speech data convolved with the impulse responses recorded in the three environments specified by the different reverberation times: $T_R = 0$ ms, 150 ms ($P = 1200$), and 300 ms ($P = 2400$).

The layout of the room we used to measure the impulse responses is shown in Fig. 2. We used a two-element array with interelement spacing of 4 cm. The speech signals arrived from two directions, $-30°$ and $40°$. An example of the measured room impulse response used in the simulation is shown in Fig. 3. Note that the contribution of the direct sound was about 8 dB when T_R was 150 ms, and about 6 dB when T_R was 300 ms.

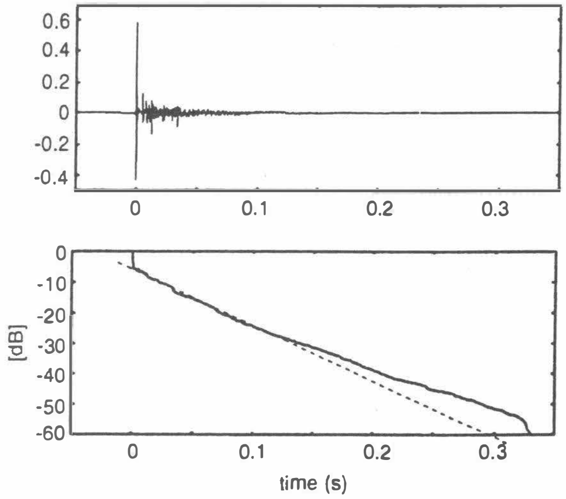

Figure 3: Measured impulse response h_{11} used in the simulation (T_R = 300 ms).

Two sentences spoken by two male and two female speakers selected from the ASJ continuous speech corpus for research were used as the original speech. The lengths of these mixed speech signals were about eight seconds each. We used the beginning of three seconds of the mixed data for learning according to equation (4), and the entire eight second data for separation.

In these experiments, we changed the frame size T from 32 to 2048 and investigated performance for each condition. The sampling rate was 8 kHz, the frame shift was half of frame size T, and the analysis window was hamming window. To solve the permutation problem, we used the blind beamforming algorithm proposed by Kurita et al [9].

3.2. Experimental results

The experimental results are shown in Fig. 4. In order to evaluate the performance for different frame size T with different reverberation time T_R, we used the *noise reduction rate* (NRR), defined as the output signal-to-noise ratio (SNR) in dB minus the input SNR in dB.

$$\begin{aligned}\mathrm{NRR}_i &= \mathrm{SNR}_{Oi} - \mathrm{SNR}_{Ii} \\ \mathrm{SNR}_{Oi} &= 10\log\frac{\sum_\omega |A_{ii}(\omega)S_i(\omega)|^2}{\sum_\omega |A_{ij}(\omega)S_j(\omega)|^2} \quad (6) \\ \mathrm{SNR}_{Ii} &= 10\log\frac{\sum_\omega |H_{ii}(\omega)S_i(\omega)|^2}{\sum_\omega |H_{ij}(\omega)S_j(\omega)|^2} \quad (7)\end{aligned}$$

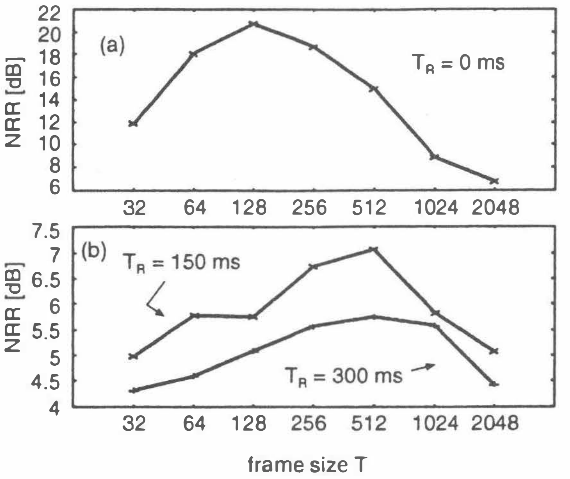

Figure 4: Result of NRR for different frame sizes.

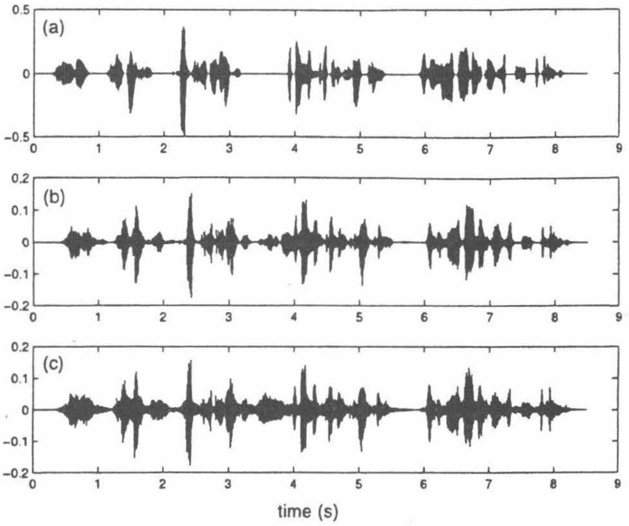

Figure 5: Separated signal for T_R = 150 ms. (a) original signal s_1. (b) separated signal y_1. T = 512. (c) separated signal y_1. T = 2048.

where $A(\omega) = W(\omega)H(\omega)$ and $i \neq j$. These values were averaged for the whole six combinations with respect to speakers, and NRR$_1$ and NRR$_2$ were averaged for the sake of convenience.

In the non-reverberant tests, the maximum NRR of 20.7 dB was obtained when T = 128 [Fig. 4(a)]. In the reverberant tests, the maximum NRR of 7.1 dB was obtained using T = 512 when the T_R was 150 ms, and the maximum NRR of 5.7 dB was obtained using T = 512 when the T_R was 300 ms [Fig. 4(b)]. The short frame functioned far better than the long frame, even for long room reverberation. Fig. 5 shows the difference of separation performance for T = 512 and T = 2048 for T_R = 150 ms. Separation was good when T = 512, but the separation was not enough and distortion occurred when T = 2048.

Even for long room reverberation, the condition $P \ll T$ is useless, and a shorter frame size T is best.

4. DISCUSSIONS

In the previous section, we showed that a longer frame size T failed. In this section, we discuss the reason why a shorter frame length T is best, and fundamental limitations of frequency domain BSS.

4.1. Optimum frame size for the frequency domain BSS

The condition $P \ll T$ has been much considered [7, 8], without success. We will discuss the reason, paying attention to the two frame sizes T, 2048 and 512.

In the frequency domain BSS framework, the signal we can use is not $x(n)$ but $X(\omega, m)$. If the frame size T is 2048 (256 ms): 1) two original signals are less independent in each frequency, thus independent assumption is hard to hold any longer; 2) frequency resolution is high, therefore the two speeches do not always exist simultaneously in the same frequency bin. The performance degrades since we cannot separate one speech or no speech using equation (4); and 3) one frame contains several consonants and vowels. Therefore, the speech is no longer stationary in the frame.

On the other hand, if the frame size is 512 (64 ms): 1) The time resolution and frequency resolution are good for speech; 2) one frame contains several fundamental periods. Therefore, the speech is stationary in the frame.

4.2. Fundamental limitation of frequency domain BSS

It is well known that an unmixing matrix $W(\omega)$ can at best be obtained up to a scale and a permutation. Before the permutation and scaling problem, however, we must note that the BSS algorithm cannot always solve the dereverberation/deconvolution problem in itself [10].

In the BSS framework, what the unmixing matrix $W(\omega)$ can do is to make $Y_1(\omega)$ and $Y_2(\omega)$ independent. W can minimize the second term of (4), and W becomes a solution of

$$\begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} = \begin{bmatrix} c_1 & 0 \\ 0 & c_2 \end{bmatrix}, \quad (8)$$

where c_1 and c_2 are arbitrary complex constants. This means that W is not always an inverse system of the mixing system H.

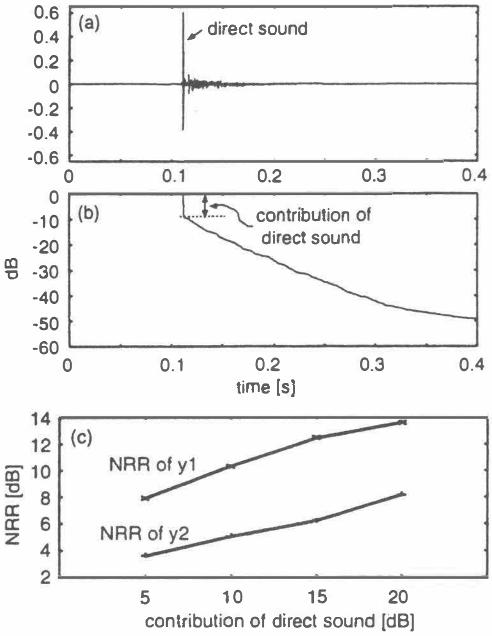

Figure 6: Relationship between the contribution of a direct sound and the separation performance. T_R = 300ms, T = 512. (a) example of an impulse response, (b) energy decay curve, (c) separation performance.

In the frequency domain approach, a delay in the impulse response is transformed in a phase shift in each frequency. If we further understand this unmixing system $\boldsymbol{W}$ in view of microphone array, we can form a directivity pattern in each frequency. The adaptation in BSS forms an adaptive null beamformer toward the jammer. Since we can control the phase shift only for the direction of the direct (biggest) sound, we can form only one null toward the jammer in the case of two microphones. As a result, separation performance is fundamentally limited by the direct to reverberant sound ratio.

Fig. 6 shows the performance when the contribution of the direct sound is changed artificially. From Fig. 6, the performance decreases with the increase of the contribution of the direct sound. This is the same characteristic as the adaptive null beamformer, *i.e.*, the inverse filter of the room impulse response is not achieved in the BSS criteria.

Incidentally, in our experiments (Fig. 4), the separation performance worsened when the frame size was 32 and 64 (4 and 8 ms). This is because the frame was too short to control the phase shift to form a null beamformer.

5. CONCLUSIONS

We have shown that it is useless to be constrained by the condition, $P \ll T$, where T is the frame size of FFT and P is the length of the room impulse responses. From our experiments, frame size of 256 or 512 is best even for long room reverberation of T_R = 150 and 300 ms. This is because, in the BSS framework, we cannot achieve dereverberation/deconvolution, *i.e.*, we cannot identify the inverse filter of the room impulse responses both for the target signal and jammer signal. Because BSS mainly considers the direct sounds, the separation performance is fundamentally limited by the direct to reverberant sound ratio.

The longer the reverberation time, the more difficult it is to achieve good separation performance. Future work will focus on finding a solution for the separation problem in a real environment.

6. ACKNOWLEDGEMENTS

We would like to thank Dr. Shigeru Katagiri and Dr. Kiyohiro Shikano for their continuous encouragement. We also thank Dr. Futoshi Asano for detailed discussions.

REFERENCES

[1] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no.6, pp. 1129–1159, 1995.

[2] S. Haykin, "Unsupervised adaptive filtering," A wiley-interscience publication, 2000.

[3] T. W. Lee, "Independent component analysis - Theory and applications," Kluwer academic publishers, 1998.

[4] M. Kawamoto, A. K. Barros, A. Mansour, K. Matsuoka, and N. Ohnishi, "Real world blind separation of convolved non-stationary signals," *Proc. ICA99*, pp. 347-352, Jan. 1999.

[5] S. Ikeda and N. Murata, "A method of ICA in time-frequency domain," *Proc. ICA99*, pp. 365-370, Jan. 1999.

[6] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, pp. 21-34, 1998.

[7] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 3, pp. 320-327, May 2000.

[8] M. Z. Ikram and D. R. Morgan, "Exploring permutation inconsistency in blind separation of speech signals in a reverberant environment," *Proc. ICASSP2000*, pp. 1041-1044, Jun. 2000.

[9] S. Kurita, H. Saruwatari, S. Kajita, K. Takeda, and F. Itakura, "Evaluation of blind signal separation method using directivity pattern under reverberant conditions," *Proc. ICASSP2000*, pp. 3140-3143, Jun. 2000.

[10] E. Weinstein, M. Feder, and A. V. Oppenheim, "Multi-channel signal separation by decorrelation," *IEEE Trans. Speech Audio Processing*, vol. 1, no. 4, pp. 405-413, Oct. 1993.