REDUCING MUSICAL NOISE BY A FINE-SHIFT OVERLAP-ADD METHOD APPLIED TO SOURCE SEPARATION USING A TIME-FREQUENCY MASK

Shoko Araki *Shoji Makino* *Hiroshi Sawada* *Ryo Mukai*

NTT Communication Science Laboratories, NTT Corporation
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan
Email: shoko@cslab.kecl.ntt.co.jp

ABSTRACT

Musical noise is a typical problem with blind source separation using a time-frequency mask. In this paper, we report that a fine-shift and overlap-add method reduces the musical noise without degrading the separation performance. The effectiveness was confirmed by results of a the listening test undertaken in a room with a reverberation time of RT_{60} =130 ms.

1. INTRODUCTION

In this paper, we consider the blind source separation (BSS) of speech signals realized by utilizing a time-frequency mask. The time-frequency mask is widely used especially to solve the underdetermined BSS problem where N source signals outnumber M sensors. We can extract/separate signals by using the time-frequency mask; however musical noise caused by the mask is a typical problem.

Several methods have been proposed for underdetermined BSS (e.g., [1–4]), and they rely on the sparseness of the source signals. In [1], the authors employ a time-frequency binary mask to extract each signal from the mixtures. The use of binary masks causes too much discontinuous zero-padding to the extracted signals, and they contain loud musical noise.

The authors of [2] also define a time-frequency binary mask by employing the ratio between an observation and the output of an adaptive beamformer, which reduces the target signal. Then they extract the target signal with this time-frequency binary mask. In a previous work [3], we used a time-frequency binary mask to remove the $N - M$ signals from the observations. We then separated the extracted mixtures by independent component analysis (ICA). Moreover, an ML estimation based ICA method has been widely studied recently (e.g.,[4–6]), where the sources are estimated after mixing matrix estimation with the l_1-norm minimization approach.

With these approaches [2–6], there is less zero-padding to the extracted signals than with the binary mask only approach [1]. However, with the $N - M$ source removal method [3] and with the l_1-norm approach [4–6], the $N-M$ components still become zero at each time-frequency point. Therefore, we still hear musical noise in their outputs.

Musical noise has been widely considered in the field of single channel speech enhancement with spectral subtraction (e.g.,[7, 8]). It is said that musical noise is heard when an output has isolated peaks and/or short ridges in its spectrogram [8]. In underdetermined BSS, we should also consider the musical noise problem because source separation with a time-frequency mask is also a sort of speech enhancement. In this paper, we report that musical noise is reduced by a fine-shift and overlap-add method. By using the fine-shift and overlap-add method, we attempt to obtain a gradual change of the spectrogram and reduce musical noise. The effectiveness was confirmed by undertaking a listening test in a room with a reverberation time of RT_{60} =130 ms.

2. UNDERDETERMINED BSS

2.1. Formulation

In real environments, N source signals s_i observed by M sensors are modeled as convolutive mixtures

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

where h_{ji} is the L-taps impulse response from a source i to a sensor j. The goal of underdetermined BSS ($N > M$) is to obtain separated signals $y_k(n)$ $(k = 1, \cdots, N)$ using only the information provided by observations $x_j(n)$. In the underdetermined scenario, sources are assumed to have a sparse distribution.

This paper employs a time-frequency domain approach because speech signals are sparser in the time-frequency domain than in the time domain [5] and convolutive mixture problems can be converted into instantaneous mixture problems at each frequency. In the time-frequency domain, observed signals are modeled as

$$X_j(f, m) = \sum_{i=1}^{N} H_{ji}(f) S_i(f, m) \ \ (j = 1, \cdots, M),$$

where $H_{ji}(f)$ is a transfer function from a source i to a sensor j, $S_i(f, m)$ and $X_j(f, m)$ denote short-time Fourier transformed sources and observed signals, respectively. f is the frequency and m is the time-dependence of the short-time Fourier transformation (STFT).

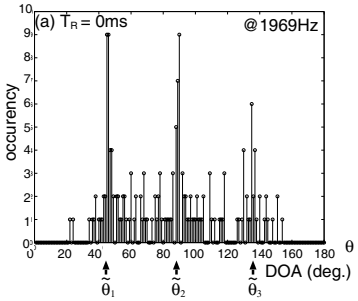

Fig. 1. Example histogram. Male-male-female combination with STFT frame size T = 512. RT_{60} = 0 ms.

2.2. Time-frequency mask estimation

Although the fine-shift and overlap-add method should be valid for various approaches [1–6], here we utilize a basic binary mask method such as that described in [1].

First, time domain signals $x_j(n)$ are converted into the time-frequency domain with a T-point STFT:

$$X_j(f,m) = \sum_{r=0}^{T-1} w(r)x_j(r+mR)e^{-\jmath 2\pi fr} \qquad (1)$$

where $f = (0, \frac{1}{T}f_s, \cdots, \frac{T-1}{T}f_s)$, f_s is a sampling frequency, $w(r)$ is a window and R is the window shift size. Here, we use a shift of $R = T/S$ where S is the shift rate.

Then a time-frequency mask is estimated using the sparseness of sources. When signals are sufficiently sparse, we can assume that sources do not overlap very often. Therefore, we can classify the observation sample points $X_j(f,m)$, and extract each signal with the time-frequency mask. To classify the observations, we use the estimated direction of arrival (DOA) $\theta(f,m) = \cos^{-1}\frac{\varphi(f,m)c}{2\pi fd}$ where $\varphi(f,m) = \angle\frac{X_i(f,m)}{X_j(f,m)}$ $(i \neq j)$ is the phase difference between two observations, d is the microphone space, and c is the speed of sound. The DOA histogram has N clusters (Fig. 1) and each cluster corresponds to one source. Using the average DOAs of these clusters $\tilde{\theta}_1, \tilde{\theta}_2, \cdots, \tilde{\theta}_N$, we define a time-frequency binary mask

$$M_k(f,m) = \begin{cases} 1 & \tilde{\theta}_k - \Delta \leq \theta(f,m) \leq \tilde{\theta}_k + \Delta \\ 0 & \text{otherwise} \end{cases} \qquad (2)$$

which extracts the estimated signal of $S_k(f,m)$. In the equation, Δ is an extraction range parameter and here we use the standard deviation σ_k of cluster k for this parameter.

2.3. Output reconstruction with overlap-add method

Next with the time-frequency mask (2), we obtain the output signal $Y_k(f,m) = M_k(f,m)X_j(f,m)(j \in \{1,\cdots,M\})$ at each time-frequency point.

Finally, the output signals in the time domain are reconstructed using the overlap-add method,

$$y_k(n) = \frac{1}{C}\sum_{l=0}^{S-1} y_k^{m+l}(n) \qquad (3)$$

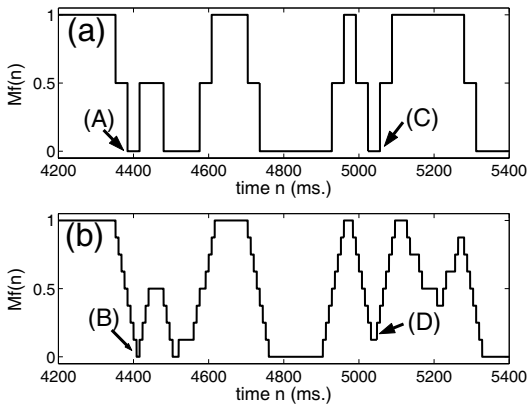

Fig. 2. Examples of overlapped time-frequency mask. (a)S = 2, (b)S = 8. T = 512.

where C is a constant that is decided by the window $w(n)$ and the shift rate S,

$$y_k^m(n) = \begin{cases} w(n-mR)y(n) = \sum_f Y(f,m)e^{\jmath 2\pi fr} \\ \qquad (mR \leq n \leq mR+T-1) \\ 0 \quad \text{(otherwise)}, \end{cases}$$

and $r = n - mR$. We should utilize an appropriate window $w(n)$ for the overlap-add. That is, the window should satisfy the condition that the sum of all the windows in $y_k^m(n)$ should be 1: $\sum_{l=0}^{S-1} w(n-(m+l)R) = 1$. Here we utilize the hanning window $w(n) = 0.5 - 0.5 * \cos(\frac{2\pi n}{T}), (n = 0, \cdots, T-1)$ and $C = 0.5S$, which fulfill the condition.

3. PROPOSAL: UTILIZING A FINE-SHIFT

3.1. Fine-Shift

In (1), a half shift $R = T/2$ (S=2) is usually used. However, we can also use a smaller shift $R = T/S$ $(S > 2)$. In this paper, we use a smaller shift, i.e., a *fine-shift* for the whole process (Secs. 2.2 and 2.3), to see how it affects musical noise.

From (3), we can see that the output signal y is obtained as the sum of S frames. When we use a fine-shift $S > 2$, the output signal is averaged over S frames and can be smoothed. Therefore, we can expect smoothed outputs, i.e., less musical noise.

3.2. Effect of fine-shift and overlap-add

At each frequency f, the time-frequency mask $M(f,m)$ extracts a signal of duration T $(mR \leq n \leq mR+T-1)$ and the extracted signals are overlapped according to (3). In order to see what happens in each frequency, we look at an overlapped time-frequency mask in this section. Figure 2 shows an example of an overlapped time-frequency mask

$$M_f(n) = \frac{1}{S}\sum_{m=\lceil\frac{n-T+1}{R}\rceil}^{\lfloor\frac{n}{R}\rfloor} M(f,m) \qquad (4)$$

at a frequency of 1078 Hz.

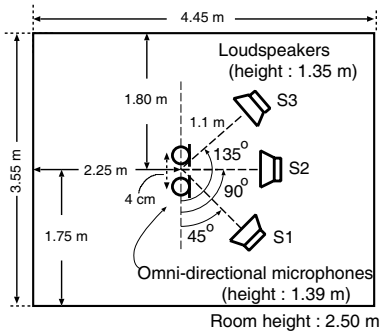

Fig. 3. Room for reverberant tests. RT_{60} = 130 ms.

One shift length is $R = T/S$ = 32 ms for $S = 2$ and 8 ms for $S = 8$, because we used $T = 512$ at a sampling frequency $f_s = 8$ kHz. When $S = 2$ we observe gaps of 32 ms [(A) and (C) in Fig. 2]. Our auditory system can recognize a gap of over around 22 ms for the sound at 500 Hz [9]. Therefore, we think that these gaps can be recognized as musical noise. However, when $S = 8$, the gap length is 8 ms [(B) in Fig. 2], and so the gap cannot be recognized. Moreover, at (D) in Fig. 2, the overlapped mask does not become zero, and its value changes gradually around (D). We believe this is why we can reduce musical noise with the fine-shift.

Moreover, the magnitude of one step is 0.125 when $S = 8$, and 0.5 when $S = 2$ (This is always true because $M(f, m) \in \{1, 0\}$). That is, we rarely have a sudden change of mask level when $S = 8$ and therefore we rarely have isolated peaks and/or short ridges in the output spectrogram. This is also one of the reasons for the musical noise being reduced by the fine-shift.

4. EXPERIMENTS

4.1. Experimental conditions

We evaluated the separation performance in anechoic tests and reverberant tests. For the anechoic tests (RT_{60} = 0 ms), we mixed speech signals using the transfer function $H_{ji}(f) = \exp(j2\pi f \tau_{ji})$, where $\tau_{ji} = \frac{d_j}{c}\cos\theta_i$, d_j is the position of the j-th microphone, c is the speed of sound, and θ_i is the direction of the i-th source. The source directions were $45°$, $90°$ and $135°$. For the reverberant tests, the speech data were convolved with impulse responses recorded in the room illustrated in Fig. 3 whose reverberation time was RT_{60} = 130 ms. As the original speech, we used Japanese sentences of around 7 seconds in length spoken by male and female speakers. We investigated three combinations of speakers and averaged the results.

The STFT frame size T was 512 at a sampling rate of 8 kHz. We changed the frame shift from $64(= T/8)$ to $256(= T/2)$.

4.2. Performance measures

We used the signal to interference ratio (SIR) and the signal to distortion ratio (SDR) as measures of separation performance and sound quality, respectively:

$$\mathrm{SIR}_i = 10\log\frac{\sum_n y_{is_i}^2(n)}{\sum_n(\sum_{i\neq j} y_{is_j}(n))^2},$$

$$\mathrm{SDR}_i = 10\log\frac{\sum_n x_{ks_i}^2(n)}{\sum_n(x_{ks_i}(n)-\alpha y_{is_i}(n-D))^2},$$

where y_i is an estimation of s_i, and y_{is_j} is the output of the whole separating system at y_i when only s_j is active, and $x_{ks_i} = h_{ki} * s_i$ ($*$ is a convolution operator). α and D are parameters used to compensate for the amplitude and phase difference between x_{ks_i} and y_{is_i}.

Table 1. Results for RT_{60} =0 ms. S: shift rate

S	SIR	SDR	MOS
2	17.0	7.9	1.9
4	18.3	8.7	2.8
8	18.7	9.0	3.3

Table 2. Results for RT_{60} =130 ms. S: shift rate

S	SIR	SDR	MOS
2	12.2	6.4	1.6
4	13.8	7.4	2.5
8	14.3	7.7	3.1

Since SDR cannot evaluate musical noise [7], we also conducted a subjective test and obtained the mean opinion score (MOS). The listening tests were undertaken with 20 listeners. Each listener was asked about the audibility of musical noise, and they awarded a score from one (clearly audible) to five (not audible) for each output signal.

The SIR and SDR values and the MOSs were averaged over three speaker combinations.

4.3. Experimental results

Tables 1 and 2 show the SIR and SDR results when we changed the shift rate S. We can see that the SDR values increase without any reduction in the SIR values. Note that the SIR values are not reduced by the fine-shift.

Moreover, Tables 1 and 2 also show the MOS. An analysis of variance confirmed that there were significant differences among the MOS for each shift rate ($F(2, 537) = 160.8, p < .0001$ when RT_{60}=0 ms and $F(2, 537) = 136.1$, $p < .0001$ when RT_{60}=130 ms). Our subjective test confirmed that the fine-shift reduced the audible musical noise.

We also confirmed that the fine-shift is effective for reducing musical noise when with Roman's method [2] and Araki's method [3]. Some sound examples can be found at [10].

5. DISCUSSIONS

We can reduce musical noise by using a fine-shift. This is because output signals are averaged over S frames. Figure 4 shows spectrograms for $S = 2$ and $S = 8$. We can see the isolated peaks and ridges when $S = 2$. By contrast, the isolated components are smoothed when $S = 8$.

Figure 5 shows examples of smoothed signals at a frequency of 1078 Hz. 'org' and 'shift2' and 'shift8' show the amplitudes of the original signal, and the outputs $|Y(f, m)|$ when $S = 2$ and $S = 8$, respectively. The plot labeled 'LPF(butt)' is the amplitude of $Y(f, m)$ filtered with a low pass filter (Butterworth filter) when $S = 2$. Although we

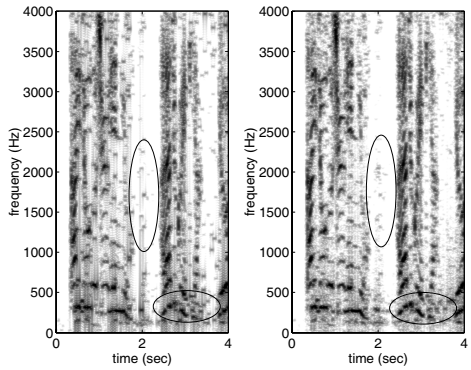

Fig. 4. Example spectrograms. Left: $S = 2$, right: $S = 8$.

expected the smoothing effect with LPF, there was an inappropriate increase in the output signal.

'smooth2' and 'smooth4' are the amplitudes of the averaged waveforms of $Y(f,m)$ when $S = 2$, that is, $Y(f,m) \leftarrow \frac{1}{K}\sum_{l=0}^{K-1} Y(f,m-l)$ and $K = 2$ for 'smooth2' and $K = 4$ for 'smooth4'. The smoothed waveforms have smaller amplitudes than the 'shift2', and the waveform has changed in 'smooth4'. This is because the speech signal is very sparse at each frequency, making this smoothing unsuitable. Moreover, with 'smooth2' and 'smooth4', the output $Y(f,m)$ is averaged at a two-times over-sampling rate of T/S ($S = 2$). This is different from the sum of (3) when $S = 4$ ('shift4') or $S = 8$ ('shift8'), that is, four-times or eight-times over-sampling.

We may also be able to use a temporally smoothed time-frequency mask $M(f,m)$, which we can obtain, for example, by shading a binary mask. However, we may pick up interference at such shaded time-frequency points. Therefore, the SIR values may decrease. By contrast, the overlap-add with the fine-shift method estimates the time-frequency mask $M(f,m)$ at each (over-sampled) time point m so that only one signal is picked up. Therefore, the fine-shift does not affect the SIR values.

Note that the fine-shift increases the calculation cost because the fine-shift is the same as over-sampling. We should select the shift rate S according to the application.

6. CONCLUSION

We evaluated the effectiveness with which a fine-shift reduces musical noise, which is a problem in the underdetermined BSS. The fine-shift and overlap-add method reduces musical noise effectively without destroying the separation performance. We have already proposed using a spatially continuous mask [11] to reduce musical noise. The fine-shift and overlap-add method realizes a temporally relatively continuous mask. Therefore, we expect that we can reduce musical noise more effectively by using a mask that is both spatially and temporally continuous.

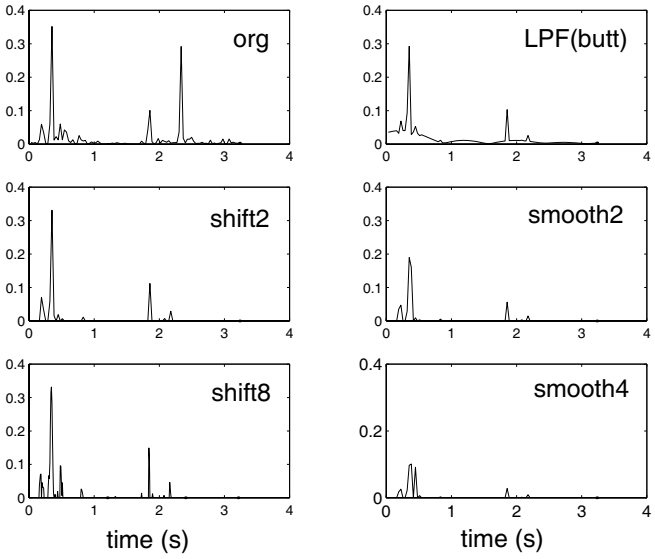

Fig. 5. Signal example at a frequency (1078Hz). 'org':original, 'shift2':$S = 2$, 'shift8':$S = 8$, 'LPF(butt)':signal filtered by Butterworth filter, 'smooth2' and 'smooth4': average for 2 or 4 sample points respectively.

7. REFERENCES

[1] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. on SP*, vol. 52, no. 7, pp. 1830-1847, 2004.

[2] N. Roman and D. Wang, "Binaural sound segregation for multisource reverberant environments," *Proc. ICASSP2004*, pp. 373-386, 2004.

[3] S. Araki, S. Makino, A. Blin, R. Mukai and H. Sawada, "Blind separation of more speech than sensors with less distortion by combining sparseness and ICA," *Proc. IWAENC2003*, pp. 271–274, 2003.

[4] F. J. Theis, C. G. Puntonet and E. W. Lang, "A histogram-based overcomplete ICA algorithm," *Proc. ICA2003*, pp. 1071-1076, 2003.

[5] P. Bofill and M. Zibulevsky, "Blind separation of more sources than mixtures using sparsity of their short-time Fourier transform," *Proc. ICA2000*, pp. 87-92, 2000.

[6] S. Winter, H. Sawada, S. Araki and S. Makino, "Overcomplete BSS for convolutive mixtures based on hierarchical clustering," ICA2004, (to appear).

[7] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Trans. on SAP*, vol. 7, no. 2, pp. 126-137, 1999.

[8] Z. Goh, K-C. Tan and B. T. G. Tan, "Postprocessing method for suppressing musical noise generated by spectral subtraction," *IEEE Trans. on SAP*, vol. 6, no. 3, pp. 287-292, 1998.

[9] B. C. J. Moore, *An introduction to the psychology of hearing*, 3rd Ed., Academic Press, 1989.

[10] http://www.kecl.ntt.co.jp/icl/signal/araki/fineshift.html

[11] S. Araki, S. Makino, H. Sawada and R. Mukai, "Underdetermined blind speech separation with directivity pattern based continuous mask and ICA," *Proc. EUSIPCO2004*, pp. 1991-1994, 2004.