Cepstral Smoothing of Separated Signals for Underdetermined Speech Separation

Yumi Ansai*†, Shoko Araki†, Shoji Makino*, Tomohiro Nakatani†, Takeshi Yamada*,
Atsushi Nakamura† and Nobuhiko Kitawaki*
*Graduate School of Systems and Information Engineering, University of Tsukuba
1-1-1 Tennoudai, Tsukuba-shi, Ibaraki 305-8573, Japan
†NTT Communication Science Laboratories, NTT Corporation
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan
Email: ansai@mmlab.cs.tsukuba.ac.jp

Abstract— **Musical noise is a typical problem with blind source separation using a time-frequency mask. Recently, the cepstral smoothing of spectral masks (CSM) was proposed. Based on the idea of smoothing in the cepstral domain, this paper proposes the cepstral smoothing of separated signals (CSS) on the assumption that a cepstral representation better reflects the characteristics of speech signals than those of masks (or filter gains). We also report a comparative evaluation study of CSM and CSS with other musical noise reduction methods. Our experimental results show that CSM is effective for musical noise reduction, but the target speech was relatively distorted. On the other hand, our proposed CSS produced less distorted target signals with the same musical noise reduction as CSM.**

I. Introduction

Blind source separation (BSS) is a source signal estimation approach that uses only the mixed signal information observed at each sensor. The BSS technique for speech signals dealt with in this paper has many applications, including hands-free teleconference systems.

In this paper, we focus on an approach that relies on the sparseness of source signals, more specifically, the time-frequency binary mask (BM) approach [1]–[4], because the time-frequency mask is widely used to solve the underdetermined BSS problem where N source signals outnumber M sensors. However, the use of binary masks makes the separated signals discontinuous and this causes audible musical noise.

Recently, to overcome this problem, a new musical noise reduction approach was proposed for speech separation [5]. In this approach, temporal cepstral smoothing of spectral masks (CSM) was applied to binary masks in the time-frequency domain. It was shown that CSM effectively reduces the musical noise for separated signals. However, [5] provides only a few example results, and does not compare the quality of the approach with that of other musical noise reduction methods. In this paper, we apply this approach to an underdetermined case and compare its quality with that of sounds obtained with the other methods.

Furthermore, inspired by CSM, we propose a new approach. To take advantage of the feature of cepstral domain representation, we propose applying cepstral smoothing to separated signals (CSS) obtained by the BM approach, because we believe that the cepstral representation appropriately reflects the characteristics of speech signals rather than those of binary masks. Another advantage of CSS is that it is applicable to any source separation and noise reduction method, including single channel noise suppression, because the smoothing is performed directly on the separated speech signal. We compare the proposed CSS approach with CSM, and verify the former's effectiveness.

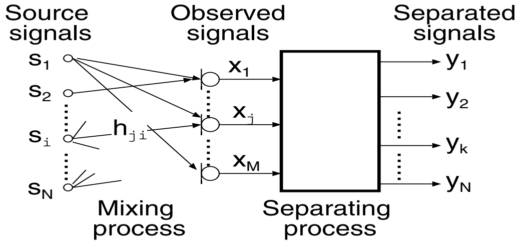


Fig. 1. Block diagram of underdetermined BSS. ($N > M$)

II. Problem description

A. Mixing process

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

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

where x_j is the signal observed by a sensor j, and h_{ji} is the L-taps impulse response from a source i to a sensor j (Fig. 1). In this paper, we consider the underdetermined case ($N > M$) and assume that N and M are known. The goal is to obtain separated signals y_i that are an estimation of s_i using only the information provided by observations x_j.

We employ a time-frequency domain approach because, in the time-frequency domain, speech signals are sparser than in the time domain [6], and convolutive mixture problems can be converted into instantaneous mixture problems at each frequency. In the time-frequency domain, observations are modeled as

$$X_j(f,m) = \sum_{i=1}^{N} H_{ji}(f)S_i(f,m) \ \ (j=1,\ldots,M), \quad (2)$$

978-1-4244-5309-2/10/$26.00 ©2010 IEEE

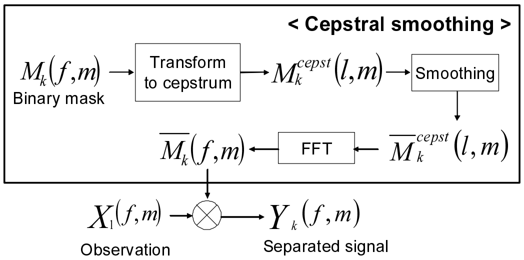
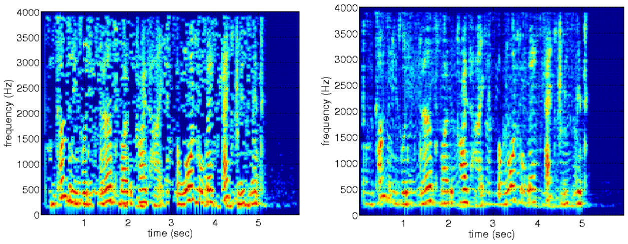


Fig. 2. Block diagram of conventional cepstral smoothing of spectral masks (CSM).



(a) BM

(b) CSM

Fig. 3. Spectrograms of separated signals.

where $H_{ji}(f)$ is a transfer function from a source i to a sensor j, $S_i(f,m)$ and $X_j(f,m)$ respectively denote short-time Fourier transformed source signals and observations. f is frequency and m is the time frame index.

B. Separating process

We employ the time-frequency BM approach [4] for separation. With this approach, we assume that signals are sufficiently sparse, and that at most one source is dominant at each time-frequency slot. If these assumptions hold, the phase differences between sensor observations have N clusters. Because an individual cluster corresponds to an individual source, we can separate each signal by collecting the observation signal at time-frequency points in each cluster. We perform observation vector $\mathbf{X}(f,m) = [X_1(f,m),\dots,X_m(f,m)]^T$ clustering with the k-means algorithm and design the mask [4],

$$M_k(f,m) = \begin{cases} 1 & \mathbf{X}(f,m) \in C_k, \\ M_{min} & \text{otherwise}, \end{cases} \tag{3}$$

where k is the number of sources, C_k denotes the cluster for source k, and M_{min} is a small value, e.g., 0. Then, we obtain separated signals by applying this binary mask to one of the observations,

$$Y_k(f,m) = M_k(f,m)X_j(f,m). \tag{4}$$

Finally, we obtain separated signals y_k by employing an inverse short-time Fourier transform (STFT) and the overlap-and-add method.

However, using binary masks makes the separated signals discontinuous and this cause audible musical noise.

III. Conventional method

In this section, we review the conventional approach for the cepstral smoothing of spectral masks (CSM) [5]. Figure 2 shows a block diagram of CSM.

The cepstral representation of the mask is obtained as

$$M_k^{cepst}(l,m) = DFT^{-1}\{\ln(M_k(f,m))|_{f=0,\dots,F-1}\}, \tag{5}$$

where l is the quefrency bin index, $DFT\{\cdot\}$ denotes the discrete Fourier transform operator, and F is the length of the transform. $M_k(f,m)$ is designed in (3) where we set $M_{min} = 0.01$. Then, temporal and recursive smoothing is applied to M_k^{cepst},

$$\overline{M}_k^{cepst}(l,m) = \beta_l \overline{M}_k^{cepst}(l,m-1) + (1-\beta_l)M_k^{cepst}(l,m). \tag{6}$$

With cepstral smoothing (6), to distinguish between the speech characteristics and unwanted random peaks, which are perceived as musical noise, the smoothing constants β_l are selected separately for the different quefrency bins l as

$$\beta_l = \begin{cases} \beta_{env} & \text{if } l \in \{0,\dots,l_{env}\}, \\ \beta_{pitch} & \text{if } l = l_{pitch}, \\ \beta_{peak} & \text{if } l \in \{(l_{env}+1),\dots,F/2\}\backslash\{l_{pitch}\}. \end{cases} \tag{7}$$

For the lower bin $l \in \{0,\dots,l_{env}\}$, because the values of $M_k^{cepst}(l,m)$ denote the spectral envelope of $M_k(f,m)$, β_{env} should have a very low value to maintain the envelope. In the same way, a relatively low value β_{pitch} is applied to the bin $l = l_{pitch}$ corresponding the pitch frequency in $M_k(f,m)$. The other bins denote the fine structure of $M_k(f,m)$ that contains the unwanted random peaks with a high probability. Therefore, strong smoothing should be performed with a high β_{peak} value ($> \beta_{pitch}$).

For the time-frame m, l_{pitch} is chosen as the cepstral bin that implements the following equation:

$$l_{pitch} = \underset{l}{\operatorname{argmax}}\{M_k^{cepst}(l,m)|l_{low} \le l \le l_{high}\}. \tag{8}$$

The range $\{l_{low}, l_{high}\}$ is determined as the possible pitch frequencies between 70 and 500 Hz.

For bins $l > F/2$, $\overline{M}_k^{cepst}(l,m)$ is determined by the symmetry assumption of the DFT. Then, a smoothed spectral mask $\overline{M}_k(f,m)$ in the time-frequency domain is obtained by using DFT and an exponential function.

$$\overline{M}_k(f,m) = \exp(\mathrm{DFT}\{\overline{M}_k^{cepst}(l,m)|_{l=0,\dots,F-1}\}). \tag{9}$$

Finally, we use this smoothed mask to obtain the separated signals according to (4).

Figure 3 shows the spectrograms of separated signals obtained with BM and CSM. With BM (Fig. 3 (a)), isolated random peaks were noticeable and a lot of musical noise is present. On the other hand, with CSM (Fig. 3 (b)), the isolated peaks were smoothed, and so CSM was confirmed to be effective in reducing the musical noise.

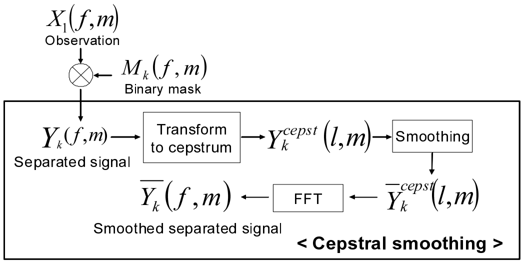


Fig. 4. Block diagram of proposed cepstral smoothing of separated signals (CSS).

TABLE I

PARAMETERS BASED ON [5]

$f_s = 8\text{kHz}$	$l_{env} = 16$	$\beta_{env} = 0$
$F = 512$	$l_{low} = 32$	$\beta_{pitch} = 0.4$
$M_{min} = 0.01$	$l_{high} = 228$	$\beta_{peak} = 0.8$

IV. PROPOSED METHOD

With the CSM, spectral masks were smoothed in the cepstral domain. However, it is not obvious whether the cepstral representation of the mask reflects the speech characteristics, e.g., the envelope and pitch information on which (7) depends. Because spectral masks are binary, i.e., 1 or 0, they are only estimation of the target speech existence. In this section, inspired by CSM, we propose applying cepstral smoothing to the separated speech signals (CSS), by assuming that the cepstral representation of the speech signal provides more reliable speech characteristics than that of the masks.

Figure 4 shows the block diagram of the proposed CSS. The difference between the proposed CSS approach and CSM is that CSS applies cepstral smoothing to separated speech signals obtained by BM.

First, a separated signal is transformed into the cepstral domain as

$$Y_k^{cepst}(l,m) = DFT^{-1}\{\ln(Y_k(f,m))|_{f=0,\ldots,F-1}\}, \quad (10)$$

where $Y_k(f,m)$ is the separated signal defined by (4). Then temporal and recursive smoothing is applied to Y_k^{cepst},

$$\overline{Y}_k^{cepst}(l,m) = \beta_l \overline{Y}_k^{cepst}(l,m-1) + (1-\beta_l)Y_k^{cepst}(l,m). \quad (11)$$

The conditions of β_l are the same as (7), but the equation for l_{pitch} has to be changed,

$$l_{pitch} = \underset{l}{\operatorname{argmax}}\{Y_k^{cepst}(l,m)|l_{low} \leq l \leq l_{high}\}. \quad (12)$$

For bins $l > F/2$, $\overline{Y}_k^{cepst}(l,m)$ is determined by the symmetry assumption of the DFT. Then, we obtain the smoothed signals $\overline{Y}_k(l,m)$ by applying DFT and an exponential function to $\overline{Y}_k^{cepst}(l,m)$. Finally, the smoothed separated signals y_k are obtained by employing an inverse STFT and the overlap-and-add method.

TABLE II

THE VALUES OF β_l

		original	case 1	case 2	case 3
CSM	β_{pitch}	0.4	0.4	0.2	0.2
	β_{peak}	0.8	0.6	0.8	0.6
CSS	β_{pitch}	–	0.4	0.4	0.4
	β_{peak}	–	0.8	0.5	0.4

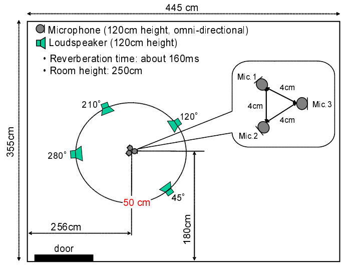

445 cm
Microphone (120cm height, omni-directional)
Loudspeaker (120cm height)
• Reverberation time: about 160ms
• Room height: 250cm
Mic.1
4cm
4cm
Mic.3
4cm
Mic.2
210°
120°
280°
45°
50 cm
355cm
256cm
180cm
door

Fig. 5. Experimental conditions

V. EXPERIMENT AND EVALUATION

A. Experiment 1 : Comparison of CSM and CSS

First, we evaluated our proposed CSS and compared its performance with that of CSM. We use the same parameters β_l as [5] for the CSM (see TABLE I). Additionally, we evaluated the CSM with different smoothing coefficients β_l. For CSS, we tuned β_l, and utilized three sets of coefficients. The variations in β_l are shown in TABLE II. It should be noted that β_{env} must be very low to maintain the envelope of the speech signal, and both β_{pitch} and β_{peak} are particularly related to smoothing strength.

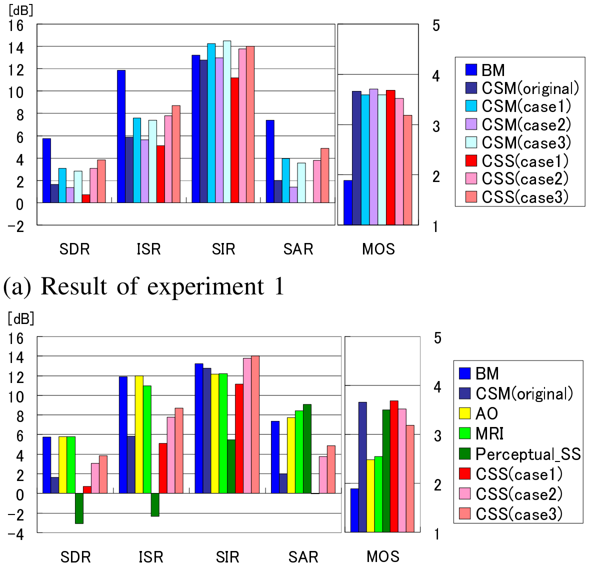
B. Experiment 2 : Comparison of CSM and CSS with other musical noise reduction methods

In [5], the performance of the approach was not compared with that of other musical noise reduction methods. Therefore in this paper, we compare the qualities of BM, CSM, and CSS with those of the following three methods.

- Adding Original Observations (AO) : A small sound from observation $X_1(f,m)$ is added to the separated signals.
- Mask Regeneration by Image processing (MRI) : Derived from image processing, the separated signal spectrum $Y_k(f,m)$ is smoothed by using adjacent time-frequency slots as follows:

$$\begin{aligned}\overline{Y}_k(f,m) = {} & 0.5Y_k(f,m) \\ & + 0.125\{Y_k(f-1,m) + Y_k(f+1,m) \quad (13) \\ & + Y_k(f,m-1) + Y_k(f,m+1)\}.\end{aligned}$$

- Perceptual_SS [7] : Modifies the masks based on human perceptual characteristics using the critical band.



(a) Result of experiment 1

(b) Result of experiment 2

Fig. 6. Evaluation results

C. Experimental conditions

The speech data were convolved with impulse responses recorded in the experimental room (Fig. 5). The reverberation time was about 160 ms.

We performed objective and subjective evaluations. The objective measures were the four kinds of distortion measure proposed in [8].

- Signal to Distortion Ratio (SDR): total distortion of following three
- Source Image to Spatial distortion Ratio (ISR): linear distortion
- Source to Interference Ratio (SIR): distortion from interference signals (non-target speaker voices)
- Sources to Artifacts Ratio (SAR): non-linear distortion

The Mean Opinion Score (MOS) was used as the subjective measure. The MOS score was obtained by employing listening tests. These listening tests focused solely on the amount of musical noise.

D. Results

Figure 6 (a) shows the results comparing CSM and CSS. For comparison, we also evaluated the performance of BM without cepstral smoothing. Although the SIR values with CSM and CSS are almost the same as that of BM, ISR and SAR of CSM and CSS degrade than BM. This distortion issue will be discussed in the next section. When we compare CSM and CSS, we can see that CSS provides us better performance as regards ISR and SAR if we choose β_l appropriately. This means that CSS holds the speech characteristics more accurately than CSM. We can also see that the MOS score of CSS is the same as that of CSM.

Figure 6 (b) shows the results of the cepstral smoothing approaches (CSM and CSS) and other methods mentioned in Section V-B. The ISR and SAR scores for CSM and CSS were lower than those obtained with other methods. On the other hand, the MOS scores of CSM and CSS, which focused on musical noise, were higher, which means the musical noise was reduced more effectively with CSM and CSS than with the other methods.

E. Discussion

As mentioned in the previous subsection, cepstral smoothing approaches (CSM and CSS) effectively reduce the musical noise. However, the ISR and SAR results showed that the cepstral smoothing approaches cause different kinds of distortion from musical noise. We observed that the cepstral smoothing sometimes removes not only the musical noise components but also the target speech signal, and causes some distortion. We also had a feeling of increased distortion like reverberation after the cepstral smoothing. Such kinds of distortion were evaluated by ISR and SAR.

In the experiments, it was also shown that CSS provides better ISR and SAR than CSM, and an equivalent MOS. When we listen to the separated signal with CSS, the target removal and reverberant distortions were smaller than that of CSM. This means that cepstral smoothing of the separated speech spectrum can protect the speech characteristics more effectively than CSM.

Moreover, as we can see from Figs. 2 and 4, the calculation time for CSS is the same as for CSM.

VI. Conclusion

For underdetermined BSS, we proposed smoothing the separated signals in the cepstral domain and evaluated its quality. Compared with CSM, CSS obtained less distorted signals, and also made it possible to reduce musical noise. Thus CSS is shown effective for musical noise reduction.

Acknowledgment

We thank Dr. Keisuke Kinoshita for detailed discussions.

References

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

[2] N. Roman and D. Wang, "Binaural sound segregation for multisource reverberant environments," *Proc. ICASSP 2004*, vol. Ⅱ, pp. 373-376, May 2004.

[3] S. Rickard and O. Yilmaz, "On the W-Disjoint orthogonality of speech," *Proc. ICASSP 2002*, vol. 1, pp. 529-532, May 2002.

[4] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Processing*, vol. 77, no. 8, pp. 1833-1847, Aug. 2007.

[5] N. Madhu, C. Breithaupt, and R. Martin, "Temporal smoothing of spectral masks in the cepstral domain for speech separation," in *Proc. ICASSP 2008*, pp. 45-48, Mar. 2008.

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

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

[8] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. P. Rosca, "First stereo audio source separation evaluation campaign: data, algorithms and results," in *Proc. ICA 2007*, pp. 552-559, Sept. 2007.