Stereo-input Speech Recognition using Sparseness-based Time-frequency Masking in a Reverberant Environment

Yosuke Izumi[1], *Kenta Nishiki*[1]*, *Shinji Watanabe*[2], *Takuya Nishimoto*[1], *Nobutaka Ono*[1], *Shigeki Sagayama*[1]

[1] Department of Information Physics and Computing, University of Tokyo 7-3-1, Hongo, Bunkyo-ku, Tokyo, Japan.
[2] NTT Communication Science Laboratories†2-4 Hikaridai Seikacho, Soraku-gun, Kyoto, Japan.
{izumi, nishiki, nishi, onono, sagayama}@hil.t.u-tokyo.ac.jp,
watanabe@cslab.kecl.ntt.co.jp

Abstract

We present noise robust automatic speech recognition (ASR) using sparseness-based underdetermined blind source separation (BSS) technique. As a representative underdetermined BSS method, we utilized time-frequency masking in this paper. Although time-frequency masking is able to separate target speech from interferences effectively, one should consider two problems. One is that masking does not work well in noisy or reverberant environment. Another is that masking itself might cause some distortion of the target speech. For the former, we apply our time-frequency masking method [7] which can separate the target signal robustly even in noisy and reverberant environment. Next, investigating the distortion caused by time-frequency masking, we reveal following facts through experiments: 1) soft mask is better than binary mask in terms of recognition performance and 2) cepstral mean normalization (CMN) reduces the distortion, especially for that caused by soft mask. At the end, we evaluate the recognition performance of our method in noisy and reverberant real environment.

Index Terms: time-frequency mask, speech sparseness, blind source separation, stereo-input, robust ASR

1. Introduction

Noise robustness is a very significant aspect of automatic speech recognition (ASR) because its performance severely degrades due to the noise which unavoidably exists in our living space. Several simple and effective techniques to suppress stationary noise, e.g. spectral subtraction (SS) for additivity noise [1, 2] and cepstral mean normalization (CMN) for channel distortion[3], have been developped so far. Recently, many speech enhancement methods using microphone array have been proposed as the front-end to refine the ASR robustness for nonstationary noise [4, 5]. Especially stereo-input ASR becomes a promissing approach since existing devices, such as normal PC and IC recorder, have a two-channel input. Therefore, this paper focuses on development of noise robust ASR techniques by using two-channel input devices.

In a real environment, we often hear interferences with target speech. And locations of the target and interferences are usually not known. In addition, the number of sound sources might be greater than that of microphones in the scenario of two-channel devices. Therefore, we should deal with a underdetermined blind source separation (BSS) problem. BSS is defined as a problem to separate multiple source signals from mixtures without any prior information about mixing process. One can separate sources using estimated inverse of the mixing matrix if the number of sources is equal to or less than that of mixtures. However, in the case where sources outnumber mixtures, i.e., underdetermined case, one can not separate them even if appropriate mixing matrices are estimated. In that sense, underdetermined BSS is a hard problem but matches our scenario.

Time-frequency masking based on speech sparseness is an effective approach for underdetermined BSS. Here sparseness means a property of speech that its energy is concentrated in a small area of time-frequency plain. Most of masking methods assume that individual source does not overlap in the time-frequency domain and obtain the target signal by multiplying an appropriate mask by the observation. Time-frequency mask can be classified into two types: 1) binary mask which has a value 0 or 1 and 2) soft mask which has a continuous value $[0, 1]$ at each time-frequency bin respectively. The cue to design masks is the time delay between two-channel observed signals. However, it is disturbed by background noise and reverberation because they are not sparse and comes from various directions. Additionaly, time-frequency masking itself might cause some distortion to the target speech signal from the viewpoint of the ASR.

Although it is well known that time-frequency masking causes distortion called musical noise, we will show it is very effective to suppress interference as the front-end of ASR system in this paper. We have developed a time-frequency masking method based on maximum likelihood estimation and it has good separation performance in terms of SNR. We will show that our soft masking method is not only able to separate target signal robustly but also performs less distortion for ASR system than binary masking method. Moreover, the remained distortion can be further reduced by CMN and consequently, recognition performance improves considerably. To show the effectiveness of our method, we performed detail investigation through connected digit recognition tasks in reverberant situation in both of simulation and real environments. It contains comparison between conventional binary, our binary and soft masking separation and investigation of the effect of CMN.

†current affiliation: NTT Information Sharing Platform Laboratories 3-9-11 Midori-cho, Musashino-city, Tokyo, Japan.

2. Sparseness-based time-frequency masking in a reverberant environment

2.1. Time-frequency mask estimation by EM algorithm

In this section, we introduce our time-frequency masking method [7] briefly. We assume that target speech and interferences are sparse in the time-frequency domain. Here sparsity means that the energy of signal is distributed in a small area in that domain. For example, speech is sparse because of its specific structure along temporal and frequency axes. Hence, one can assume that each signal does not overlap and only a single source is active at each time-frequency bin. Assuming the two microphones have the same sensitivity and the difference of source position causes only the time difference of arrival, the two-channel observation $\boldsymbol{M}(\tau,\omega) = (M_L(\tau,\omega), M_R(\tau,\omega))^T$ can be written by

$$\boldsymbol{M}(\tau,\omega) = S_k(\tau,\omega)\boldsymbol{b}_k(\omega) + \boldsymbol{N}(\tau,\omega), \tag{1}$$

where $\boldsymbol{b}_k = (1, \exp(j\omega\delta_k))$ is the transfer function for the active source $S_k(\tau,\omega)$ to arrive at microphones directly as a plane wave and $\boldsymbol{N}(\tau,\omega) = (N_L(\tau,\omega), N_R(\tau,\omega))$ is the noise component which contains reverberation and background noise. The subscript k represents the index of active source at (τ,ω). For simplisity, we will omit (τ,ω) and denote just S_k, $\boldsymbol{M}$, $\boldsymbol{N}$ and so on if there is no ambiguity. Note that we can easily extend the model to one with the attenuation of spherical sound wave.

Our method estimates time-frequency masks based on maximum likelihood. Modeling that $\boldsymbol{N}$ follows the Gaussian distribution, the likelihood that kth source is active at (τ,ω) can be written by

$$\begin{aligned} &p(\boldsymbol{M}|\delta_k, \sigma^2, S_k) \qquad (2)\\ &= \frac{1}{2\pi|V|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{M} - S_k\boldsymbol{b}_k)^H V^{-1}(\boldsymbol{M} - S_k\boldsymbol{b}_k)\right), \end{aligned}$$

where ^{H} means Hermitian transposition, σ^2 is the noise power and V represents the noise covariance matrix which we applied diffuse sound field model:

$$V = \mathrm{E}[\boldsymbol{N}\boldsymbol{N}^H] = \sigma^2 \begin{pmatrix} 1 & \mathrm{sinc}(\omega D/c) \\ \mathrm{sinc}(\omega D/c) & 1 \end{pmatrix}, \tag{3}$$

where D is the distance between microphones and c represents sound velocity.

Our problem is to estimate which source is active at each (τ,ω) through maximizing log likelihood defined on the whole time-frequency plain:

$$LL\left(\boldsymbol{M}|\, \delta_1, \cdots, \delta_K, \sigma^2, S_1, \cdots, S_K\right), \tag{4}$$

where K means the number of sources. In other words, parameters we have to estimate are not only $\delta_k, \sigma^2, S_k \quad (k = 1, \cdots, K)$, but also the index k. This is a typical missing data problem which can be solved by the EM algorithm.

Applying the EM algorithm, maximum likelihood parameters can be estimated by iteration of calculating so-called Q function (E step) and maximizing Q function (M step). Update rules are derived as below.

E step:

$$m_{\tau,\omega,k} \leftarrow \frac{r_k\, p(\boldsymbol{M}|\delta_k, \sigma^2, S_k)}{\sum_{k'} r_{k'}\, p(\boldsymbol{M}|\delta_{k'}, \sigma^2, S_{k'})} \tag{5}$$

$$Q(\delta_k, \sigma^2, S_k) = \sum_{\tau,\omega,k} m_{\tau,\omega,k} \log p(\boldsymbol{M}|\delta_k, \sigma^2, S_k), \tag{6}$$

where r_k means the a priori probability that kth source is active and is subject to $\sum_k r_k = 1$.

M step:

$$r_k \leftarrow \frac{\sum_{\tau,\omega} m_{\tau,\omega,k}}{\sum_{\tau,\omega,k'} m_{\tau,\omega,k'}}, \tag{7}$$

$$S_k \leftarrow \frac{\boldsymbol{b}_k^H V^{-1}\boldsymbol{M}}{\boldsymbol{b}_k^H V^{-1}\boldsymbol{b}_k}, \tag{8}$$

$$\begin{aligned} \sigma^2 \leftarrow\ & \frac{1}{2C} \sum_{\tau,\omega,k} \frac{m_{\tau,\omega,k}}{1 - \mathrm{sinc}^2(\omega D/c)} \\ & \times \left(\boldsymbol{M}^H V^{-1}\boldsymbol{M} - \frac{\boldsymbol{b}_k^H V^{-1}\boldsymbol{M}}{\boldsymbol{b}_k^H V^{-1}\boldsymbol{b}_k} \right) \end{aligned} \tag{9}$$

$$\delta_k \leftarrow \operatorname*{argmax}_{\delta_k} Q(\delta_k, \sigma^2, S_k), \tag{10}$$

where C represents the number of all time-frequency bins. Since the update rule of δ_k cannot be derived analiticaly and it is within a range $[-D/c, D/c]$, we updated δ_k by discrete scannning of maximum of Q function. The iteration is performed until the increment of log likelihood shown in Eq. (4) becomes smaller than a certain threshold.

2.2. Binary and soft masking

We can design masks in two ways, i.e., binary and soft masks. Although both methods have been proposed in BSS context, comparison of them in terms of speech recognition have rarely been performed. Generally, binary mask, which is also called as hard mask, has a value 1 at time-frequency bin where the target speech is active, and has 0 in other area. It just switches passing and cutting all energy at each bin. And soft mask, or ratio mask, has a continuous value within the range $[0, 1]$ at each bin. It is known that soft masking shows best performance if it has a value defined as a ratio between power of the target source and noise [9]. We defined two types of masking based on our algorithm described in the previous section as below:

Binary masking:

$$\hat{S}_k(\tau,\omega) = \begin{cases} \dfrac{\boldsymbol{b}_k^H V^{-1}\boldsymbol{M}}{\boldsymbol{b}_k^H V^{-1}\boldsymbol{b}_k} & (m_{\tau,\omega,k} > m_{\tau,\omega,k'}) \\ 0 & (\text{otherwise}) \end{cases}. \tag{11}$$

Soft masking:

$$\begin{aligned} E[S_k] &= \sum_{k'} m_{\tau,\omega,k} E_{k'}[S_k] \\ &= m_{\tau,\omega,k} E_k[S_k] + \sum_{k' \neq k} m_{\tau,\omega,k'} E_{k'}[S_k] \\ &= m_{\tau,\omega,k} \frac{\boldsymbol{b}_k^H V^{-1}\boldsymbol{M}}{\boldsymbol{b}_k^H V^{-1}\boldsymbol{b}_k}. \end{aligned} \tag{12}$$

Binary mask has the value $\boldsymbol{b}_k^H V^{-1}/\boldsymbol{b}_k^H V^{-1}\boldsymbol{b}_k$ at (τ,ω) if the kth target signal has the largest likelihood. Otherwise it has zero. On the other hand, soft mask has a continuous value between zero and $\boldsymbol{b}_k^H V^{-1}/\boldsymbol{b}_k^H V^{-1}\boldsymbol{b}_k$ depending on $m_{\tau,\omega,k}$ which corresponds to the probability that kth signal is active. Also, the soft-masked signal can be derived as the expectation of S_k.

3. Recognition experiment

3.1. Experiment objective and condition

We have performed two experiments. First, we compared recognition performances between binary mask and soft mask.

Table 1: Sounds used in the experiments.

	S_1	S_2	S_3
N1	target speech	baby crying	cleaner
N2	target speech	reading voice	impact noise

Table 2: experiment condition

	t-f masking	speech recognition
sampline frequency	16 kHz	
pre-emphasis		$1 - 0.97z^{-1}$
frame size	64 ms	25 ms
shift	32 ms	10 ms
window function	Hamming window	

And secondly we have investigated recognition performance of our method in both noisy and reverberant environment of both simulation and real situation.

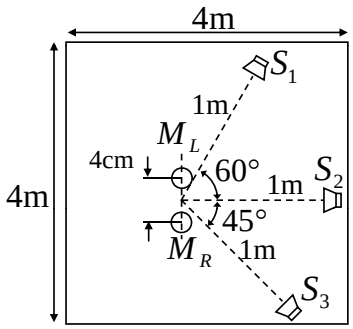
Our recognition task was Japanese connected digit recognition which conforms to CENSREC-4, the reverberant speech database [10]. The target speech for evaluation was randomly selected from clean speech of CENSREC-4 which contains 6,556 words uttered by 52 people. Two sets of interference noises were selected from a noise database SMILE2004 [11], as shown in Table. 1. Impact noise was a interval randomly picked up from a long signal which we made by cutting out 5 sec from "claps", "auto door", "printing", "dragging a chair" and "wind bell" and connecting them with 0.5 sec silence. In the simulation experiment, devices were aligned as shown in Fig. 1(a) and reverberation was simulated by the mirror method in three cases where the reverberation time were 0ms, 270ms and 468ms, respectively. We have also performed an experiment in a real environment with the same condition. Signals observed in this real room was contaminated by interferences and reverberations. The noise level was -3 to -7 dB in terms of SNR against original clean target speech. The target speech and two interferences were aligned as Fig. 1(b). We recorded their mixture with a commercially available IC recorder. The room was a normal office room, which has no noise insulation device installed.

Experimental conditions of masking and recognition are shown in Table. 2. We first obtained the time-frequency representation of signals and performed the masking. Next, we resynthesized separated signals and recognized them. As a typical conventional method, we also obtained speech signals separated by a binary masking method called DUET [6]. Extracted features were 38 dimension which includes MFCC (12 dimension), ΔMFCC, $\Delta\Delta$MFCC, $\Delta \log$ power and $\Delta\Delta \log$ power. In addition, we have also investigated the effect of CMN [3]. The cepstral mean is estimated by computing the average of each cepstral parameter across each masked speech.

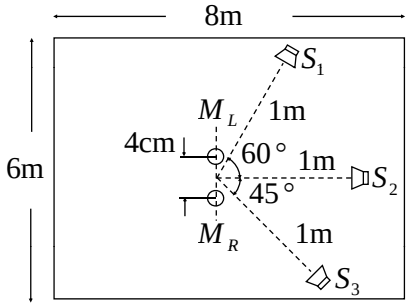
We used HTK ver 3.4 for training and recognition. The acoustic model was trained with clean 8,440 speeches of CENSREC-4 uttered by 55 people. It showed 99.3% of word accuracy (WA) for clean original speech.

3.2. Preliminary experiment to evaluate the separation

Before the recognition experiments, we confirmed separation performance of our method in terms of SNR by using simulation data. Table. 3 shows averaged SNRs of all noise set and reverberation conditions. We can see that binary mask and soft mask based on our algorithm show almost the same performance and



(a) Simulation. Height of microphones and sources are 1.5 m.



(b) Real room. Height of microphones and sources are 0.9 m.

Figure 1: The alingment of three sources (S_1, S_2, and S_3) and two microphones (M_L, M_R). The height of room is both 3 m.

Table 3: Comparison of separation performance among three two-channel BSS methods.

SNR[dB]	separated signal	improvement from mixture
DUET	5.87	11.34
Binary mask	8.13	13.60
Soft mask	8.29	13.76

they are better than that of DUET. Our algorithm is robust at least in terms of the SNR measure compared to DUET.

3.3. Comparison of results between binary and soft mask

In this section, we have compared distortions caused by binary and soft masks. Actually, separated signal through masking has various type of distortions such as one due to interferences, reverberation and mask itself. To investigate the effect of each mask itself, we multiplied masks by clean speeches which is estimated from noisy mixtures of simulation data. Table. 4 shows WAs which were calculated by averaging scores of all environments. While we can see that both of binary and soft masks caused distortion and degraded speech recognition rates, the degradation of soft mask was less than that of binary mask. And we can also see that CMN reduced the distortion greatly, especially in the soft mask case.

This results shows that soft mask is favorable as a front-end of speech recognition and CMN reduces the distortion caused by soft mask very well. In other words, the distortion caused by soft mask can be approximated by stationary multiplicative noise in the mel log spectrum domain.

3.4. Recognition result of simulation

We performed speech recognition experiments by using no mask (recognizing original mixture), DUET, binary mask and

Table 5: Recognition performance (WA(%)) in a simulation environment. BMask means binary mask and SMask means soft mask

method	0ms		270ms		468ms		Ave.
	N1	N2	N1	N2	N1	N2	
NoMask	19.9	12.2	11.6	9.1	10.3	8.2	11.9
NoMask+CMN	45.1	34.5	26.3	25.0	18.7	17.6	27.9
DUET	23.5	27.7	19.9	14.4	12.9	11.2	18.2
DUET+CMN	63.7	70.3	57.9	46.3	30.7	27.0	42.3
BMask	23.8	34.5	14.4	19.2	11.5	13.0	19.4
BMask+CMN	62.2	85.6	37.4	56.9	24.1	31.6	49.6
SMask	39.6	40.8	32.1	29.4	23.7	21.8	31.3
SMask+CMN	**93.2**	**95.7**	**89.9**	**88.6**	**74.7**	**71.7**	**85.6**

Table 6: Recognition performance (WA(%)) in a real environment.

	N1	N2	Ave.
NoMask	13.7	5.9	9.8
NoMask+CMN	37.9	13.4	25.7
DUET	13.1	11.6	12.4
DUET+CMN	33.1	32.4	32.8
BMask	7.2	10.3	8.8
BMask+CMN	31.6	46.8	39.2
SMask	19.2	16.4	17.8
SMask+CMN	**64.0**	**59.9**	**62.0**

Table 4: Comparison of speech recognition performance (WA(%)) between two time-frequency masking methods.

method	WA (%)
Binary mask	28.0
Binary mask+CMN	75.9
Soft mask	41.0
Soft mask+CMN	96.9

soft mask in a simulated reverberant environment. The results are shown in Table. 5. Our method, or soft mask with CMN, showed the best performance in all situation. On an average, our method achieved WA of 85.8%, which was a considerable improvement compared to 42.3% of DUET with CMN case and 49.6% of binary mask with CMN case. All masking results without CMN were not much different in terms of WA while they showed different performances in terms of SNR in Table. 3. Although soft mask and binary mask outperformed DUET in terms of SNR, only soft mask with CMN outperforms greatly in terms of WA. From this result, we can see that CMN also reduced the distortion of residual interferences which separated signal by soft masking contains. We can consider this is because separated signals by binary mask were too rough to preserve the envelope of power spectrum. While soft mask also caused some distortion, its error was less than that of binary mask because it has a continuous value at each time-frequency bin and keeps speech envelope flexibly.

3.5. Recognition result of real situation

We also performed recognition experiment in a real environment to investigate the performance of our method. The result is shown in Table. 6. Similarly to the simulation case, soft mask with CMN outperforms other methods greatly. We can see that separation by masks failed from the result of all masks comparing the case of simulation. Background noise had large power, which is the difference from the simulation, degraded mask separation performance. Nevertheless, we can also see that CMN reduces the distortion and make a good performance of WA. On an average, soft mask with CMN achieved WA of 62.0%, which improved from 32.8% of DUET with CMN case and 39.2% of binary mask with CMN case. This result shows that our method works well as a front-end of ASR even in a real environment.

4. Conclusion

In this paper, we presented a noise-robust stereo-input ASR system using sparseness-based BSS technique in a reverberant environment. We applied our noise-robust time-frequency masking method as the front-end of ASR system, then we showed that soft mask is better than binary mask in terms of recognition rate. The reason of this is not clear yet and we will perform a detailed analysis in the future work. Next, we revealed the fact that CMN reduces the distortion caused by masks, especially for that caused by soft mask greatly. At the end, we evaluate the recognition performance of our method in noisy and reverberant real environment. The experiment result showed that our soft masking method followed by CMN has a good recognition performance.

5. References

[1] S.F. Boll, "Supression of acoustic noise in speech using spctral subtraction," IEEE Trans. Acoustics, Speech, and Signal Processing, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[2] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," Proc. ICASSP, vol. 4, pp. 208–211, Apr. 1979.

[3] F.H. Liu, R.M. Stern, X. Huang, and A. Acero, "Efficient cepstral normalization for robust speech recognition," Proc. ARPA Workshop on Human Language Technology, pp. 69–74, Mar. 1993.

[4] M. Matassoni, M. Omologo, and D. Giuliani, "Handsfree speech recognition using a filtered clean corpus and incremental hmm adaptation," Proc. ICASSP, pp. 1407–1410, Apr. 2000.

[5] Y. Takahashi, T. Takatani, H. Saruwatari, and K. Shikano, "Blind spatial subtraction array with independent component analysis for hands-free speech recognition," Proc. IWAENC, Sep. 2006.

[6] O. Yilmaz, and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," IEEE Transaction on Signal Processing, vol. 52, no. 7, pp. 1830–1847, 2004.

[7] Y. Izumi, N. Ono, and S. Sagayama, "Sparseness- based 2ch BSS using the EM algorithm in reverberant environment," Proc. WASPAA, pp. 147–150, Oct. 2007.

[8] R.K. Cook, R.V. Waterhouse, R.D. Berendt, S. Edelman, and M.C. Thompson, "Measurement of correlation coefficients in reverberant sound fields," JASA, vol. 27, no. 6, pp. 1072–1077, 1955.

[9] Y. Li and D. Wang, "On the optimality of ideal binary time-frequency masks," Proc. ICASSP, pp. 3501–3504, Apr. 2008.

[10] M. Nakayama, T. Nishiura, Y. Denda, N. Kitaoka, K. Yamamoto, T. Yamada, S. Tsuge, C. Miyajima, M. Fujimoto, T. Takiguchi, S. Tamura, T. Ogawa, S. Matsuda, S. Kuroiwa, K. Takeda, and S. Nakamura, "CENSREC-4: Development of evaluation framework for distant-talking speech recognition under reverberant environments," Proc. Interspeech, pp. 968–971, Sep. 2008.

[11] K. Kawai, K. Fujimoto, T. Iwase, H. Yasuoka, T. Sakuma, and Y. Hidaka, "Development of a sound source database for environmental/architectural acoustics: Introduction of SMILE 2004 (sound ma- terial in living environment 2004)," Proc. ICA, pp. 1561–1564, 2004.