Improving the robustness of multiple input spectrogram inversion *

○ Dongxiao Wang† Hirokazu Kameoka‡ Koichi Shinoda†
† Tokyo Institute of Technology
‡ NTT Communication Science Laboratories

1 Introduction

We focus on the single channel source separation (SCSS) problem where multiple sources are collected by a single sensor. While most of its studies have neglected the phase information, it becomes essential for some applications such as hearing aid devices.

To estimate phases, Gunawan *et al.*[1] proposed multiple input spectrogram inversion (MISI). While it significantly improves the quality of recovered signals when the true magnitude spectrogram of each underlying source is given, the performance drops drastically when only erroneous magnitude spectrograms or only some of them are available.

In this paper, we propose an algorithm that is robust against the uncertainty of the magnitude spectrum. In addition, it is flexible enough to allow for the presence of residual components. It also deals with the different degrees of uncertainty among the component signals.

2 Previous Studies

Here we briefly introduce Multiple Input Spectrogram Inversion (MISI). When given the mixture signal $\tilde{\mathbf{y}}$ and magnitude spectrogram $\mathbf{a}_j$ of each source j, this algorithm estimates signal $\tilde{\mathbf{c}}_j$ and phase $\boldsymbol{\phi}_j$ by solving the following optimization problem:

$$\text{minimize} \quad \mathcal{J}(\boldsymbol{\phi}, \tilde{\mathbf{c}}) = \sum_j \left\| \mathbf{a}_j \odot \boldsymbol{\phi}_j - \mathbf{W}\tilde{\mathbf{c}}_j \right\|_2^2, \tag{1}$$

$$\text{subject to} \quad \sum_j \tilde{\mathbf{c}}_j = \tilde{\mathbf{y}}. \tag{2}$$

Here $\mathbf{W}$ denotes short time Fourier transform (STFT). Since it is difficult to solve them directly, it utilizes an iterative process:

$$\tilde{\mathbf{c}}_j \leftarrow \mathbf{W}^+ \left(\mathbf{a}_j \odot \boldsymbol{\phi}_j\right) + \frac{1}{J}\left(\tilde{\mathbf{y}} - \sum_{j'} \mathbf{W}^+ \left(\mathbf{a}_{j'} \odot \boldsymbol{\phi}_{j'}\right)\right), \tag{3}$$

$$\boldsymbol{\phi}_j \leftarrow \angle \mathbf{W}\tilde{\mathbf{c}}_j, \tag{4}$$

Here $\mathbf{W}^+$ denotes the inverse STFT.

In MISI, the summation of each source should exactly be the mixture signal. This constraint is often too hard to be satisfied. Another observation is that the error between mixture signal $\tilde{\mathbf{y}}$ and the sum of estimated signal $\sum_{j'} \mathbf{W}^+(\mathbf{a}_{j'} \odot \boldsymbol{\phi}_{j'})$ is equally distributed over all sources. This also may not be practical since this error can be different from source to source.

3 Modified MISI

To mitigate the two problems mentioned above, we propose a new algorithm, Modified MISI (M-MISI), which employs a different objective function:

$$\begin{aligned} \text{minimize} \quad \mathcal{I}(\boldsymbol{\phi}) = &\sum_j \left\| \mathbf{a}_j \odot \boldsymbol{\phi}_j - \mathbf{W}\mathbf{W}^+(\mathbf{a}_j \odot \boldsymbol{\phi}_j) \right\|_2^2 \\ &+ \lambda \left\| \mathbf{y} - \sum_j \mathbf{a}_j \odot \boldsymbol{\phi}_j \right\|_2^2 . \end{aligned} \tag{5}$$

Here $\mathbf{y}$ is a vector of the complex spectrogram of the mixture signal. The first term of (5) measures how much the estimated phase spectrogram satisfies the constraint introduced by overlapping windowing function in STFT. The second term measures the distance between the mixture spectrogram and the sum of estimated spectrogram of each source, relaxes the hard constraint (2) in the original MISI. Note that it is not always necessary to satisfy (2).

In addition we introduce weight $\boldsymbol{\beta}$ for each component signal where $\sum_j \beta_{j,f,n} = 1, 0 < \beta_{j,f,n} < 1$. Let $y_{f,n} = \sum_j x_{j,f,n}$, and the vector notation of $x_{j,f,n}$ be $\mathbf{x}_j$. By solving (5) in the similar way with [2], we get the update rules of each variable:

$$\tilde{\mathbf{c}}_j \leftarrow \mathbf{W}^+(\mathbf{a}_j \odot \boldsymbol{\phi}_j) \tag{6}$$

$$\mathbf{x}_j \leftarrow \mathbf{a}_j \odot \boldsymbol{\phi}_j + \boldsymbol{\beta}_j \odot \left(\mathbf{y} - \sum_{j'} \mathbf{a}_{j'} \odot \boldsymbol{\phi}_{j'}\right) \tag{7}$$

$$\boldsymbol{\phi}_j \leftarrow \angle \left(\frac{\boldsymbol{\beta}_j \odot \mathbf{W}\tilde{\mathbf{c}}_j + \lambda \mathbf{x}_j}{\boldsymbol{\beta}_j + \lambda \mathbf{1}}\right) \tag{8}$$

$$\boldsymbol{\beta}_j \leftarrow \frac{\text{abs}(\mathbf{x}_j - \mathbf{a}_j \odot \boldsymbol{\phi}_j)}{\sum_{j'} \text{abs}(\mathbf{x}_{j'} - \mathbf{a}_{j'} \odot \boldsymbol{\phi}_{j'})}, \tag{9}$$

where $\mathbf{1}$ is an all-ones vector, $\div$ denotes the element-wise division and abs$(\cdot)$ denotes an operation that

*多入力スペクトログラムインバージョンの頑健性の改善、王冬暁、亀岡弘和、篠田浩一

Table 1: Comparison of SDR, SAR and SIR among Wiener filtering, MISI and Modified-MISI. Each value represents the best value in 60 iterations, averaged over 100 experiments.

Noise SSNR	−10 dB			0 dB			+10 dB			+20 dB		
	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR
Wiener	12.13	19.55	13.19	12.14	19.41	13.28	12.56	19.98	13.66	12.56	20.09	13.63
MISI	2.87	5.95	8.90	7.84	13.35	11.85	14.34	20.53	17.96	22.79	28.39	26.24
M-MISI	11.95	24.56	12.79	13.27	26.46	14.03	17.45	30.31	18.11	24.45	35.26	25.04

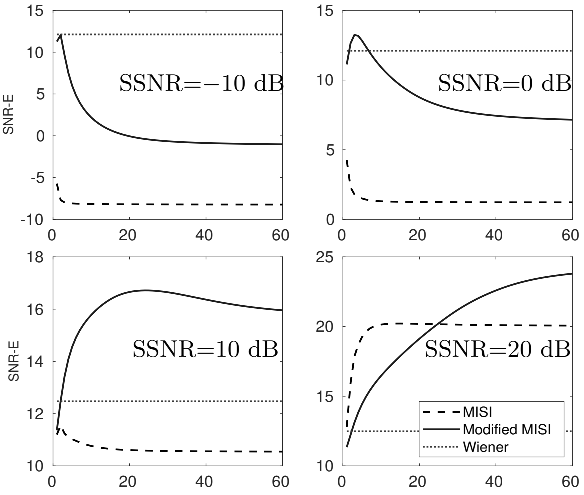

Fig. 1: Comparison of SNR among Wiener filtering, MISI and Modified-MISI.

takes the absolute value of each element of a vector. And we choose λ as the ratio between the two terms of the objective function.

4 Experimental Evaluation

4.1 Setup

We evaluate the proposed algorithm on synthesized data assuming that only imperfect magnitude spectrogram is known beforehand. Here we randomly add Gamma noise to the magnitude spectrogram of each source with some specific Spectral Signal-to-Noise Ratio (SSNR) defined as: SSNR $\equiv 10\log_{10}\left(\frac{\sum_{f,t} s_{f,t}^2}{\sum_{f,t} n_{f,t}^2}\right)$, where s and n stand for signal and noise, respectively. The mixtures are generated from the ATR speech database [3]. All utterances are re-sampled to 16 kHz.

We use Signal-to-Distortion, Interference and Artifacts Ratio (SDR, SIR and SAR) as evaluation metrics. We also report SNR in estimation, SNR-E defined as the SNR between ground truth signal s and the difference between s and the estimated signal s_e. We compare our method with MISI and Wiener filtering.

4.2 Experimental Results

Table 1 shows the overall comparison between M-MISI, MISI, and Wiener filtering. In most cases the proposed M-MISI gave significantly better performance than MISI and Wiener filtering.

Figure 1 compares SNR-E among those methods. MISI and M-MISI performed poorly with SNR −10 dB and 0 dB. Our method achieved a larger SNR-E with SNR 10 dB and 20 dB.

5 Conclusion

We have proposed a new phase reconstruction algorithm, M-MISI, which employs a soft objective function. This algorithm is more robust when magnitude spectra of each source are erroneous. In our experiment, the proposed M-MISI showed a significant improvement when we added gamma noise on the simulated data. In future, we would like to tackle the case when the number of recording channels (microphones) is more than one.

References

[1] D. Gunawan and D. Sen, “ Iterative phase estimation for the synthesis of separated sources from single-channel mixtures, ”IEEE Signal Processing Letters, vol. 17, no. 5, pp. 421424, May 2010.

[2] J. Le Roux, H. Kameoka, N. Ono, and S. Sagayama, “ Fast signal reconstruction from magnitude stft spectro- gram based on spectrogram consistency, ” in Proc. Int. Conf. Digital Audio Effects, vol. 10, 2010.

[3] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, “ Atr japanese speech database as a tool of speech recognition and synthe- sis, ” Speech communication, vol. 9, no. 4, pp. 357363, 1990.