INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

A Modified Algorithm for Multiple Input Spectrogram Inversion

Dongxiao Wang[1], *Hirokazu Kameoka*[2], *Koichi Shinoda*[1]

[1]Tokyo Institute of Technology, Japan
[2]NTT Communication Science Laboratories, Japan

dongxiao@ks.cs.titech.ac.jp, kameoka.hirokazu@lab.ntt.co.jp, shinoda@c.titech.ac.jp

Abstract

We propose a new algorithm to estimate the phase of speech signal in the mixture of audio sources under the assumption that the magnitude spectrum of each source is given. The previous method, multiple input spectrogram inversion algorithm (MISI), often performs poorly when the magnitude spectrograms estimated are not accurate. This may be because it imposes a strict constraint that the summation of source waveforms should be exactly the same as the mixture waveform. Our proposing algorithm employs a new objective function in which this constraint is relaxed. In this objective function, the difference between the summation of source waveforms and the mixture waveform is the target to be minimized. The performance of our method, modified MISI is evaluated on two different experimental settings. In both settings it improves the audio source separation performance compared to MISI.

Index Terms: Phase estimation, MISI, source separation, synthesis

1. Introduction

Single channel source enhancement (SCSE) has been widely applied in hearing-aids, teleconferencing and speech recognition systems. In these applications, a mixture of target signal and background noise is collected by a single microphone, and noise reduction methods are applied to remove the background noise.

SCSE has been intensively studied (for a detailed survey, see [1]). Most approaches use short time Fourier transform (STFT) to transfer time domain mixture signals to spectra, then estimate the magnitude of target source by several different ways. Next, they recover the waveform of target source by the inverse short time Fourier transform (iSTFT) where the phase of mixture is used for the phase of each individual sources. Of course the phase of each source is different from that of the mixture, and so, this process cannot be justified.

The STFT of a signal is its redundant representation when we use an analysis window which largely overlaps with nearby windows; The spectral coefficients of successive frames are correlated. Griffin and Lim [2] utilizes this fact to recover phase from magnitude spectrum. For multi-source situation, however, separated sources are not independent with each other, *i.e.* all of the individual sources should exactly add up to the mixture signal. Gunawan and Sen [3] proposed multiple input spectrogram inversion (MISI) to utilize this information. While this approach works well when the true magnitude of each source is given, the performance significantly degrades when inaccurate magnitudes are given.

To address this problem, different approaches have been proposed for different settings. A method named ISSIR is introduced in [4] for informed source separation (ISS), where magnitudes and phases are jointly updated only when the T-F bin is "active". Another joint estimation approach named *consistent wiener filtering* [5] combines the consistency constraint with a loss between the magnitude estimation and classical wiener filter output, resulting in a promising performance when the given magnitude spectra are estimated from spectral subtraction. In recent research [6], MISI procedure is unfolded and integrated into the chimera++ network, which improves the magnitude-only methods. We still have a strong motivation of finding a robust algorithm for the phase estimation, since it can be used in a source separation setting, as a post-processing procedure after applying magnitude oriented approaches, such as [7–13].

In this paper, we propose an algorithm which is robust against the inaccurate magnitude spectrum, by relaxing the strict constraint as an additional term in the objective function.

2. Griffin & Lim Algorithm

Let $\mathbf{x} = [x(0), \ldots, x(T-1)]^\top \in \mathbb{R}^T$ be a time domain signal, $f = 1, \ldots, F$ and $n = 1, \ldots, N$ be frequency and frame indices, respectively, where F is the number of frequency bins in each frame, T is the length of the signal, and N is the number of frames. Also let $\mathbf{w}_{f,n} = [w_{f,n}(0), \ldots, w_{f,n}(T-1)]^\top$ be a complex sinusoid of frequency ω_f modulated by the window function of the n-th frame. Note that $\mathbf{w}_{f,n}$ is padded with zeros over the range outside the frame window. Then, the complex Fourier coefficient $c_{f,n}$ of the f-th bin at the n-th frame is the inner product between $\mathbf{x}$ and $\mathbf{w}_{f,n}$, namely $c_{f,n} = \mathbf{w}_{f,n}^{\mathsf{H}}\mathbf{x}$, where $\cdot^{\mathsf{H}}$ denotes Hermitian transpose. Let $\mathbf{c} \in \mathbb{C}^{FN}$ be a vector obtained by stacking $c_{f,n}$ over all the frequency bins F of all the N frames. Then the relationship between $\mathbf{c}$ and $\mathbf{x}$ can be written as

$$\mathbf{c} = \mathbf{W}\mathbf{x}. \tag{1}$$

Here $\mathbf{W}$ is a $FN \times T$ matrix in which each row is $\mathbf{w}_{f,n}^{\mathsf{H}}$, $\mathbf{c}$ is a complex spectrogram and $\mathbf{W}$ is the the matrix of complex exponentials functions used in the short time Fourier transform (STFT) modulated by the window function.

We typically use overlapping windows, FN is larger than T, and accordingly, $\mathbf{c}$ is a redundant representation of the signal $\mathbf{x}$. The inverse STFT of $\mathbf{c}$ can be written using the pseudo-inverse matrix $\mathbf{W}^+$ of $\mathbf{W}$:

$$\begin{aligned}\mathbf{W}^+\mathbf{c} &= \operatorname*{argmin}_x \|\mathbf{c} - \mathbf{W}\mathbf{x}\|_2^2 \\ &= (\mathbf{W}^{\mathsf{H}}\mathbf{W})^{-1}\mathbf{W}^{\mathsf{H}}\mathbf{c}.\end{aligned} \tag{2}$$

$\mathbf{c}$ is consistent iff $\mathbf{c} = \mathbf{W}\mathbf{W}^+c$. Griffin and Lim [2] utilize this consistency constraint to estimate the phase of each frequency component.

http://dx.doi.org/10.21437/Interspeech.2019-3242

Let $\mathbf{a}$ and $\boldsymbol{\phi}$ be the magnitude and phase spectra of $\mathbf{c}$, respectively. Namely $\mathbf{c} = \mathbf{a} \odot \boldsymbol{\phi}$, where $\phi_{f,n} \equiv e^{i\theta_{f,n}}$ and $\odot$ denotes an element-wise product. Here we estimate $\boldsymbol{\phi}$ assuming $\mathbf{a}$ is known. Then we can obtain the phase $\boldsymbol{\phi}$ by minimizing the following objective function $\mathcal{J}(\boldsymbol{\phi})$:

$$\mathcal{J}(\boldsymbol{\phi}) = \left\|\mathbf{a} \odot \boldsymbol{\phi} - \mathbf{W}\mathbf{W}^{+}(\mathbf{a} \odot \boldsymbol{\phi})\right\|_2^2. \tag{3}$$

Instead of solving Eq.(3) directly, we introduce another variable $\tilde{\mathbf{c}}$ and minimize the following function:

$$\mathcal{J}^{+}(\boldsymbol{\phi}, \tilde{\mathbf{c}}) \equiv \|\mathbf{a} \odot \boldsymbol{\phi} - \mathbf{W}\tilde{\mathbf{c}}\|_2^2. \tag{4}$$

We iteratively update ϕ and $\tilde{\mathbf{c}}$ using the principle of majorization and minimization [14]:

$$\tilde{\mathbf{c}} \leftarrow \underset{\tilde{\mathbf{c}} \in \mathbb{R}^T}{\operatorname{argmin}} \|\mathbf{a} \odot \boldsymbol{\phi} - \mathbf{W}\tilde{\mathbf{c}}\|_2^2 = \mathbf{W}^{+}(\mathbf{a} \odot \boldsymbol{\phi}), \tag{5}$$

$$\boldsymbol{\phi} \leftarrow \underset{\boldsymbol{\phi}}{\operatorname{argmin}} \|\mathbf{a} \odot \boldsymbol{\phi} - \mathbf{W}\tilde{\mathbf{c}}\|_2^2 = \angle \mathbf{W}\tilde{\mathbf{c}}, \tag{6}$$

where $\angle \cdot$ denotes an operation that divides each element of a vector by its absolute value.

3. Multiple Input Spectrogram Inversion

For reconstructing individual source signals that compose a mixture signal, Gunawan and Sen [3] proposed a closed loop method called multiple input spectrogram inversion (MISI) where the magnitude spectra of each source is assumed to be given.

Let $\tilde{\mathbf{y}}$ be the mixture of each source, $\mathbf{a}_j$ and $\boldsymbol{\phi}_j$ be the magnitude and phase spectra of source j, respectively. We introduce another variable $\tilde{\mathbf{c}}_j$ for source j, then MISI iteratively updates $\tilde{\mathbf{c}}_j$ and $\boldsymbol{\phi}_j$:

$$\tilde{\mathbf{c}}_j \leftarrow \mathbf{W}^{+}\left(\mathbf{a}_j \odot \boldsymbol{\phi}_j\right) + \frac{1}{J}\left(\tilde{\mathbf{y}} - \sum_{j'} \mathbf{W}^{+}\left(\mathbf{a}_{j'} \odot \boldsymbol{\phi}_{j'}\right)\right), \tag{7}$$

$$\boldsymbol{\phi}_j \leftarrow \angle \mathbf{W}\tilde{\mathbf{c}}_j, \tag{8}$$

Although it is not explicitly stated in [3], this algorithm solves the following constrained optimization problem:

$$\text{minimize} \qquad \mathcal{J}(\boldsymbol{\phi}, \tilde{\mathbf{c}}) = \sum_j \|\mathbf{a}_j \odot \boldsymbol{\phi}_j - \mathbf{W}\tilde{\mathbf{c}}_j\|_2^2, \tag{9}$$

$$\text{subject to} \qquad \sum_j \tilde{\mathbf{c}}_j = \tilde{\mathbf{y}}. \tag{10}$$

The derivation can be checked in

In MISI, the summation of individual sources should exactly be the mixture signal. This constraint is often too hard to be satisfied and may lead to incorrect updates when only erroneous magnitude spectra are available. Another observation from (7) 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. While we assume that the true magnitude spectrum for each source is given, the magnitude spectrum of one source may be more reliable than that of other sources in real situations.

4. Modified MISI

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

$$\begin{aligned}\mathcal{I}(\boldsymbol{\phi}) \quad &= \sum_j \left\|\mathbf{a}_j \odot \boldsymbol{\phi}_j - \mathbf{W}\mathbf{W}^{+}(\mathbf{a}_j \odot \boldsymbol{\phi}_j)\right\|_2^2 \\ &\quad + \lambda \left\|\mathbf{y} - \sum_j \mathbf{a}_j \odot \boldsymbol{\phi}_j\right\|_2^2.\end{aligned} \tag{11}$$

Here $\mathbf{y}$ is a vector of the complex spectrogram of the mixture signal. The first term in $\mathcal{I}(\boldsymbol{\phi})$ measures how exactly $\mathbf{a}_j \odot \boldsymbol{\phi}_j$ satisfies the constraint from the redundancy, which corresponds to (9) and the second term represents the error between $\mathbf{y}$ and $\sum_j \mathbf{a}_j \odot \boldsymbol{\phi}_j$, which plays a similar role as (10), but transferred into time-frequency domain. The motivation is that in the T-F domain, we can handle each T-F bin separately. While (11) has the same consistency constraint with [5] where it is denoted as $\mathcal{F}$, the second term of (11) is different with [5] where they calculate the quadratic loss between the current estimation and the classical wiener filter output.

Instead of equally distributing the error into each source, here we introduce weight $\boldsymbol{\beta}$ for each T-F bin:

$$\sum_j \beta_{j,f,n} = 1, \quad 0 < \beta_{j,f,n} < 1. \tag{12}$$

To solve this optimization problem, we also employ majorization-minimization as in Section (2). We first find an auxiliary function of the objective and solve it iteratively.

First let $y_{f,n}$ be

$$y_{f,n} = \sum_j x_{j,f,n}, \tag{13}$$

then from Jensen's inequality, we can show that

$$\begin{aligned}&\left\|\mathbf{y} - \sum_j \mathbf{a}_j \odot \boldsymbol{\phi}_j\right\|_2^2 \\ &= \sum_{f,n} \left|y_{f,n} - \sum_j a_{j,f,n}\phi_{j,f,n}\right|^2 \\ &= \sum_{f,n} \left|\sum_j \frac{\beta_{j,f,n}}{\beta_{j,f,n}}\left(x_{j,f,n} - a_{j,f,n}\phi_{j,f,n}\right)\right|^2 \\ &\le \sum_{f,n}\sum_j \beta_{j,f,n}\left|\frac{1}{\beta_{j,f,n}}\left(x_{j,f,n} - a_{j,f,n}\phi_{j,f,n}\right)\right|^2 \\ &= \sum_{f,n,j} \frac{1}{\beta_{j,f,n}}\left|x_{j,f,n} - a_{j,f,n}\phi_{j,f,n}\right|^2,\end{aligned} \tag{14}$$

Here, the equality in (14) holds when

$$x_{j,f,n} = a_{j,f,n}\phi_{j,f,n} + \beta_{j,f,n}\left(y_{f,n} - \sum_{j'} a_{j',f,n}\phi_{j',f,n}\right). \tag{15}$$

In the same way as in Section 2, we have

$$\begin{aligned}&\left\|\mathbf{a}_j \odot \boldsymbol{\phi}_j - \mathbf{W}\mathbf{W}^{+}(\mathbf{a}_j \odot \boldsymbol{\phi}_j)\right\|_2^2 \\ &\le \|\mathbf{a}_j \odot \boldsymbol{\phi}_j - \mathbf{W}\tilde{\mathbf{c}}_j\|_2^2 = \sum_{f,n}\left|a_{j,f,n}\phi_{j,f,n} - c_{j,f,n}\right|^2,\end{aligned} \tag{16}$$

where $c_{j,f,n} = \mathbf{w}_{f,n}^{\mathsf{H}} \tilde{\mathbf{c}}_j$. Here the equality in (16) holds when

$$\tilde{\mathbf{c}}_j = \mathbf{W}^{+}(\mathbf{a}_j \odot \boldsymbol{\phi}_j). \tag{17}$$

From (14) and (16), we can use

$$\begin{aligned}\mathcal{J}(\boldsymbol{\Phi}, \mathbf{x}, \tilde{\mathbf{c}}) = &\sum_{j,f,n} |a_{j,f,n}\phi_{j,f,n} - c_{j,f,n}|^2 \\ &+ \sum_{j,f,n} \frac{\lambda}{\beta_{j,f,n}} |x_{j,f,n} - a_{j,f,n}\phi_{j,f,n}|^2\end{aligned} \tag{18}$$

as an auxiliary function for optimizing $\mathcal{I}(\boldsymbol{\phi})$. where d is other terms that are not related to the solution.

We have discussed how to update $\tilde{\mathbf{c}}_j$ and $\mathbf{x}_j$ respectively. Next we explore how to update the parameters $\boldsymbol{\phi}_j$, and $\boldsymbol{\beta}_j$.

First, for $\boldsymbol{\phi}_j$, since this term appears in both of the terms in (18), we merge those two terms together. Since the optimal value of $\phi_{j,f,n}$ can be found independently, we omit the subscript in order to keep the notation uncluttered. Then the objective function for each (j, f, n) is given by

$$|a\phi - c|^2 + \frac{\lambda}{\beta} |x - a\phi|^2 . \tag{19}$$

By expanding it to a second order polynomial of $a\phi$, (19) becomes

$$\begin{aligned}&\frac{(\beta+\lambda)}{\beta}\left[a^2|\phi|^2 - \frac{(\beta\bar{c}+\lambda\bar{x})a\phi + (\beta c+\lambda x)a\bar{\phi}}{(\beta+\lambda)}\right. \\ &\left.+\frac{(\beta|c|^2+\lambda|x|^2)}{(\beta+\lambda)}\right].\end{aligned} \tag{20}$$

Here $\bar{\cdot}$ denotes the complex conjugate. Then by completing the square, (20) becomes

$$\frac{(\beta+\lambda)}{\beta}\left|\frac{\beta c+\lambda x}{\beta+\lambda} - a\phi\right|^2 + d, \tag{21}$$

where d is other terms that does not related to the optimal solution of ϕ. (21) achieves its minimum value when both terms inside have the same phase which is $\phi = \angle\frac{\beta c+\lambda x}{\beta+\lambda}$. For $\boldsymbol{\beta}$, by minimizing (14) with constraint (12) when fixing $x_{j,f,t}$ and $\phi_{j,f,n}$, we have

$$\beta_{j,f,t} = \frac{|x_{j,f,t} - a_{j,f,t}\phi_{j,f,t}|}{\left|\sum_{j'} x_{j',f,t} - a_{j',f,t}\phi_{j',f,t}\right|}. \tag{22}$$

Thus we can iteratively update the parameters:

$$\tilde{\mathbf{c}}_j \leftarrow \mathbf{W}^{+}(\mathbf{a}_j \odot \boldsymbol{\phi}_j) \tag{23}$$

$$\mathbf{x}_j \leftarrow \mathbf{a}_j \odot \boldsymbol{\phi}_j + \boldsymbol{\beta}_j \odot \left(\mathbf{y} - \sum_{j'} \mathbf{a}_{j'} \odot \boldsymbol{\phi}_{j'}\right) \tag{24}$$

$$\boldsymbol{\phi}_j \leftarrow \angle\Big((\boldsymbol{\beta}_j \odot \mathbf{W}\tilde{\mathbf{c}}_j + \lambda\mathbf{x}_j) \oslash (\boldsymbol{\beta}_j + \lambda\mathbf{1})\Big) \tag{25}$$

$$\boldsymbol{\beta}_j \leftarrow \left(abs(\mathbf{x}_j - \mathbf{a}_j \odot \boldsymbol{\phi}_j) \oslash \sum_{j'} abs(\mathbf{x}_{j'} - \mathbf{a}_{j'} \odot \boldsymbol{\phi}_{j'})\right), \tag{26}$$

where $\mathbf{1}$ is an all-ones vector, $\oslash$ denotes the element-wise division and abs$(\cdot)$ denotes an operation that takes the absolute value of each element of a vector. As for the value of λ, it is set to be a hyperparameter to be optimized.

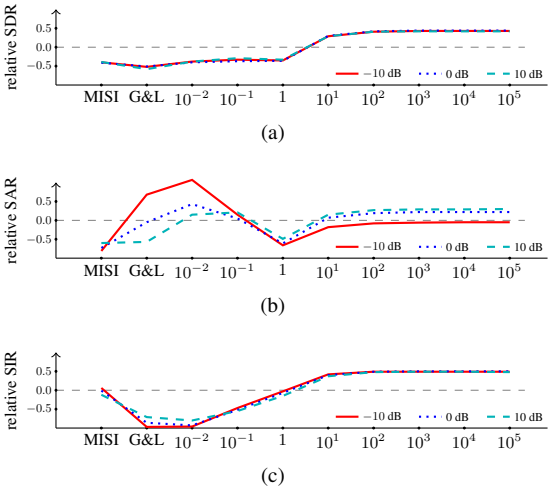

Figure 1: *Output SDR, SAR and SIR of different λ values in **Spectral Subtraction** case, averaged over all mixtures for each input SNR, subtracting the mean value for each input SNR. Numbers on the x-axis stand for different λ values in M-MISI.*

5. Experimental Evaluation

5.1. Setup

We evaluate the signal reconstruction performance of M-MISI in two different conditions, 1) *Binary Mask* and 2) *Spectrum Subtraction*. In both conditions, we assume that only imperfect magnitude spectra is known beforehand. In 1), The magnitude spectrum of each source is estimated by applying a binary mask to the mixture magnitude spectra. Here the ideal binary mask is computed with

$$\text{IBM}(t,f) = \begin{cases} 1 & \text{if SNR}(t,f) \geq \theta \\ 0 & \text{otherwise} \end{cases} \tag{27}$$

where t and f are the time and frequency indices respectively, the local SNR for each T-F bin is defined by

$$\text{SNR}(t,f) = 10\log_{10}\frac{|S(t,f)|^2}{|N(t,f)|^2}, \tag{28}$$

where S and N are the spectrogram of the clean speech and noise, respectively. θ is a local criterion(LC) in dB, which is chosen to be 0 dB. To add errors to the IBM values, we randomly flip the value with different rate between 0 and 20%, with 5% step. In 2), we only assume knowing the time average of power spectrum of the noise signal, and use power spectrum subtraction to recover the value of speech signal.

For both conditions, the mixtures are generated by mixing a speech signal taken from the ATR speech database [15], and a stationary noise from the noise database published by Japanese Standards Association (JSA) [16]. All of the conversational sentences in ATR (115 audio clips in total) are used for generating mixtures. The noise dataset contains 20 types of in-door noises including air conditioner, washing machine, and dryer. All utterances are re-sampled to 16kHz. The overlap ratio of each frame is set to 50%. The window size is set to 20ms.

For each condition, the Signal-to-Distortion, Interference and Artifacts ratio (SDR, SIR and SAR) are used as evaluation metrics. The performance of the proposed method is compared with MISI and Griffin & Lim algorithm.

Table 1: *Comparison of the SDR, SAR and SIR for the G&L[2] algorithm, MISI[3] and proposed Modified-MISI, with λ set to 10^3 for each input SNR, in 1)* ***Binary Mask*** *and 2)* ***Spectral Subtraction*** *conditions*

Input SNR		−10 dB			0 dB			+10 dB		
		SDR	SAR	SIR	SDR	SAR	SIR	SDR	SAR	SIR
Binary Mask	G&L[2]	5.74	**10.59**	9.88	13.62	**16.81**	17.67	**20.19**	**22.11**	25.00
	MISI[3]	5.76	10.11	10.18	13.40	16.07	17.87	19.03	20.36	25.10
	MMISI (Proposed)	**6.01**	10.27	**10.33**	**13.68**	16.46	**18.02**	19.71	21.27	**25.20**
Spectral Subtractioin	G&L[2]	−6.46	**5.28**	−4.91	3.83	9.98	5.55	13.25	17.96	15.26
	MISI[3]	−6.34	3.79	−3.88	3.93	9.30	6.40	13.44	17.93	15.85
	MMISI (Proposed)	**−5.52**	4.55	**−3.44**	**4.79**	**10.25**	**6.91**	**14.26**	**18.82**	**16.46**

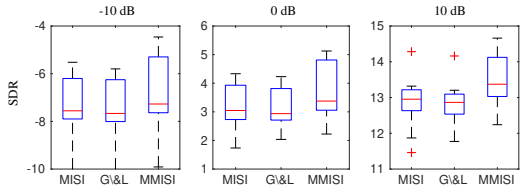

Figure 2: *The box plot showing the SDR performance of the Spectral Subtraction case*

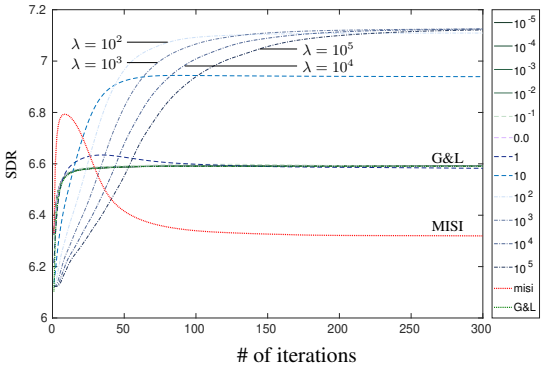

Figure 3: *Output SDR of MISI, G&L($\lambda = 0$), and M-MISI with different λ, as a function of number of iterations.*

5.2. Experimental Results

Figure 1 compares different λ values in terms of SDR, SAR and SIR, relative to the average value of each input SNR. Note that the G&L algorithm corresponds to the case where $\lambda = 0$. It is shown for each input SNR, the proposed method outperforms MISI when $\lambda > 10$. While the proposed method achieves similar performances when $\lambda > 10^2$, we find that the convergence becomes slower when λ becomes greater, which could be further confirmed from Figure 3. In other experiments, we choose $\lambda = 10^3$ for a best trade off between execution time and performance. Another observation from Figure 3 is that it is difficult to stop MISI when it achieves the best performance, since the performance degrades drastically after it reaches the peak at the very beginning of the execution, while M-MISI is assured to improve the SDR by running for more iterations, at a cost of more computational time.

Table 1 shows the comparison between M-MISI, MISI, and Griffin & Lim algorithm. For each experiment, the phase spectra of each source is computed by running 200 iterations of all of the 3 algorithms. In condition 1), the same experiment is conducted with different error rate, as mentioned in Section 5.1, and the upper rows show the average values of all experiments. Since G&L algorithm does not involve error distribution, it achieves the best SAR values in most of the times. For condition 2), we show further the statistics in Figure 2 for each input SNR. While in both conditions, the separation quality is dominated by the quality of magnitude estimations, our proposed method still achieved 0.4 dB in condition 1) and 1 dB in condition 2) compared with MISI.

6. Conclusion

We have proposed a new phase recovery algorithm, M-MISI, which employs a soft objective function. This algorithm is robust against imperfect magnitude spectra of the estimated sources. In our experiment, the proposed method achieved 1 dB improvement when we only use spectral subtraction to recover the magnitude spectrograms, and 0.4 dB on average when we use binary mask as the magnitude estimator. In future, we need to combine our method with recent source separation methods involving deep learning. Also, we would like to tackle the case when the number of recording channels (microphones) is more than one.

7. Appendix

7.1. The derivation of MISI

To confirm that (9) and (10) lead to the solution given in (7) and (8), we first define the Lagrangian as

$$\mathcal{L}(\boldsymbol{\phi}, \mathbf{c}, \gamma) = \mathcal{J}(\boldsymbol{\phi}, \tilde{\mathbf{c}}) + \gamma^{\mathsf{H}} \left(\sum_j \tilde{\mathbf{c}}_j - \tilde{\mathbf{y}} \right).$$

By setting the partial derivative of the Lagrangian with respect to $\tilde{\mathbf{c}}_j^*$ at zero, we obtain

$$\mathbf{W}^{\mathsf{H}} \mathbf{W} \tilde{\mathbf{c}}_j = \mathbf{W}^{\mathsf{H}} (\mathbf{a}_j \odot \boldsymbol{\phi}_j) - \gamma.$$

By summing this expression over j, we get

$$J\gamma = \sum_j \mathbf{W}^{\mathsf{H}} (\mathbf{a}_j \odot \boldsymbol{\phi}_j) - \mathbf{W}^{\mathsf{H}} \mathbf{W} \tilde{\mathbf{y}}.$$

Hence, when $\{\boldsymbol{\phi}_j\}_j$ is fixed, (9) is minimized with respect to $\tilde{\mathbf{c}}_j$ subject to (10) when

$$\tilde{\mathbf{c}}_j = \mathbf{W}^{+} (\mathbf{a}_j \odot \boldsymbol{\phi}_j) + \frac{1}{J} \left(\tilde{\mathbf{y}} - \sum_j \mathbf{W}^{+} (\mathbf{a}_j \odot \boldsymbol{\phi}_j) \right).$$

It is straightforward to show that when $\{\tilde{\mathbf{c}}_j\}$ is fixed, (9) is minimized with respect to $\boldsymbol{\phi}_j$ when

$$\boldsymbol{\phi}_j = \angle \mathbf{W} \tilde{\mathbf{c}}_j.$$

8. References

[1] R. C. Hendriks, T. Gerkmann, and J. Jensen, "Dft-domain based single-microphone noise reduction for speech enhancement: A survey of the state of the art," *Synthesis Lectures on Speech and Audio Processing*, vol. 9, no. 1, pp. 1–80, 2013.

[2] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, April 1984.

[3] 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. 421–424, May 2010.

[4] N. Sturmel and L. Daudet, "Informed Source Separation using Iterative Reconstruction," *arXiv:1202.2075 [cs]*, Feb. 2012, arXiv: 1202.2075. [Online]. Available: http://arxiv.org/abs/1202.2075

[5] J. L. Roux and E. Vincent, "Consistent Wiener Filtering for Audio Source Separation," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 217–220, Mar. 2013.

[6] Z.-Q. Wang, J. L. Roux, D. Wang, and J. R. Hershey, "End-to-end speech separation with unfolded iterative phase reconstruction," *arXiv preprint arXiv:1804.10204*, 2018.

[7] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[8] M. Davies and C. James, "Source separation using single channel ica," *Signal Processing*, vol. 87, no. 8, pp. 1819 – 1832, 2007, independent Component Analysis and Blind Source Separation. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0165168407000151

[9] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *Interspeech*, sep 2006. [Online]. Available: http://www2.imm.dtu.dk/pubdb/p.php?4511

[10] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE transactions on audio, speech, and language processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[11] J. R. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 31–35.

[12] Y. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," *arXiv preprint arXiv:1607.02173*, 2016.

[13] M. Kolbæk, D. Yu, Z.-H. Tan, J. Jensen, M. Kolbaek, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 10, pp. 1901–1913, 2017.

[14] D. R. Hunter and K. Lange, "A tutorial on mm algorithms," *The American Statistician*, vol. 58, no. 1, pp. 30–37, 2004.

[15] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "Atr japanese speech database as a tool of speech recognition and synthesis," *Speech communication*, vol. 9, no. 4, pp. 357–363, 1990.

[16] J. S. Association, "Design guidelines for alarm sound of consumer products-environmental sound database," *JIS/TR S 0001*, 2002. [Online]. Available: https://ci.nii.ac.jp/naid/10011785283/