MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

End-to-End Speech Separation with Unfolded Iterative Phase Reconstruction

Wang, Zhong-Qiu; Le Roux, Jonathan; Hershey, John

TR2018-051 July 10, 2018

Abstract

This paper proposes an end-to-end approach for single-channel speaker-independent multi-speaker speech separation, where time-frequency (T-F) masking, the short-time Fourier transform (STFT), and its inverse are represented as layers within a deep network. Previous approaches, rather than computing a loss on the reconstructed signal, used a surrogate loss based on the target STFT magnitudes. This ignores reconstruction error introduced by phase inconsistency. In our approach, the loss function is directly defined on the reconstructed signals, which are optimized for best separation. In addition, we train through unfolded iterations of a phase reconstruction algorithm, represented as a series of STFT and inverse STFT layers. While mask values are typically limited to lie between zero and one for approaches using the mixture phase for reconstruction, this limitation is less relevant if the estimated magnitudes are to be used together with phase reconstruction. We thus propose several novel activation functions for the output layer of the T-F masking, to allow mask values beyond one. On the publiclyavailable wsj0-2mix dataset, our approach achieves state-ofthe-art 12.6 dB scale-invariant signal-to-distortion ratio (SISDR) and 13.1 dB SDR, revealing new possibilities for deep learning based phase reconstruction and representing a fundamental progress towards solving the notoriously-hard cocktail party problem.

arXiv

End-to-End Speech Separation with Unfolded Iterative Phase Reconstruction

Zhong-Qiu Wang[1,2]*, Jonathan Le Roux*[1]*, DeLiang Wang*[2,3]*, John R. Hershey*[1]

[1]Mitsubishi Electric Research Laboratories (MERL), USA
[2]Department of Computer Science and Engineering, The Ohio State University, USA
[3]Center for Cognitive and Brain Sciences, The Ohio State University, USA

{wangzhon,dwang}@cse.ohio-state.edu, leroux@merl.com

Abstract

This paper proposes an end-to-end approach for single-channel speaker-independent multi-speaker speech separation, where time-frequency (T-F) masking, the short-time Fourier transform (STFT), and its inverse are represented as layers within a deep network. Previous approaches, rather than computing a loss on the reconstructed signal, used a surrogate loss based on the target STFT magnitudes. This ignores reconstruction error introduced by phase inconsistency. In our approach, the loss function is directly defined on the reconstructed signals, which are optimized for best separation. In addition, we train through unfolded iterations of a phase reconstruction algorithm, represented as a series of STFT and inverse STFT layers. While mask values are typically limited to lie between zero and one for approaches using the mixture phase for reconstruction, this limitation is less relevant if the estimated magnitudes are to be used together with phase reconstruction. We thus propose several novel activation functions for the output layer of the T-F masking, to allow mask values beyond one. On the publicly-available wsj0-2mix dataset, our approach achieves state-of-the-art 12.6 dB scale-invariant signal-to-distortion ratio (SI-SDR) and 13.1 dB SDR, revealing new possibilities for deep learning based phase reconstruction and representing a fundamental progress towards solving the notoriously-hard cocktail party problem.

Index Terms: deep clustering, chimera++ network, iterative phase reconstruction, cocktail party problem.

1. Introduction

Recent years have witnessed exciting advances towards solving the cocktail party problem. The inventions of deep clustering [1, 2, 3], deep attractor networks [4, 5] and permutation free training [1, 2, 6, 7] have dramatically improved the performance of single-channel speaker-independent multi-speaker speech separation, demonstrating overwhelming advantages over previous methods including graphical modeling approaches [8], spectral clustering approaches [9], and CASA methods [10].

However, all of these conduct separation on the magnitude in the time-frequency (T-F) domain and directly use the mixture phase for time-domain re-synthesis, largely because phase is difficult to estimate. It is well-known that this incurs a phase inconsistency problem [11, 12, 13], especially for speech processing, where there is typically at least half overlap between consecutive frames. This overlap makes the STFT representation of a speech signal highly redundant. As a result, the enhanced STFT representation obtained using the estimated magnitude and mixture phase would not be in the consistent STFT domain, meaning that it is not guaranteed that there exists a time-domain signal having that STFT representation.

To improve the consistency, one stream of research is focused on iterative methods such as the classic Griffin-Lim algorithm [14], multiple input spectrogram inverse (MISI) [15], ISSIR [16], and consistent Wiener filtering [17], which can recover the clean phase to some extent starting from the mixture phase and a good estimated magnitude by iteratively performing STFT and iSTFT [13]. There are some previous attempts at naively applying such iterative algorithms as a post-processing step on the magnitudes produced by deep learning based speech enhancement and separation [18, 19, 20, 3]. However, this usually only leads to small improvements, even though the magnitude estimates from DNNs are reasonably good. We think that this is possibly because the T-F masking is performed without being aware of the later phase reconstruction steps and hence may not produce spectral structures that are appropriate for iterative phase reconstruction.

This study hence proposes a novel end-to-end speech separation algorithm that trains through iterative phase reconstruction via T-F masking for signal-level approximation. On the publicly-available wsj0-2mix corpus, our algorithm reaches 12.6 dB scale-invariant SDR, which surpasses the previous best by a large margin and is comparable to the oracle 12.7 dB result obtained using the so-called ideal ratio mask (IRM). Our study shows, for the first time and based on a large open dataset, that deep learning based phase reconstruction leads to tangible and large improvements when combined with state-of-the-art magnitude-domain separation.

2. Chimera++ Network

To elicit a good phase via phase reconstruction, it is necessary to first obtain a good enough magnitude estimate. Our recent study [3] proposed a novel multi-task learning approach combining the regularization capability of deep clustering with the ease of end-to-end training of mask inference, yielding significant improvements over the individual models.

The key idea of deep clustering [1] is to learn a high-dimensional embedding vector for each T-F unit using a powerful deep neural network (DNN) such that the embeddings of the T-F units dominated by the same speaker are close to each other in the embedding space while farther otherwise. This way, simple clustering methods like k-means can be applied to the learned embeddings to perform separation at run time. More specifically, the network computes a unit-length embedding vector $v_i \in \mathbb{R}^{1\times D}$ corresponding to the i^{th} T-F element. Similarly, $y_i \in \mathbb{R}^{1\times C}$ is a one-hot label vector representing which source in a mixture dominates the i^{th} T-F unit. Vertically stacking these, we form the embedding matrix $V \in \mathbb{R}^{TF\times D}$ and the label matrix $Y \in \mathbb{R}^{TF\times C}$. The embeddings are learned

Part of this work was done while Z.-Q. Wang was an intern at MERL.

by approximating the affinity matrix from the embeddings:

$$\mathcal{L}_{\text{DC,classic}} = \|VV^{\mathsf{T}} - YY^{\mathsf{T}}\|_{\text{F}}^2 \quad (1)$$

Our recent study [3] suggests that an alternative loss function, which whitens the embedding in a k-means objective, leads to better separation performance.

$$\begin{aligned}\mathcal{L}_{\text{DC,W}} &= \|V(V^{\mathsf{T}}V)^{-\frac{1}{2}} - Y(Y^{\mathsf{T}}Y)^{-1}Y^{\mathsf{T}}V(V^{\mathsf{T}}V)^{-\frac{1}{2}}\|_{\text{F}}^2 \\ &= D - \text{tr}\big((V^{\mathsf{T}}V)^{-1}V^{\mathsf{T}}Y(Y^{\mathsf{T}}Y)^{-1}Y^{\mathsf{T}}V\big) \quad (2)\end{aligned}$$

To learn the embeddings, bi-directional LSTM (BLSTM) is usually used to model the context information from past and future frames. The network architecture is shown at the bottom of Fig. 1, where the DC embedding layer is a fully-connected layer with a non-linearity such as a logistic sigmoid, followed by unit-length normalization for each frequency.

Another permutation-free training scheme was proposed for mask-inference networks first in [1], and was later found to be working very well in [2] and [6]. The idea is to train a mask-inference network to minimize the minimum loss over all permutations. Following [7], the phase-sensitive mask (PSM) [21] is used as the training target. It is common in phase-sensitive spectrum approximation (PSA) to truncate the unbounded mask values. Using $\text{T}_a^b(x) = \min(\max(x, a), b)$, the truncated PSA (tPSA) objective is

$$\begin{aligned}\mathcal{L}_{\text{tPSA}} = \min_{\pi\in\mathcal{P}} \sum_c \Big\| &\hat{M}_{\pi(c)} \odot |X| \\ &- \text{T}_0^{\gamma|X|}\left(|S_c| \odot \cos(\angle S_c - \angle X)\right)\Big\|_1, \quad (3)\end{aligned}$$

where $\angle X$ is the mixture phase, $\angle S_c$ the phase of the c-th source, $\mathcal{P}$ the set of permutations on $\{1, \ldots, C\}$, $|X|$ the mixture magnitude, $\hat{M}_c$ the c-th estimated mask, $|S_c|$ the magnitude of the c-th reference source, $\odot$ denotes element-wise matrix multiplication, and γ is a mask truncation factor. Sigmoidal activation together with $\gamma = 1$ is commonly used in the output layer of T-F masking. To endow the network with more capability, multiple activation functions that can work with $\gamma > 1$ will be discussed in Section 3.4.

Following [22], our recent study [3] proposed a chimera++ network combining the two approaches via multi-task learning, as illustrated in the bottom of Fig. 1. The loss function is a weighted sum of the deep clustering loss and the mask inference loss.

$$\mathcal{L}_{\text{chi}_\alpha^{++}} = \alpha\mathcal{L}_{\text{DC,W}} + (1-\alpha)\mathcal{L}_{\text{tPSA}} \quad (4)$$

Only the MI output is needed to make predictions at run time.

3. Proposed Algorithms

3.1. Iterative Phase Reconstruction

There are multiple target sources to be separated in each mixture in our study. The Griffin-Lim algorithm [14] only performs iterative reconstruction for each source independently. In [3], we therefore proposed to utilize the MISI algorithm [15] (see Algorithm 1) to reconstruct the clean phase of each source starting from the estimated magnitude of each source and the mixture phase, where the sum of the reconstructed time-domain signals after each iteration is constrained to be the same as the mixture signal. Note that the estimated magnitudes remain fixed during iterations, while the phase of each source are iteratively reconstructed. In [3], the phase reconstruction was only added as a post-processing, and it was not part of the objective function during training, which remained computed on the time-frequency representation of the estimated signal, prior to resynthesis. In this paper, we go several steps further.

Input : Mixture time-domain signal x, mixture complex spectrogram X, mixture phase $\angle X$, enhanced magnitudes $\hat{A}_c = \hat{M}_c \circ |X|$ for $c = 1, \ldots, C$, and iteration number K
Output : Reconstructed phase $\hat{\theta}_c^{(K)}$ and signal $\hat{s}_c^{(K)}$ for $c = 1, \ldots, C$

$\hat{s}_c^{(0)} = \text{iSTFT}(\hat{A}_c, \angle X)$, for $c = 1, \ldots, C$;
for $i = 1, \ldots, K$ **do**
  $\delta^{(i-1)} = x - \sum_{c=1}^{C} \hat{s}_c^{(i-1)}$;
  $\hat{\theta}_c^{(i)} = \angle\text{STFT}\big(\hat{s}_c^{(i-1)} + \frac{\delta^{(i-1)}}{C}\big)$, for $c = 1, \ldots, C$;
  $\hat{s}_c^{(i)} = \text{iSTFT}(\hat{A}_c, \hat{\theta}_c^{(i)})$, for $c = 1, \ldots, C$;
end

Algorithm 1: Iterative phase reconstruction based on MISI. STFT($\cdot$) extracts the STFT magnitude and phase of a signal, and iSTFT($\cdot$,$\cdot$) reconstructs a time-domain signal from a magnitude and a phase.

3.2. Waveform Approximation

The first step in phase reconstruction algorithms such as MISI is to reconstruct a waveform from a time-frequency domain representation using the inverse STFT. We thus consider a first objective function computed on the waveform reconstructed by iSTFT, denoted as waveform approximation (WA), and represent iSTFT as various layers on top of the mask inference layer, so that end-to-end optimization can be performed. The label permutation problem is resolved by minimizing the minimum L_1 loss of all the permutations at the waveform level. We denote the model trained this way as WA. The objective function to train this model is

$$\mathcal{L}_{\text{WA}} = \min_{\pi\in\mathcal{P}} \sum_c \Big\|\hat{s}_{\pi(c)}^{(0)} - s_c\Big\|_1, \quad (5)$$

where s_c denotes the time-domain signal of source c, and $\hat{s}_c^{(0)}$ denotes the c-th time-domain signal obtained by inverse STFT from the combination of the c-th estimated magnitude and the mixture phase. Note that mixture phase is still used here and no phase reconstruction is yet performed. This corresponds to the initialization step in Algorithm 1.

In [23], a time-domain reconstruction approach is proposed for speech enhancement. However, their approach only trains a feed-forward mask-inference DNN through iDFT separately for each frame using squared error in the time domain. By Parseval's theorem, this is equivalent to optimizing the mask for minimum squared error in the complex spectrum domain, when using the noisy phases, as in [21], proposed in the same conference. A follow-up work [19] of [23] supplies clean phase during training. However, this makes their approach equivalent to conventional magnitude spectrum approximation [24], which does not perform as well as the phase-sensitive mask [25]. Closest to the above WA objective, an adaptive front-end framework was recently proposed [26] in which the STFT and its inverse are subsumed by the network, along with the noisy phase, so that training is effectively end-to-end in the time-domain. The proposed method then replaces the STFT and its inverse by trainable linear convolutional layers. Unfortunately the paper does not compare training through the STFT to the conventional method so the results are uninformative about this direction.

3.3. Unfolded Iterative Phase Reconstruction

We further unfold the iterations in the MISI algorithm as various deterministic layers in a neural network. This can be achieved

by further growing several layers representing STFT and iSTFT operations on top of the mask inference layer. By performing end-to-end optimization that trains through MISI, the network can become aware of the later iterative phase reconstruction steps and learn to produce estimated magnitudes that are well-suited to that subsequent processing, hence producing better phase estimates for separation. The model trained this way is denoted as WA-MISI-K, where $K \geq 1$ is the number of unfolded MISI iterations. The objective function is

$$\mathcal{L}_{\text{WA-MISI-K}} = \min_{\pi \in \mathcal{P}} \sum_{c} \left\| \hat{s}^{(K)}_{\pi(c)} - s_c \right\|_1, \tag{6}$$

where $\hat{s}^{(K)}_c$ denotes the c-th time-domain signal obtained after K MISI iterations as described in Algorithm 1. The whole separation network, including unfolded phase reconstruction steps at the output of the mask inference head of the Chimera++ network, is illustrated in Fig. 1. The STFT and iSTFT can be easily implemented using modern deep learning toolkits as deterministic layers efficiently computed on a GPU and through which backpropagation can be performed.

A recent study by Williamson et al. [27, 28] proposed a complex ratio masking approach for phase reconstruction and speech enhancement, where a feed-forward DNN is trained to predict the real and imaginary components of the ideal complex filter in the STFT domain, i.e., $M_c = S_c/X = |S_c|e^{j(\angle S_c - \angle X)}/|X|$ for source c for example. The real component is equivalent to the earlier proposed phase-sensitive mask [21], which contains patterns clearly predictable from energy-based features [21, 25]. However, recent studies along this line suggest that the patterns in the imaginary component are too random to predict [29], possibly because it is difficult for a learning machine to determine the sign of $\sin(\angle S_c - \angle X)$ only from energy-based features. In contrast, the $\cos(\angle S_c - \angle X)$ in the real component is typically much smaller than one for T-F units dominated by other sources and close to one otherwise, making itself predictable from energy-based features. The proposed method thus only focuses on estimating a mask in the magnitude domain and uses the estimated magnitude to elicit better phase through iterative phase reconstruction.

Another recent trend is to avoid the phase inconsistency problem altogether by operating in the time domain, using convolutional neural networks [30, 31], WaveNet [32], generative adversarial networks [33], or encoder-decoder architectures [34]. Although they are promising approaches, the current state-of-the-art approach for supervised speech separation is via T-F masking [35, 3]. The proposed approach is expected to produce even better separation if the phase can be reconstructed.

3.4. Activation Functions with Values Beyond One

Sigmoidal units are dominantly used in the output layer of deep learning based T-F masking [36, 35], partly because they can model well data with bi-modal distribution [37], such as the IRM [38] and its variants [36]. Restricting the possible values of the T-F mask to lie in $[0, 1]$ is also reasonable when using the mixture phase for reconstruction: indeed, T-F mask values larger than one would in theory be needed in regions where interferences between sources result in a mixture magnitude smaller than that of a source; but the mixture phase is also likely to be different from the phase of that source in such regions, in which case it is more rewarding in terms of objective measure to oversuppress than to go even further in a wrong direction. This is no longer valid if we consider phase reconstruction in the optimization. Moreover, capping the mask values to

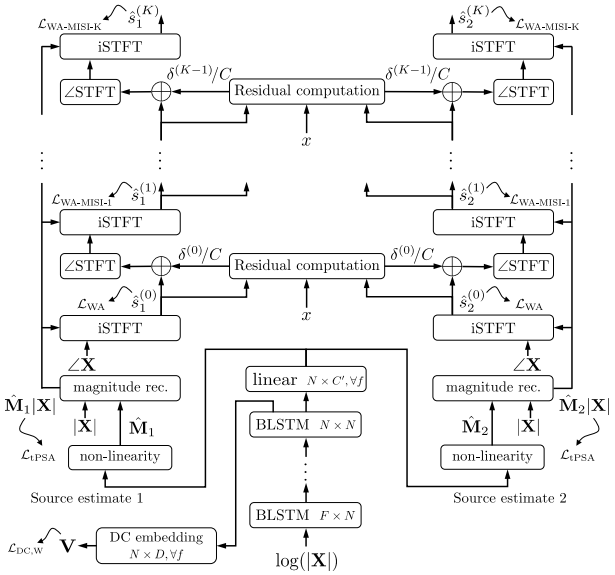

Figure 1: *Training through K MISI iterations.*

be between zero and one is more likely to take the enhanced magnitude further away from the consistent STFT domain, posing potential difficulties for later phase reconstruction.

To obtain clean magnitudes, the oracle mask should be $|S_c|/|X|$ (also known as the FFT mask in [38] or the ideal amplitude mask in [21]). Clearly, this mask can go beyond one, because the underlying sources, although statistically independent, may have opposite phase at a particular T-F unit, therefore cancelling with each other and producing a mixture magnitude that is smaller than the magnitude of a given source. It is likely much harder to predict the mask values of such T-F units, but we believe that it is still possible based on contextual information.

In our study, we truncate the values in PSM to the range $[0, 2]$ (i.e., $\gamma = 2$ in Eq. (3)), as only a small percentage of mask values goes beyond this range. Multiple activation functions can be utilized in the output layer. We here consider:

- doubled sigmoid: sigmoid non-linearity multiplied by 2;
- clipped ReLU: ReLU non-linearity clipped to $[0, 2]$;
- convex softmax: the output non-linearity is a three-dimensional softmax for each source at each T-F unit. It is used to compute a convex sum between the values 0, 1, and 2: $y = [x_0, x_1, x_2][0, 1, 2]^T$ where $[x_0, x_1, x_2]$ is the output of the softmax. This activation function is designed to model the three modes concentrated at 0, 1 and 2 in the histogram of the PSM.

4. Experimental Setup

We validate the proposed algorithms on the publicly-available wsj0-2mix corpus [1], which is widely used in many speaker-independent speech separation tasks. It contains 20,000, 5,000 and 3,000 two-speaker mixtures in its 30 h training, 10 h validation, and 5 h test sets, respectively. The speakers in the validation set (closed speaker condition, CSC) are seen during training, while the speakers in the test set (open speaker condition, OSC) are completely unseen. The sampling rate is 8 kHz.

Our neural network contains four BLSTM layers, each with 600 units in each direction. A dropout of 0.3 is applied on the output of each BLSTM layer except the last one. The network is trained on 400-frame segments using the Adam algorithm. The

Table 1: *SI-SDR (dB) performance on wsj0-2mix.*

Approaches	CSC	OSC
$\mathcal{L}_{\text{DC,W}}$	10.4	10.4
$\mathcal{L}_{\text{tPSA}}$	10.1	10.0
$\mathcal{L}_{\text{chi}_\alpha^{++}}$ (sigmoid)	11.1	11.2
+ Griffin-Lim-5	11.2	11.3
+ MISI-5	11.4	11.5
+ $\mathcal{L}_{\text{WA}}$	11.6	11.6
+MISI-5	11.6	11.6
+$\mathcal{L}_{\text{WA-MISI-5}}$	12.4	12.2
$\mathcal{L}_{\text{chi}_\alpha^{++}}$ (doubled sigmoid)	10.0	10.0
+$\mathcal{L}_{\text{WA}}$	11.5	11.4
+$\mathcal{L}_{\text{WA-MISI-5}}$	12.5	12.3
$\mathcal{L}_{\text{chi}_\alpha^{++}}$(clipped RelU)	10.4	10.4
+$\mathcal{L}_{\text{WA}}$	11.7	11.7
+$\mathcal{L}_{\text{WA-MISI-5}}$	12.6	12.4
$\mathcal{L}_{\text{chi}_\alpha^{++}}$(convex softmax)	11.0	11.1
+$\mathcal{L}_{\text{WA}}$	11.8	11.8
+$\mathcal{L}_{\text{WA-MISI-5}}$	**12.8**	**12.6**

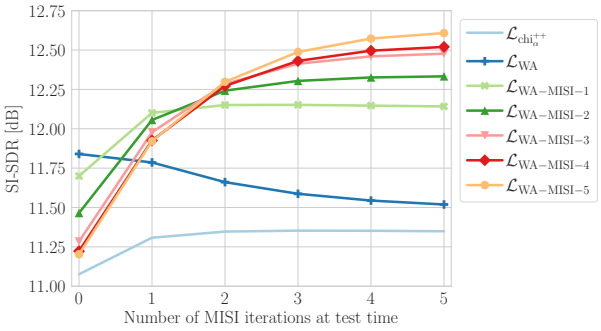

Figure 2: *SI-SDR vs number of MISI iterations at test time*

window length is 32 ms and the hop size is 8 ms. The square root Hann window is employed as the analysis window and the synthesis window is designed accordingly to achieve perfect reconstruction after overlap-add. A 256-point DFT is performed to extract 129-dimensional log magnitude input features. We first train the chimera++ network with α set to 0.975. Next, we discard the deep clustering branch (i.e., we set α to 0) and train the network with $\mathcal{L}_{\text{WA}}$. Subsequently, the network is trained using $\mathcal{L}_{\text{WA-MISI-1}}$, then $\mathcal{L}_{\text{WA-MISI-2}}$, and all the way to $\mathcal{L}_{\text{WA-MISI-K}}$, where here $K = 5$, as performance saturated after five iterations in our experiments. We found this curriculum learning strategy to be helpful. At run time, for the models trained using $\mathcal{L}_{\text{WA-MISI-5}}$, we run MISI with 5 iterations, while results for other models are obtained without phase reconstruction unless specified.

We report the performance using *scale-invariant SDR* (SI-SDR) [1, 2, 5, 39], as well as the SDR metric computed using the bss_eval_sources software [40] because it is used by other groups. We believe SI-SDR is a more proper measure for single-channel instantaneous mixtures [39].

5. Evaluation Results

Table 1 reports the SI-SDR results on the wsj0-2mix dataset. We first present the results using sigmoidal activation. The chimera++ network obtains significantly better results than the individual models (11.2 dB vs. 10.4 dB and 10.0 dB SI-SDR). With the mixture phase and estimated magnitudes, performing five iterations of MISI pushes the performance to 11.5 dB, while 11.3 dB is obtained when applying five iterations of Griffin-Lim on each source independently, as is reported in [3]. Performing end-to-end optimization using $\mathcal{L}_{\text{WA}}$ improves the results to 11.6 dB from 11.2 dB, without requiring phase reconstruction post-processing. Further applying MISI post-processing for five iterations (MISI-5) on this model however does not lead to any improvements, likely because the mixture phase is used during training and the model compensates for it without expecting further processing. In contrast, training the network through MISI using $\mathcal{L}_{\text{WA-MISI-5}}$ pushes the performance to 12.2 dB.

Table 2: *Comparison with other systems on wsj0-2mix.*

Approaches	SI-SDR (dB)		SDR (dB)	
	CSC	OSC	CSC	OSC
Deep Clustering [1, 2]	-	10.8	-	-
Deep Attractor Networks [4, 5]	-	10.4	-	10.8
PIT [6, 7]	-	-	10.0	10.0
TasNet [34]	-	10.2	-	10.5
Chimera++ Networks [3]	11.1	11.2	11.6	11.7
+ MISI-5 [3]	11.4	11.5	12.0	12.0
WA (proposed)	11.8	11.8	12.3	12.3
WA-MISI-5 (proposed)	**12.8**	**12.6**	**13.2**	**13.1**
Oracle Masks:				
Magnitude Ratio Mask	12.5	12.7	13.0	13.2
+ MISI-5	13.5	13.7	14.1	14.3
Ideal Binary Mask	13.2	13.5	13.7	14.0
+ MISI-5	13.1	13.4	13.6	13.8
PSM	16.2	16.4	16.7	16.9
+ MISI-5	18.1	18.3	18.5	18.8
Ideal Amplitude Mask	12.6	12.8	12.9	13.2
+ MISI-5	26.3	26.6	26.8	27.1

Among the three proposed activation functions, the convex softmax performs the best, reaching 12.6 dB SI-SDR. It thus seems effective to model the multiple peaks in the histogram of the truncated PSM, and important to produce estimated magnitudes that are closer to the consistent STFT domain. As expected, activations going beyond 1 only become beneficial when training through phase reconstruction.

In Fig. 2, we show the evolution of the SI-SDR performance of the convex softmax models trained with different objective functions against the number of MISI iterations at test time (0 to 5). Training with $\mathcal{L}_{\text{WA}}$ leads to a magnitude that is very well suited to iSTFT, but not to further MISI iterations. As we train for more MISI iterations, performance starts lower, but reaches higher values with more test-time iterations.

Table 2 lists the performance of competitive approaches on the same corpus, along with the performance of various oracle masks with or without applying MISI for five iterations. The first three algorithms use mixture phase directly for separation. The fourth one, time-domain audio separation network (TasNet), operates directly in the time domain. Our result is 1.1 dB better than the previous state-of-the-art by [3] in terms of both SI-SDR and SDR.

6. Concluding Remarks

We have proposed a novel end-to-end approach for single-channel speech separation. Significant improvements are obtained by training the T-F masking network through an iterative phase reconstruction procedure. Future work includes applying the proposed methods to speech enhancement, considering the joint estimation of magnitude and an initial phase that improves upon the mixture phase, and improving the estimation of the ideal amplitude mask. We shall also consider alternatives to the waveform-level loss, such as errors computed on the magnitude spectrograms of the reconstructed signals.

7. References

[1] J. R. Hershey, Z. Chen, and J. Le Roux, "Deep Clustering: Discriminative Embeddings for Segmentation and Separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[2] Y. Isik, J. Le Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-Channel Multi-Speaker Separation using Deep Clustering," in *Proc. Interspeech*, 2016.

[3] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Alternative Objective Functions for Deep Clustering," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[4] Z. Chen, Y. Luo, and N. Mesgarani, "Deep Attractor Network for Single-Microphone Speaker Separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[5] Y. Luo, Z. Chen, and N. Mesgarani, "Speaker-Independent Speech Separation with Deep Attractor Network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2018.

[6] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation Invariant Training of Deep Models for Speaker-Independent Multi-talker Speech Separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[7] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multi-Talker Speech Separation with Utterance-Level Permutation Invariant Training of Deep Recurrent Neural Networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 2017.

[8] J. R. Hershey, S. Rennie, P. A. Olsen, and T. T. Kristjansson, "Super-Human Multi-Talker Speech Recognition: A Graphical Modeling Approach," *Computer Speech & Language*, vol. 24, no. 1, 2010.

[9] F. Bach and M. Jordan, "Learning Spectral Clustering, with Application to Speech Separation," *The Journal of Machine Learning Research*, vol. 7, 2006.

[10] D. Wang and G. J. Brown, *Eds., Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Hoboken, NJ: Wiley-IEEE Press, Sep. 2006.

[11] J. Le Roux, N. Ono, and S. Sagayama, "Explicit consistency constraints for STFT spectrograms and their application to phase reconstruction," in *Proc. ISCA Workshop on Statistical and Perceptual Audition (SAPA)*, Sep. 2008.

[12] N. Sturmel and L. Daudet, "Signal Reconstruction from STFT Magnitude: A State of the Art," in *International Conference on Digital Audio Effects*, 2011.

[13] T. Gerkmann, M. Krawczyk-Becker, and J. Le Roux, "Phase Processing for Single-Channel Speech Enhancement: History and Recent Advances," *IEEE Signal Processing Magazine*, vol. 32, no. 2, 2015.

[14] D. W. Griffin and J. S. Lim, "Signal Estimation from Modified Short-Time Fourier Transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, 1984.

[15] D. Gunawan and D. Sen, "Iterative Phase Estimation for the Synthesis of Separated Sources from Single-Channel Mixtures," in *IEEE Signal Processing Letters*, 2010.

[16] N. Sturmel and L. Daudet, "Informed Source Separation using Iterative Reconstruction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 1, Jan. 2013.

[17] J. Le Roux and E. Vincent, "Consistent Wiener Filtering for Audio Source Separation," *IEEE Signal Processing Letters*, vol. 20, no. 3, Mar. 2013.

[18] K. Han, Y. Wang, D. Wang, W. S. Woods, and I. Merks, "Learning Spectral Mapping for Speech Dereverberation and Denoising," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 6, 2015.

[19] Y. Zhao, Z.-Q. Wang, and D. Wang, "A Two-stage Algorithm for Noisy and Reverberant Speech Enhancement," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[20] K. Li, B. Wu, and C.-H. Lee, "An Iterative Phase Recovery Framework with Phase Mask for Spectral Mapping with an Application to Speech Enhancement," in *Proc. Interspeech*, 2016.

[21] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-Sensitive and Recognition-Boosted Speech Separation using Deep Recurrent Neural Networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[22] Y. Luo, Z. Chen, J. R. Hershey, J. Le Roux, and N. Mesgarani, "Deep Clustering and Conventional Networks for Music Separation: Stronger Together," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[23] Y. Wang and D. Wang, "A Deep Neural Network for Time-Domain Signal Reconstruction," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[24] F. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller, "Discriminatively Trained Recurrent Neural Networks for Single-channel Speech Separation," in *Proc. IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2014.

[25] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech Enhancement with LSTM Recurrent Neural Networks and its Application to Noise-Robust ASR," in *Proc. International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, 2015.

[26] S. Venkataramani and P. Smaragdis, "End-to-end source separation with adaptive front-ends," in *arXiv preprint arXiv:1705.02514*, 2017.

[27] D. S. Williamson, Y. Wang, and D. Wang, "Complex Ratio Masking for Monaural Speech Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2016.

[28] D. S. Williamson and D. Wang, "Time-Frequency Masking in the Complex Domain for Speech Dereverberation and Denoising," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 25, no. 7, 2017.

[29] D. S. Williamson and D. Wang, "Speech Dereverberation and Denoising using Complex Ratio Masks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[30] S.-W. Fu, Y. Tsao, X. Lu, and H. Kawai, "End-to-End Waveform Utterance Enhancement for Direct Evaluation Metrics Optimization by Fully Convolutional Neural Networks," *arXiv preprint arXiv:1709.03658*, 2017.

[31] ——, "Raw Waveform-Based Speech Enhancement by Fully Convolutional Networks," in *arXiv preprint arXiv:1703.02205*, 2017.

[32] K. Qian, Y. Zhang, S. Chang, X. Yang, M. Hasegawa-Johnson, D. Florencio, and M. Hasegawa-Johnson, "Speech Enhancement using Bayesian WaveNet," in *Proc. Interspeech*, 2017.

[33] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech Enhancement Generative Adversarial Network," in *Proc. Interspeech*, 2017.

[34] Y. Luo and N. Mesgarani, "TasNet: Time-Domain Audio Separation Network for Real-Time, Single-Channel Speech Separation," in *arXiv preprint arXiv:1711.00541*, 2017.

[35] D. Wang and J. Chen, "Supervised Speech Separation Based on Deep Learning: An Overview," in *arXiv preprint arXiv:1708.07524*, 2017.

[36] Z.-Q. Wang and D. Wang, "Recurrent Deep Stacking Networks for Supervised Speech Separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[37] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[38] Y. Wang, A. Narayanan, and D. Wang, "On Training Targets for Supervised Speech Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, 2014.

[39] J. Le Roux, J. R. Hershey, S. T. Wisdom, and H. Erdogan, "SDR – half-baked or well done?" Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA, Tech. Rep., 2018.

[40] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 4, Jul. 2006.