The Phasebook: Building Complex Masks via Discrete Representations for Source Separation

Le Roux, J.; Wichern, G.; Watanabe, S.; Sarroff, A.; Hershey, J.

TR2019-008 March 29, 2019

Abstract

Deep learning based speech enhancement and source separation systems have recently reached unprecedented levels of quality, to the point that performance is reaching a new ceiling. Most systems rely on estimating the magnitude of a target source, either directly or by computing a real-valued mask to be applied to a time-frequency representation of the mixture signal. A limiting factor in such approaches is a lack of phase estimation: the phase of the mixture is most often used when reconstructing the estimated time-domain signal. We propose to estimate phase using "phasebook", a new type of layer based on a discrete representation of the phase difference between the mixture and the target. We also introduce "combook", a similar type of layer that directly estimates a complex mask. Wepresent various training and inference schemes involving these representations, and explain in particular how to include them in an endto-end learning framework. We also present an oracle study to assess upper bounds on performance for various types of masks using discrete phase representations. We evaluate the proposed methods on the wsj0-2mix dataset, a well-studied corpus for single-channel speaker-independent speaker separation, matching the performance of state-of-the-art mask-based approaches without requiring additional phase reconstruction steps.

IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)

THE PHASEBOOK: BUILDING COMPLEX MASKS VIA DISCRETE REPRESENTATIONS FOR SOURCE SEPARATION

Jonathan Le Roux, Gordon Wichern, Shinji Watanabe, Andy Sarroff, John R. Hershey

Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA

ABSTRACT

Deep learning based speech enhancement and source separation systems have recently reached unprecedented levels of quality, to the point that performance is reaching a new ceiling. Most systems rely on estimating the magnitude of a target source, either directly or by computing a real-valued mask to be applied to a time-frequency representation of the mixture signal. A limiting factor in such approaches is a lack of phase estimation: the phase of the mixture is most often used when reconstructing the estimated time-domain signal. We propose to estimate phase using "phasebook", a new type of layer based on a discrete representation of the phase difference between the mixture and the target. We also introduce "combook", a similar type of layer that directly estimates a complex mask. We present various training and inference schemes involving these representations, and explain in particular how to include them in an end-to-end learning framework. We also present an oracle study to assess upper bounds on performance for various types of masks using discrete phase representations. We evaluate the proposed methods on the wsj0-2mix dataset, a well-studied corpus for single-channel speaker-independent speaker separation, matching the performance of state-of-the-art mask-based approaches without requiring additional phase reconstruction steps.

***Index Terms*—** source separation, deep learning, phase estimation, discrete representation, mask inference

1. INTRODUCTION

The field of speech separation and speech enhancement has witnessed dramatic improvements in performance with the recent advent of deep learning-based techniques [1–11]. Most of these algorithms rely on the estimation of some sort of time-frequency (T-F) mask to be applied to the T-F representation of an input mixture signal, the estimated signal then being resynthesized using some inverse transform. Let us denote by $\boldsymbol{X} = (x_{t,f})$, $\boldsymbol{S} = (s_{t,f})$, and $\boldsymbol{N} = (n_{t,f})$ the complex-valued time-frequency representations of a mixture signal, a target source signal, and an interference signal, respectively at time t in frequency bin f. We also denote by $\theta_{t,f} = \angle(s_{t,f}/x_{t,f})$ the phase difference between the mixture and the target source. The time-frequency representation is typically taken to be the short-time Fourier transform (STFT), such that $x_{t,f} = s_{t,f} + n_{t,f}$. The goal of speech enhancement or separation is to recover an estimate $\hat{\boldsymbol{S}} = (\hat{s}_{t,f})$ of the signal $\boldsymbol{S}$ from the mixture $\boldsymbol{X}$. We focus here on algorithms that do so by estimating a mask $\boldsymbol{C} = (c_{t,f})$ such that $\hat{s}_{t,f} = c_{t,f} x_{t,f}$. Note that the interference signal itself could also be target for separation, such as in the case of speaker separation.

Masking is motivated by the sparseness of speech, since the majority of T-F bins in a mixture of speakers will contain mostly energy from only one of the speakers. Real-valued masks are therefore typically constrained to lie between 0 and 1, with 1 indicating time-frequency bins that contain only the target speaker, and therefore also contain the correct phase. However this logic does not apply to the T-F bins that contain significant energy from more than one speaker, and handling such cases leads us back to dealing with phase.

Until recently, getting good estimates of the magnitude was already difficult enough that improving the phase estimate over the noisy phase was not seen as a priority. With the advent of recent deep learning algorithms, the magnitude estimates have improved significantly, and the noisy phase has become a limiting factor to the overall performance. Because the noisy phases are typically inconsistent with the estimated magnitudes [12, 13], the reconstructed time-domain signal has a different magnitude spectrogram from the estimated one. Further improving the magnitude estimate by making it closer to the true target magnitude may actually lead to worse results, in terms of measures such as signal to noise ratio (SNR), if nothing is done to improve the phases. Limiting the estimates to the noisy phases thus places a ceiling on the achievable SNR, and makes magnitude estimation less straightforward. Improving upon the noisy phase therefore presents an opportunity to do better magnitude estimation as well.

If one optimizes both magnitudes and phase for best signal fidelity then exploring schemes where the magnitude mask goes beyond 1 becomes a reasonable option. When signals in a mixture are out of phase with each other they can cancel in a given T-F bin. In this case the magnitudes of each source are greater than that of the mixture, and so mask values of greater than 1 are required to accurately estimate the magnitude. This was explored in [14] with the introduction of a convex softmax activation function which interpolates between the values $0, 1, 2$ to obtain a continuous representation of the interval $[0, 2]$ as the target interval for the magnitude mask, leading to significantly better performance.

Interpolating between fixed prototypes can be seen as a coarse coding of the output. We propose to apply this idea to the estimation of a phase mask or a complex mask. That is, we combine a phase codebook, or *phasebook*, with a softmax layer to build various phase representations, either discrete or continuous; we also propose to directly model a complex mask without magnitude-phase factorization by combining a complex codebook, or *combook*, with a softmax layer to build various complex mask representations. These representations are flexible and can be incorporated within optimization frameworks that are regression-based, classification-based, or a combination of both.

Related works: Discrete representations of the phase for source separation were considered in [15] and [16], within a generative model based on mixtures of Gaussians. Some works have attempted to incorporate phase modeling for deep-learning-based source separa-

S. Watanabe is now with Johns Hopkins University, Baltimore, MD, USA. A. Sarroff is now with iZotope, Cambridge, MA, USA. J. R. Hershey is now with Google Research in Cambridge, MA, USA.

tion, such as estimating the phase difference for audio-visual separation in [17], and PhaseNet [18] which estimates discretized values of the target source phase using cross-entropy training. PhaseNet is close to a particular setup of our framework; however its use of argmax makes it less amenable to end-to-end training, whereas our framework has more flexible outputs and cost functions. The so-called complex ratio mask [19], is another deep learning system which considers a range of values that are not limited to $[0, 1]$ and uses a continuous real-imaginary representation, while we here focus mainly on discrete representations involving a magnitude-phase factorization or a direct modeling of the complex value (with the real and imaginary parts considered jointly).

Another, potentially complementary, way to improve the phase is to use phase reconstruction. Recent works applied phase reconstruction as a post-processing [11], then as part of the optimization pipeline [14]. We finally trained the T-F representations used in the phase reconstruction algorithm [20]. This is the current state-of-the-art in methods relying on time-frequency representations. Recently, a version of the TasNet algorithm [21] established a new state-of-the-art benchmark on the wsj0-2mix dataset, and introduced techniques that could be adopted in our framework, such as convolution layers instead of recurrent ones, layer normalization schemes, and the use of SI-SDR as the objective instead of the L^1 waveform approximation loss that we consider. It is unclear how these techniques would influence the performance of competing methods, and we shall consider incorporating them in our framework as future work.

2. MASK DESIGN USING DISCRETE REPRESENTATIONS

We propose to rely on discrete values to build representations for a complex ratio mask, either via its factorization into magnitude and phase components or directly as a complex value. Each of the magnitude, phase, or complex masks is estimated by combining discrete values in a scalar codebook, using probabilities obtained with a softmax layer.

Consider a scalar codebook of phase values, or phasebook, denoted by $\mathcal{F}_P = \{\theta^{(1)}, \ldots, \theta^{(P)}\}$. At each T-F bin t, f, a network can estimate a softmax probability vector $p_{\boldsymbol{\phi}}(\theta_{t,f}|\boldsymbol{O}) \in \Delta^{P-1}$, where $\boldsymbol{O}$ denotes the input features, $\boldsymbol{\phi}$ the network parameters, and $\Delta^n = \big\{(t_0, \ldots, t_n) \in \mathbb{R}^{n+1} \mid \sum_{i=0}^{n} t_i = 1 \text{ and } t_i \geq 0 \text{ for all } i\big\}$ is the unit n-simplex. We consider several options for using this softmax layer output vector to build a final output, either as probabilities, to select the most likely value (*argmax*) or sample a value (*sampling*), or as weights within some interpolation scheme (*interpolation*):

- argmax: $\theta^{\text{out}}_{t,f} = \operatorname{argmax} p_{\boldsymbol{\phi}}(\theta_{t,f}|\boldsymbol{O}),$ (1)
- sampling: $\theta^{\text{out}}_{t,f} \sim p_{\boldsymbol{\phi}}(\theta_{t,f}|\boldsymbol{O}),$ (2)
- interpolation: $\theta^{\text{out}}_{t,f} = \angle \sum_j p_{\boldsymbol{\phi}}(\theta_{t,f} = \theta^{(j)}|\boldsymbol{O})\, \mathrm{e}^{\mathrm{j}\theta^{(j)}}.$ (3)

Note that the interpolation in Eq. (3) is performed in the complex domain and that taking the angle implies a renormalization step; this interpolation is illustrated in Fig. 1. An advantage of this representation is that it takes into account phase wrapping, that is, the fact that any measure of difference between phase values should be considered modulo 2π. Indeed, with either sampling or argmax selection, there is no need to introduce a notion of proximity between values; with the interpolation uniform of Eq. (3), the phase is defined by its location around the unit circle, varies continuously with the softmax probabilities, and values such as $-\pi + \epsilon$ and $\pi - \epsilon$ for small ϵ can be obtained with probabilities close to each other. This would not be the case if phase was represented directly as a real-valued angle.

We can define similar "magbook" and "combook" representations for the magnitude mask and the complex mask, again interpolating using a convex sum over the codebook values with the softmax probabilities as weights. For the magnitude, this is an extension of the classical sigmoid activation function for the case of a fixed magbook of size 2 with elements $\{0, 1\}$ (referred to here as uniform magbook 2), and an extension of the convex softmax considered in [14] for the case of a fixed magbook of size 3 with elements $\{0, 1, 2\}$ (referred to here as uniform magbook 3).

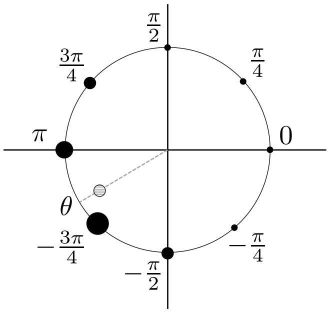

Fig. 1. Illustration of the phase interpolation scheme for a uniform phasebook with 8 elements. Softmax probabilities are displayed via the surface of each circle.

In the following, we shall call "phasebook layer" a layer computing phase values based on the outputs of a softmax layer and a phasebook via a method such as those above, and similarly for a "magbook layer" and a "combook layer". These layers allow us to define both discrete and continuous representations which can be involved in both classification-based and regression-based optimization frameworks. The continuous representations may lead to more accurate estimates, or be easier to include within an end-to-end training scheme. On the other hand, the discrete representations open the possibility to consider conditional probability relationships across variables combined with the chain rule, and may also avoid regression issues, for example where the estimated value is an interpolation of two values with high probability but itself has low probability.

3. PHASEBOOK WITH ARGMAX

To get an idea of the potential benefits of better phase modeling, we consider the argmax scheme for the phase mask, in which the system attempts to select the best codebook value at each T-F bin, and study the performance in oracle settings.

Given a phasebook $\mathcal{F}_P = \{\theta^{(1)}, \ldots, \theta^{(P)}\}$, the goal of our system is to estimate at each T-F bin (t, f) the codebook index $j_{t,f}$ such that $j_{t,f} = \operatorname{argmin}_j |m_{t,f}\mathrm{e}^{\mathrm{j}\theta^{(j)}} x_{t,f} - s_{t,f}|^2$, where $m_{t,f}$ is some estimate for the magnitude of the mask. The estimation is in fact independent of the magnitude mask value:

$$j_{t,f} = \underset{j}{\operatorname{argmin}} \cos(\theta^{(j)} - \angle(s_{t,f}/x_{t,f})). \quad (4)$$

We compare the performance of various oracle magnitude masks combined with the noisy phase, the true phase, and oracle quantized phases using uniform phasebooks with $P = 2, \ldots, 10$ elements, whose values equally partition the unit circle: $\mathcal{F}_P^{\text{uniform}} = \{0, \ldots, \frac{2p\pi}{P}, \ldots, \frac{2(P-1)\pi}{P}\}$. The oracle phases are obtained by selecting the best element in a phasebook according to Eq. 4. Performance is measured on the full wsj0-2mix evaluation set [6] using the scale-invariant signal-to-distortion ratio (SI-SDR) between the target speech and estimate [22]. We investigate the most popular magnitude masks, whose oracle performance when paired with the noisy phase was compared in [4]: ideal amplitude mask (IAM: $a^{\text{IAM}} = \frac{|s|}{|x|}$), here also considering its truncations to various thresh-

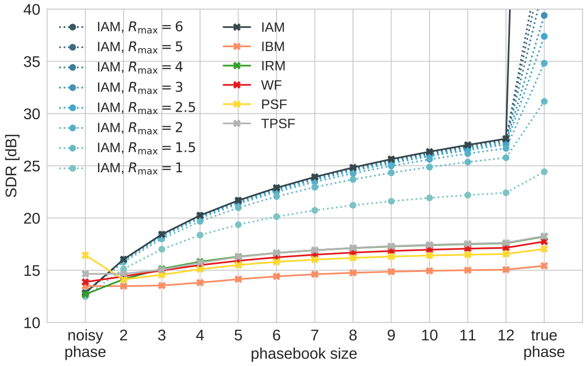

Fig. 2. Speech SI-SDR for truncated IAM and various classical masks with quantized phase difference for various phasebook sizes.

olds $R_{\max}$, phase sensitive filter (PSF: $a^{\text{PSF}} = \cos(\theta)\frac{|s|}{|x|}$), and its truncated version to $[0,1]$ (TPSF), ideal binary mask (IBM: $a^{\text{IBM}} = \delta(|s| > |n|)$), ideal ratio mask (IRM: $a^{\text{IRM}} = |s|/(|s|+|n|)$), and Wiener-filter-like mask (WF: $a^{\text{WF}} = |s|^2/(|s|^2+|n|^2)$). All these masks are real-valued, and only modify the magnitude of the mixture, and thus all masks do better with the true phase than with the noisy phase (except for PSF, which allows negative magnitudes which accommodate phase reversals in the noisy phase). The results are shown in Fig. 2.

We first notice that, apart from the unrestricted phase-sensitive mask, PSF, all masks lead to results under 15 dB when paired with the noisy phase. This confirms that the noisy phase drastically limits performance. As soon as a slightly better estimate of the phase is considered, performance significantly increases, especially for the IAM masks that consider magnitude ratio values above 1. For phases other than the noisy phase, we notice a very big jump in performance when allowing the truncation ratio to go from a classical value $R_{\max} = 1$ to an only slightly larger value $R_{\max} = 1.5$. Interestingly, very small codebook sizes already lead to high oracle performance, e.g., $P = 4$. In non-oracle conditions, of course, we need to find the right balance between upper-bound performance and classification accuracy.

4. OBJECTIVE FUNCTIONS

We consider the above representations as layers within a deep learning model for source separation, and we need to optimize the parameters $\boldsymbol{\phi}$ of the model under some objective function. We note that the codebook themselves can be considered fixed (to uniform or pre-trained values), or optimized jointly with the rest of the network. For magnitude masks we consider a fixed uniform magbook with 3 elements $\{0,1,2\}$ corresponding to the convex softmax activation proposed in [14]. We consider two types of training frameworks: train a phasebook layer for best phase accuracy using cross-entropy, after training the rest of the network separately using an objective involving the magnitude; use a phasebook or combook layer to obtain a complex mask estimate, and train the whole network jointly for best waveform domain reconstruction.

Cross-entropy on the phase: Let $\boldsymbol{j}^{\text{ref}}$ denote the reference values for the phase mask, which are the corresponding reference codebook indices obtained using Eq. (4). We can define an objective function based on the cross-entropy against the oracle codebook assignments for the softmax layer outputs of the phasebook layer as:

$$\mathcal{L}_{\text{CE-phase}}(\boldsymbol{\phi}) = -\sum_{t,f}\sum_{j}\delta(j, j^{\text{ref}}_{t,f})\log p_{\boldsymbol{\phi}}(\theta_{t,f} = \theta^{(j)}|\boldsymbol{O}). \tag{5}$$

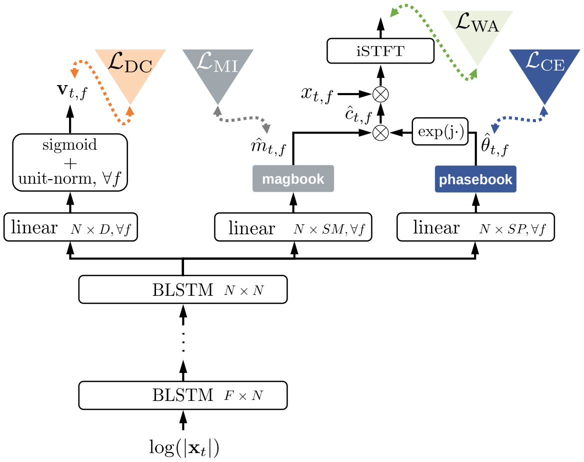
$\mathcal{L}_{\text{DC}}$
$\mathcal{L}_{\text{MI}}$
$\mathcal{L}_{\text{WA}}$
$\mathcal{L}_{\text{CE}}$
$\mathbf{v}_{t,f}$
iSTFT
$x_{t,f}$
$\hat{c}_{t,f}$
exp(j·)
$\hat{m}_{t,f}$
$\hat{\theta}_{t,f}$
sigmoid + unit-norm, $\forall f$
magbook
phasebook
linear $N \times D, \forall f$
linear $N \times SM, \forall f$
linear $N \times SP, \forall f$
BLSTM $N \times N$
BLSTM $F \times N$
$\log(|\mathbf{x}_t|)$

Fig. 3. Chimera++ network with phasebook-magbook MI head.

When using these training objectives, either sampling or argmax inference seem most appropriate for use at test time.

Time-domain objectives: Here we consider training the mask estimation networks end-to-end using a time-domain loss on the resulting signal, as proposed in [14]. That is, we use a waveform approximation (WA) objective defined on the time-domain signal $\hat{s}[l]$ reconstructed by inverse STFT from the masked mixture, using L^1 as distance. We also consider training through an unfolded phase reconstruction algorithm such as multiple input spectrogram inversion (MISI) [23], using the WA objective on the reconstructed time-domain signal $\hat{s}^{(K)}[l]$ after K iterations.

5. EXPERIMENTAL VALIDATION

We validate the proposed algorithms on the publicly available wsj0-2mix corpus [6], which is widely used for speaker-independent speech separation. 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 validation speakers are seen during training, while those in the test set are completely unseen. Sampling rate is 8 kHz.

5.1. Chimera++ network with phasebook-magbook MI head

Our system is based on the state-of-the-art chimera++ network [14], which combines within a multi-task learning framework a deep clustering head outputting a D-dimensional embedding for each T-F bin ($D = 20$ here), and a mask-inference (MI) head with convex softmax output which predicts a magnitude mask with values in $[0,2]$, here generalized to a magbook layer. T-F analysis and network training parameters are the same as in [14].

We also add a phasebook layer as a new head at the output of the final BLSTM layer, as shown in Fig. 3. The final complex mask is obtained by combining the outputs of the magbook and phasebook layers as $\hat{c}_{t,f} = \hat{m}_{t,f}e^{\mathrm{j}\hat{\theta}_{t,f}}$, and then multiplied with the complex mixture to obtain a complex T-F representation $\hat{s}_{t,f}$ of the target estimate. We still refer to the branch of the network used in computing the final output as the MI head, which now predicts a complex mask.

5.2. Training and inference schemes for phasebook

In this experiment, we start by pre-training chimera++ networks similarly to [14], i.e., with a uniform magbook 3 layer as MI head. For each of the magnitude spectrum approximation (MSA), phase-sensitive spectrum approximation (PSA), and WA losses as MI objective, we train such a network from scratch within the multi-task

Table 1. SI-SDR (dB) on the wsj0-2mix test set for various training paradigms from various pre-trained magnitude estimation networks.

	Network	Joint mag.	Mag. pretraining		
Phase estimate	Objective	training	MSA	PSA	WA
Noisy	-	×	10.5	11.1	11.8
Uniform phasebook 8 argmax	CE	×	10.7	11.1	11.8
Uniform phasebook 8 interp.	WA	×	11.2	11.1	12.0
Uniform phasebook 8 interp.	WA	✓	12.2	12.4	12.4

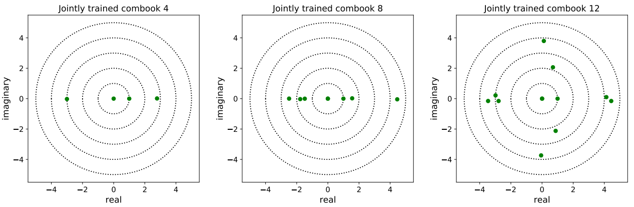

Fig. 4. Jointly trained combooks for $C \in \{4, 8, 12\}$ for chimera++ training followed by MI fine-tuning with WA objective.

learning setting involving the deep clustering and MI objectives, then discard the deep clustering head and fine-tune the MI head only. We then add a uniform phasebook 8 layer in the MI head ($P = 8$ led to best results in preliminary experiments), and we consider: (1) training only the phasebook layer with the rest of the network fixed, with the cross-entropy loss $\mathcal{L}_{\text{CE-phase}}$, and using the argmax scheme in Eq. 1 at inference time; (2) training only the phasebook layer with the rest of the network fixed, with the WA loss, using interpolation in Eq. (3) to compute the phase; and (3) training the whole network with the WA loss, again with phase interpolation.

Results are shown in Table 1 in terms of scale-invariant SDR (dB) [22] on the wsj0-2mix test set. The CE objective only provides SI-SDR improvements for networks pre-trained with the phase-unaware MSA objective, and is generally outperformed by the WA objective. This makes sense, as MSA-based magnitude estimates are likely closer to the true magnitude than those obtained with PSA and WA, which try to compensate for errors in the noisy phase; once the phasebook layer fixes these errors, which it learns to do without considering the interaction with the magnitude in the CE case, the compensation performed by the magnitude estimate may become extraneous or even detrimental. When training the phasebook layer with WA objective, the largest improvement is again observed for MSA. Finally, when allowing joint training of the magbook layer, all pretraining objectives attain their best performance, with PSA and WA obtaining slightly larger values than MSA. Overall, the WA objective with interpolation appears the most robust, both for pretraining and training. We thus focus on this configuration going forward.

5.3. Combook

We have so far considered factorized representations of the complex mask as a product of magnitude and phase masks. We now consider modeling it directly using a codebook of complex values. We train Chimera++ networks where the magnitude mask estimation layer is replaced by a complex mask estimation layer consisting of a softmax layer used to interpolate values of a combook. The networks are trained from scratch with both deep clustering and WA objectives, then fine-tuned with WA objective only. Examples of learned combooks are shown in Fig. 4 for $C \in \{4, 8, 12\}$. Interestingly, for small sizes such as $C \in \{4, 8\}$, the combook layer does not take advantage of non-real values, focusing first on covering negative values (for phase inversion), 0, and positive values. With $C = 12$, we do observe non-real values. However, the network appears inefficient in its usage of available values, learning seemingly redundant values. Table 2 compares SI-SDR results for combooks of various sizes (performance did not further improve for $c > 12$), in addition to different uniformly spaced magbook and phasebook configurations. In the current setup, the ability of the combook layer to estimate a complex mask via a single network layer works slightly better than estimating magnitude and phase via separate layers.

Table 2. SI-SDR improvement (dB) on the wsj0-2mix test set for various phasebook and combook sizes.

Codebook	SI-SDR (dB)
Jointly trained combook 4	12.1
Jointly trained combook 8	12.1
Jointly trained combook 12	12.6
Uniform magbook 3 w/ uniform phasebook 4	12.3
Uniform magbook 3 w/ uniform phasebook 8	12.4
Uniform magbook 3 w/ uniform phasebook 12	12.2

Table 3. SI-SDR improvement (dB) of recent systems on wsj0-2mix.

	MISI	SI-SDR
Approach	Iterations	[dB]
Chimera++ [11]	0	11.2
	5	11.5
Uniform magbook 3 w/ noisy phase [14]	0	11.8
	5	12.6
Unfolded MISI with learned untied transforms [20]	0	12.2
	5	**12.8**
Uniform magbook 3 w/ uniform phasebook 8	0	12.4
	5	12.6
Jointly trained combook 12	0	**12.6**
	5	12.6

5.4. Training through unfolded MISI

Following [14], for the best phasebook and combook networks, we add an unfolded MISI network with K iterations at the output of the MI head, and train using the WA-MISI-K loss function. Table 3 compares these systems with three recently proposed approaches: Chimera++ with noisy phase and MISI phase reconstruction as post-processing only [11]; Chimera++ trained through unfolded MISI phase reconstruction [14], equivalent to a uniform magbook 3 with noisy phase as initial phase; and Chimera++ with unfolded phase reconstruction with learned transforms replacing STFT and iSTFT at each layer [20]. The jointly trained combook 12 system obtains the best performance when no MISI iteration is performed, at 12.6 dB, beating the previous state-of-the-art 12.2 dB which involves further learning a transform replacing the final iSTFT [20]. If we allow ourselves 5 MISI iterations, all proposed systems reach 12.6 dB, but they are slightly outperformed by the system which learns replacements for the STFT/iSTFT transforms, with 12.8 dB. We shall leave it to future work to combine such transform learning with our proposed systems.

6. CONCLUSION AND FUTURE WORKS

We showed that both a combook layer and a combination of magbook and phasebook layers within an end-to-end framework can significantly improve performance of single-channel multi-speaker speech separation, especially reducing the need for further phase reconstruction. Future work will explore training through the argmax phasebook scheme, with the goal of introducing conditional probability relationships between T-F bins.

7. REFERENCES

[1] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, “Speech enhancement based on deep denoising autoencoder,” in *Proc. ISCA Interspeech*, 2013.

[2] F. J. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller, “Discriminatively trained recurrent neural networks for single-channel speech separation,” in *Proc. GlobalSIP Machine Learning Applications in Speech Processing Symposium*, 2014.

[3] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, “An experimental study on speech enhancement based on deep neural networks,” *IEEE Signal Processing Letters*, vol. 21, no. 1, 2014.

[4] 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)*, Apr. 2015.

[5] 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. Latent Variable Analysis and Signal Separation (LVA)*. Springer, 2015.

[6] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, “Deep clustering: Discriminative embeddings for segmentation and separation,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Mar. 2016.

[7] Y. Isik, J. Le Roux, Z. Chen, S. Watanabe, and J. R. Hershey, “Single-channel multi-speaker separation using deep clustering,” in *Proc. ISCA Interspeech*, Sep. 2016.

[8] 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)*, Mar. 2017.

[9] M. Kolbæk, 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*, vol. 25, no. 10, 2017.

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

[11] 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)*, Apr. 2018.

[12] J. Le Roux, H. Kameoka, N. Ono, A. de Cheveigné, and S. Sagayama, “Computational auditory induction by missing-data non-negative matrix factorization,” in *Proc. ISCA Workshop on Statistical and Perceptual Audition (SAPA)*, Sep. 2008.

[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, Mar. 2015.

[14] Z.-Q. Wang, J. Le Roux, D. Wang, and J. R. Hershey, “End-to-end speech separation with unfolded iterative phase reconstruction,” in *Proc. ISCA Interspeech*, Sep. 2018.

[15] S. J. Rennie, K. Achan, B. J. Frey, and P. Aarabi, “Variational speech separation of more sources than mixtures.” in *Proc. AISTATS*, 2005.

[16] A. Liutkus, C. Rohlfing, and A. Deleforge, “Audio source separation with magnitude priors: the BEADS model,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018.

[17] T. Afouras, J. S. Chung, and A. Zisserman, “The conversation: Deep audio-visual speech enhancement,” *arXiv preprint arXiv:1804.04121*, 2018.

[18] N. Takahashi, P. Agrawal, N. Goswami, and Y. Mitsufuji, “PhaseNet: Discretized phase modeling with deep neural networks for audio source separation,” *Proc. ISCA Interspeech*, 2018.

[19] D. S. Williamson, Y. Wang, and D. Wang, “Complex ratio masking for monaural speech separation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, 2016.

[20] G. Wichern and J. Le Roux, “Phase reconstruction with learned time-frequency representations for single-channel speech separation,” in *Proc. IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, Sep. 2018.

[21] Y. Luo and N. Mesgarani, “TasNet: Surpassing ideal time-frequency masking for speech separation,” *arXiv preprint arXiv:1809.07454*, Sep. 2018.

[22] J. Le Roux, S. T. Wisdom, H. Erdogan, and J. R. Hershey, “SDR – half-baked or well done?” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019.

[23] D. Gunawan and D. Sen, “Iterative phase estimation for the synthesis of separated sources from single-channel mixtures,” in *IEEE Signal Processing Letters*, 2010.