Deep clustering-based single-channel speech separation and recent advances

Aihara, Ryo; Wichern, Gordon; Le Roux, Jonathan

TR2021-020 March 23, 2021

Abstract

The recently-proposed deep clustering algorithm introduced significant advances in single-channel speaker-independent multi-speaker speech separation. In this paper, we review deep clustering and its improved method called chimera net. In addition, we describe our architectures for reducing the latency of deep clustering by combining block processing and teacher-student learning. Unfolding of a phase reconstruction algorithm and a complex mask estimation method for speech separation are also described.

Acoustical Science and Technology 2020

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

INVITED REVIEW

Deep clustering-based single-channel speech separation and recent advances

Ryo Aihara[1,*], Gordon Wichern[2] and Jonathan Le Roux[2]

[1]*Information Technology R and D Center, Mitsubishi Electric Corporation,*
5–1–1 Ofuna, Kamakura, 247–8501 Japan
[2]*Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA*

Abstract: The recently-proposed deep clustering algorithm introduced significant advances in single-channel speaker-independent multi-speaker speech separation. In this paper, we review deep clustering and its improved method called chimera net. In addition, we describe our architectures for reducing the latency of deep clustering by combining block processing and teacher-student learning. Unfolding of a phase reconstruction algorithm and a complex mask estimation method for speech separation are also described.

Keywords: Cocktail party problem, Speaker-independent speech separation, Chimera network, Low-latency, Phase estimation

PACS number: 43.60.Np, 43.60.Dh [doi:10.1250/ast.41.465]

1. INTRODUCTION

Speech separation is the task of estimating the individual speech signals that are mixed together and overlapping in a single-channel or multi-channel signal. In contrast with speech separation, we here define speech enhancement as the task of estimating speech signals that are mixed with non-speech signals. Before deep learning methods started being widely used, most speech separation approaches focused on multiple microphone scenarios [1], and single-channel speech separation remained as a challenging task. Deep learning techniques made it possible to combine feature extraction and acoustic modeling, resulting in drastically improved performance for single-channel speech enhancement [2,3]. However, in the case of single-channel speech separation, most early approaches were limited to speaker-dependent scenarios [4,5].

The main hurdle for single-channel speech separation is the so-called "permutation problem" where the correspondence between the outputs of an algorithm and the true sources is up to an arbitrary permutation [6]. Figure 1 illustrates the permutation problem occurring when separating two speakers. When speech samples of speaker pairs (A,B), (A,C), and (B,C) are used to train a speaker-independent speech separation model, it is unclear how to assign each speaker's sample to each output of the algorithm. If speaker A were assigned to the first target position for mixtures (A,B) and (A,C) while B and C were assigned to the second target, there would be no natural assignment for the case of mixture (B,C), as both of the speakers would need to be in the second position for consistency. Such speaker-based assignments are thus clearly not a proper solution in the speaker-independent case.

Deep clustering [6] represents a significant step towards solving this problem. This framework projects each time-frequency unit to a high-dimensional embedding such that the pairs of embeddings dominated by the same speaker are closer to each other while those dominated by different speakers are farther apart. Masks are then obtained by clustering these embeddings. This approach can avoid the direct estimation of time-frequency speaker masks and enables high quality speaker-independent speech separation. The success of deep clustering accelerated the research in single-channel speech separation. In recent years, direct time-frequency mask estimation methods [7,8] and time-domain speech separation have also been proposed [9].

This paper reviews single-channel speech separation methods based on deep clustering and introduces follow-up methods from our team. The rest of this paper is organized as follows: In Sect. 2, classic deep clustering is reviewed. In Sect. 3, our recent work to improve and expand deep clustering is explained. In Sect. 4, a low-latency method for single-channel speech separation is described. In Sect. 5, a single-channel speech separation approach that incorporates phase processing is introduced.

[*]e-mail: Aihara.Ryo@dx.MitsubishiElectric.co.jp

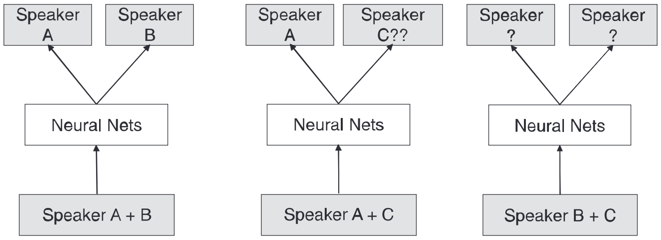

Fig. 1 Permutation problem [6,10].

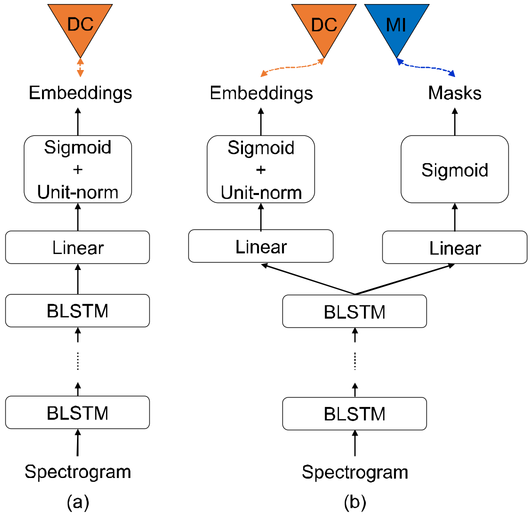

Fig. 2 (a) Deep clustering and (b) Chimera network.

2. DEEP CLUSTERING

The key idea of deep clustering [6] is to learn a high-dimensional embedding for each time-frequency bin such that the embeddings belonging to the same speaker are close to each other in the embedding space, and far from each other otherwise. This way, simple clustering methods such as k-means can be performed on the learned embeddings to perform separation at the test stage.

2.1. Formulation

We define as x a raw input signal and as $X_i = g_i(x)$, $i \in \{1, \ldots, N\}$, a feature vector indexed by an element i. In the case of audio signals, i is typically a time-frequency index (t, f), where t indexes the frame of the signal, f indexes frequency, and $X_i = X_{t,f}$ is the value of the complex spectrogram at the corresponding time-frequency bin. Motivated by the sparseness of speech, we assume that there exists a dominant speaker in each X_i even though there are multiple speakers in x. In the case of speech separation based on the estimation of time-frequency masks, a dominant speaker (a.k.a. class) or a dominance ratio of the speaker is estimated for each time-frequency bin and used to build masks to be applied to X_i.

The objective of deep clustering is to learn an embedding vector $\boldsymbol{v}_i \in \mathbb{R}^{1\times D}$ for each X_i in order to avoid the direct estimation of the time-frequency masks. Here, we consider a unit-norm embedding, so that $|\boldsymbol{v}_i|^2 = 1$. The target label is represented by $\boldsymbol{y}_i \in \mathbb{R}^{1\times C}$ mapping each element i to each of C clusters, so that $y_{i,c} = 1$ if element i is in cluster c, and $y_{i,c} = 0$ otherwise. Vertically stacking these embeddings on one hand and the target labels on the other, we form the embedding matrix $\boldsymbol{V} \in \mathbb{R}^{N\times D}$ and the label matrix $\boldsymbol{Y} \in \mathbb{R}^{N\times C}$. The matrix $\boldsymbol{Y}\boldsymbol{Y}^{\mathsf{T}}$ can then be considered as a binary affinity matrix that represents the cluster assignments in a permutation-independent way: $(\boldsymbol{Y}\boldsymbol{Y}^{\mathsf{T}})_{ij} = 1$ if elements i and j belong to the same cluster, and $(\boldsymbol{Y}\boldsymbol{Y}^{\mathsf{T}})_{ij} = 0$ otherwise. The embeddings are learned by minimizing the following objective function:

$$\begin{aligned}\mathcal{L}_{\text{DC-classic}}(\boldsymbol{V}, \boldsymbol{Y}) &= \|\boldsymbol{V}\boldsymbol{V}^{\mathsf{T}} - \boldsymbol{Y}\boldsymbol{Y}^{\mathsf{T}}\|_{\mathrm{F}}^2 \\ &= \|\boldsymbol{V}^{\mathsf{T}}\boldsymbol{V}\|_{\mathrm{F}}^2 + \|\boldsymbol{Y}^{\mathsf{T}}\boldsymbol{Y}\|_{\mathrm{F}}^2 - 2\|\boldsymbol{V}^{\mathsf{T}}\boldsymbol{Y}\|_{\mathrm{F}}^2 \qquad (1)\end{aligned}$$

where $\|\cdot\|_{\mathrm{F}}^2$ denotes the Frobenius norm, and the last expression is obtained by using a trace trick to lower the dimensionality of the computations.

Figure 2(a) shows the network architecture of deep clustering. In order to estimate embeddings, bidirectional long short-term memory (BLSTM) is used as a non-linear mapping function. At test time, k-means, which is a classical method of unsupervised clustering, is applied to the estimated embeddings to estimate binary masks. Because k takes any natural number, the estimated deep clustering network can be applied to mixtures of k speakers' speech, while the direct mask estimation method cannot. In [6,7], deep clustering is applied to the separation of three-speaker mixtures, even when only trained on two-speaker mixtures. Soft k-means can also be applied to estimate ratio masks [7]. Deep attractor network [11] is another embedding-based method, which estimates not only embeddings but also the centroids of embeddings.

2.2. Normalization of Cost Functions

Discarding or reducing the influence of time-frequency bins in silence regions is found to be very important for training deep clustering networks. The estimated mask value for such low-energy bins has little influence on the output, and their labelling is somewhat arbitrary. It is thus likely counterproductive to force the network to learn how to create embeddings for these bins. By filtering them out, the network can focus on learning embeddings for the time-frequency bins that actually contain some speech.

The square root of a weighting matrix $\boldsymbol{W} = \mathrm{diag}(w)$ is applied to (1) as follows:

$$\mathcal{L}_{\text{DC-classic},W}(\boldsymbol{V}, \boldsymbol{Y}) = \|\boldsymbol{W}^{\frac{1}{2}}(\boldsymbol{V}\boldsymbol{V}^{\mathsf{T}} - \boldsymbol{Y}\boldsymbol{Y}^{\mathsf{T}})\boldsymbol{W}^{\frac{1}{2}}\|_{\mathrm{F}}^2. \qquad (2)$$

There are multiple ways to define the weights for training. In [12], we obtained the best results with the magnitude ratio weight defined as the ratio of the mixture magnitude at time-frequency bin i over the sum of the mixture magnitudes at all bins within an utterance: $w_i = |x_i|/\sum_j |x_j|$.

As also proposed in [12], an alternative objective can be obtained by considering the objective function of the k-means algorithm applied to the embedding matrix $\boldsymbol{V}$ after normalizing it to have identity covariance. The objective function of k-means is defined as follows:

$$\hat{\boldsymbol{Y}} = \underset{\boldsymbol{Y}}{\operatorname{argmin}} \|\boldsymbol{V} - \boldsymbol{Y}(\boldsymbol{Y}^{\mathsf{T}}\boldsymbol{Y})^{-1}\boldsymbol{Y}^{\mathsf{T}}\boldsymbol{V}\|_{\mathrm{F}}^2. \tag{3}$$

Based on (3), a new objective function for deep clustering can be obtained using the whitened embedding matrix $\hat{\boldsymbol{V}} = \boldsymbol{V}(\boldsymbol{V}^{\mathsf{T}}\boldsymbol{V})^{-\frac{1}{2}}$ as follows:

$$\begin{aligned}
&\mathcal{L}_{\text{DC-W}}(\boldsymbol{V}, \boldsymbol{Y}) \\
&= \|\boldsymbol{V}(\boldsymbol{V}^{\mathsf{T}}\boldsymbol{V})^{-\frac{1}{2}} - \boldsymbol{Y}(\boldsymbol{Y}^{\mathsf{T}}\boldsymbol{Y})^{-1}\boldsymbol{Y}^{\mathsf{T}}\boldsymbol{V}(\boldsymbol{V}^{\mathsf{T}}\boldsymbol{V})^{-\frac{1}{2}}\|_{\mathrm{F}}^2 \\
&= D - \operatorname{tr}((\boldsymbol{V}^{\mathsf{T}}\boldsymbol{V})^{-1}\boldsymbol{V}^{\mathsf{T}}\boldsymbol{Y}(\boldsymbol{Y}^{\mathsf{T}}\boldsymbol{Y})^{-1}\boldsymbol{Y}^{\mathsf{T}}\boldsymbol{V}).
\end{aligned} \tag{4}$$

3. CHIMERA NETWORK

3.1. Mask Inference Learning

Mask inference (MI) is proposed in [6] as a benchmark approach for deep clustering. Its cost function can be defined as follows in the magnitude spectrum approximation (MSA) with L^2 norm case:

$$\mathcal{L}_{\text{MI-MSA},L^2} = \min_{\pi \in \mathcal{P}} \sum_c \|\hat{\boldsymbol{M}}_c \odot |\boldsymbol{X}| - |\boldsymbol{S}_{\pi(c)}|\|^2, \tag{5}$$

where $\mathcal{P}$ is the set of permutations on $\{1, \ldots, C\}$, $|\boldsymbol{X}|$ the mixture magnitude, $\hat{\boldsymbol{M}}_c$ the c-th estimated mask, $|\boldsymbol{S}_c|$ the magnitude of the c-th reference source, and $\odot$ the element-wise product. This objective function is also referred to as permutation invariant training (PIT) [8], as it calculates the cost for every permutation and picks the minimum one.

Phase-sensitive spectrum approximation (PSA) [13], which takes into account the phase difference between mixture and reference source in the cost function, is often used for speech enhancement from background noise. In speech enhancement, it is common to truncate the mask values to the range $[0, 1]$, in which case this technique is called truncated PSA (tPSA). In the case of speech separation, tPSA is adopted for MI as follows, with the L^1 norm leading to best results in [12]:

$$\begin{aligned}
\mathcal{L}_{\text{MI-tPSA},L^1} = \min_{\pi \in \mathcal{P}} \sum_c &\|\hat{\boldsymbol{M}}_c \odot |\boldsymbol{X}| \\
&- \boldsymbol{T}_0^{|\boldsymbol{X}|}(|\boldsymbol{S}_{\pi(c)}| \odot \cos(\theta_{\boldsymbol{X}} - \theta_{\pi(c)}))\|_1,
\end{aligned} \tag{6}$$

where $\theta_{\boldsymbol{X}}$ is the mixture phase, θ_c the phase of the c-th source, and $\boldsymbol{T}_a^b$ denotes a function truncating its input to the range $[a, b]$, where $a \leq b$.

3.2. Chimera Network Objective

The objective function of deep clustering can be combined with MI in a multi-task learning fashion, leveraging the regularizing property of the deep clustering loss and the simplicity of the mask-inference network. Figure 2(b) shows a chimera network architecture, where a shared stack of encoding layers such as BLSTMs is followed by separate heads for deep clustering and MI. The loss function we minimize is a weighted sum of a deep clustering loss and an MI loss:

$$\mathcal{L}_{\text{chi}} = \alpha\mathcal{L}_{\text{DC}}(V, Y) + (1 - \alpha)\mathcal{L}_{\text{MI}}, \tag{7}$$

where α is a weight for the deep clustering loss. At run time, we only use the MI output to make predictions.

4. LOW-LATENCY SPEECH SEPARATION

Unfortunately, the best performing deep clustering, mask inference, and chimera models are based on bidirectional recurrent networks (e.g., BLSTM), which require running forward and backward passes over an entire utterance before separation results can be obtained. This high-latency operation is unacceptable in many applications, e.g., as a front-end for speech recognition systems, however, simply replacing an offline BLSTM with a forward-only LSTM leads to unacceptable performance degradation. A trade-off between latency and separation performance can be achieved with the latency-controlled BLSTM (LC-BLSTM) [14,15], a block-based BLSTM where the input is cut into overlapping blocks and the latency is reduced to the block-size [16]. Another possible approach to closing the performance gap between BLSTM networks and low-latency LSTM/LC-BLSTM networks is to use teacher-student learning (also known as distillation [17]).

4.1. Latency-controlled BLSTM

BLSTMs are not practical for low-latency applications. Indeed, as illustrated on the left-hand side of Fig. 3, the forward LSTM operates from the first frame to the last frame of the input and the backward LSTM operates from the last frame to the first frame of the input. The output of each direction is concatenated and used as an input to the next layer. Therefore, one needs to wait until the BLSTM sees a whole utterance to do the computations.

In order to cope with such issues, latency-controlled BLSTM (LC-BLSTM) networks were proposed for automatic speech recognition in [14,18]. We here consider their application to speech separation, as an alternative to the low-latency approximations to BLSTM considered in [15] for speech enhancement. The LC-BLSTM architecture is illustrated in the right-hand side of Fig. 3. In LC-BLSTM, the input utterance is cut into non-overlapping blocks of fixed length N_m called the main block. Each main block has

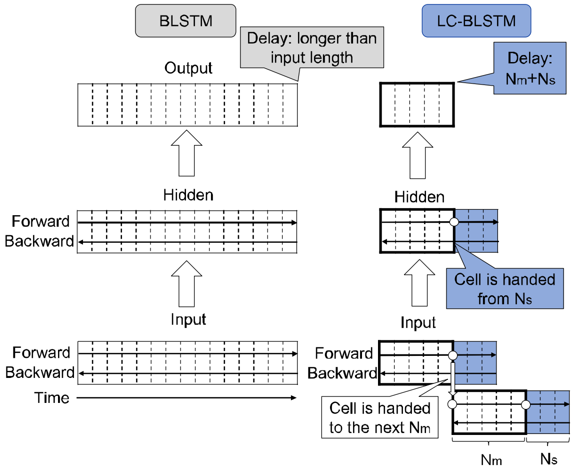

Fig. 3 BLSTM and LC-BLSTM.

a sub block of fixed length N_s, which is appended to the right context. The forward LSTM now operates from the first frame of a main block to the last frame of its sub block. The memory cell of the last frame of the main block is handed over to the next main block. The backward LSTM operates from the last frame of the sub block to the first frame of the main block, and its memory cell is always initialized with 0. The outputs of sub blocks are propagated to the last LC-BLSTM layer but are not propagated to the linear layer in Fig. 2. Also, the gradients from the sub blocks are not back-propagated. If N_s is set to 0, LC-BLSTM is equivalent to a block BLSTM where each block operates independently.

4.2. Teacher-student Deep Clustering

In order to improve the performance of low-latency models involving stacks of either LSTM or LC-BLSTM layers in place of the BLSTM stack, we consider applying teacher-student learning to deep clustering based speech separation. The procedure is illustrated in Fig. 3. A BLSTM-chimera network as presented in Sect. 3 is used as the teacher. As the student, we use either a stack of LSTM layers, which enables frame-wise operation, or a stack of LC-BLSTM layers, which enables block-wise operation. The teacher network is first trained using (7). The student network is then optimized under the following objective:

$$\mathcal{L}_{\text{stu}} = \alpha\mathcal{L}_{\text{DC}} + (1-\alpha)\mathcal{L}_{\text{MI}} + \beta\mathcal{L}_{\text{diff}}, \tag{8}$$

where $\mathcal{L}_{\text{diff}}$ denotes a distance between the weights of the final hidden layer of the teacher and student networks, and β a weight for that distance. We consider here two variants for the teacher-student distance:

$$\mathcal{L}_{\text{diff},L^p} = \|\boldsymbol{h}_N^{\text{t}} - \boldsymbol{h}_N^{\text{s}}\|_p^p, \quad p \in \{1,2\}, \tag{9}$$

where $\boldsymbol{h}_N^{\text{t}}$ and $\boldsymbol{h}_N^{\text{s}}$ denote the output of the final layer of the teacher network and that of the student network, respectively. In the above equations, $\boldsymbol{h}_N^{\text{t}}$ and $\boldsymbol{h}_N^{\text{s}}$ need to have the same number of units. If the number of units is different, a projection layer to expand or contract the dimensions accordingly can be used.

5. PHASE RECONSTRUCTION AND COMPLEX MASK ESTIMATION

As magnitude processing has improved and begun to approach oracle performance, i.e., the upper bound of source separation performance using the noisy phase, interest in processing phase for source separation has increased as well [19]. When combined with the noisy phase, separated source magnitudes may be inconsistent, i.e., no corresponding time-domain signal may exist [20,21]. Here, we describe a phase reconstruction method which is integrated into our separation algorithm [22,23]. Moreover, complex time-frequency mask estimation based on discrete representations is also introduced [24].

5.1. Phase Reconstruction

Time-frequency objectives such as (6) and (7) do not account for phase inconsistencies, so a waveform approximation (WA) objective that operates directly on the reconstructed time-domain signals $\hat{s}_1, \ldots, \hat{s}_C$ has been proposed in [22] as follows:

$$\mathcal{L}_{\text{WA}} = \min_{\pi\in\mathcal{P}} \sum_c \|\hat{s}_{\pi(c)} - s_c\|_1. \tag{10}$$

Iterative reconstruction techniques such as Griffin-Lim [25] and multiple input spectrogram inversion (MISI) [26] attempt to recover each source's clean phase by fixing its magnitude estimate and running alternating STFT and iSTFT iterations starting from the noisy phase. Applying iterative phase reconstruction as a post processing to a magnitude enhancement network often results in modest improvements in source separation performance [22].

The framework of deep unfolding [27] enables us to treat each iteration of the phase reconstruction as a layer in a neural network. Algorithm 1 describes an extension of

```
Input: Mixture signal $x$ in the time domain, estimated
       masks $\hat{M}_c$ for $c = 1, \ldots, C$, and number of
       itcrations $K$
$X = \text{STFT}_\Theta^{(0)}(x)$;
$\tilde{S}_c^{(0)} = \hat{M}_c \odot X$, for $c = 1, \ldots, C$;
for $k = 1, \ldots, K$ do
    $\hat{s}_c^{(k-1)} = \text{iSTFT}_\Theta^{(k-1)}(\tilde{S}_c^{(k-1)})$, for $c = 1, \ldots, C$;
    $\delta^{(k-1)} = x - \sum_{c=1}^{C} \hat{s}_c^{(k-1)}$;
    $\tilde{S}_c^{(k)} = |\tilde{S}_c^{(0)}| e^{\mathrm{j}\angle\text{STFT}_\Theta^{(k)}\left(\hat{s}_c^{(k-1)} + \frac{\delta^{(k-1)}}{C}\right)}$, for
      $c = 1, \ldots, C$;
end
return $\hat{s}_c^{(K)} = \text{iSTFT}_\Theta^{(K)}(\tilde{S}_c^{(K)})$, for $c = 1, \ldots, C$;
```

Algorithm 1: Unfolded MISI. $\text{STFT}_\Theta^{(k)}$ extracts a complex spectrogram of a signal, and $\text{iSTFT}_\Theta^{(k)}$ reconstructs a time-domain signal from a complex spectrogram.

the MISI algorithm considered here, where the STFT and iSTFT operations are generalized to STFT-like and iSTFT-like operations that incorporate parameters. For real-valued sequences such as audio signals, an N-point discrete Fourier transform (DFT) has $N/2+1$ unique complex coefficients. The DFT can be implemented using only real-valued operations by stacking the real and imaginary components and defining the elements of the basis matrix $W \in \mathbb{R}^{N+2 \times N}$ as

$$W_{i,n} = \begin{cases} w(n)\cos(2\pi ki/N), & i \in \left[\!\left[1, \dfrac{N}{2}+1\right]\!\right] \\ -w(n)\sin(2\pi ki/N), & i \in \left[\!\left[\dfrac{N}{2}+2, N+2\right]\!\right] \end{cases} \tag{11}$$

where we have incorporated the analysis window $w(n)$ into the basis matrix. By treating W as the weight matrix in a one-dimensional convolution layer, and setting the stride parameter of this layer equal to the hop size, we can efficiently create STFT-like layers with learnable basis matrices. The inverse DFT matrix can be defined similarly to (11) by using the synthesis window and accounting for the appropriate normalization terms. We can again implement a trainable iSTFT-like layer using transposed convolutions.

Figure 4 shows the overall block diagram of the system. In [22], the mask inference network is trained using the objective (10) while keeping the STFT and iSTFT layers fixed. Further improvement is obtained when the forward and inverse transform parameters are untied and independently learned for each STFT-like and iSTFT-like layer [23].

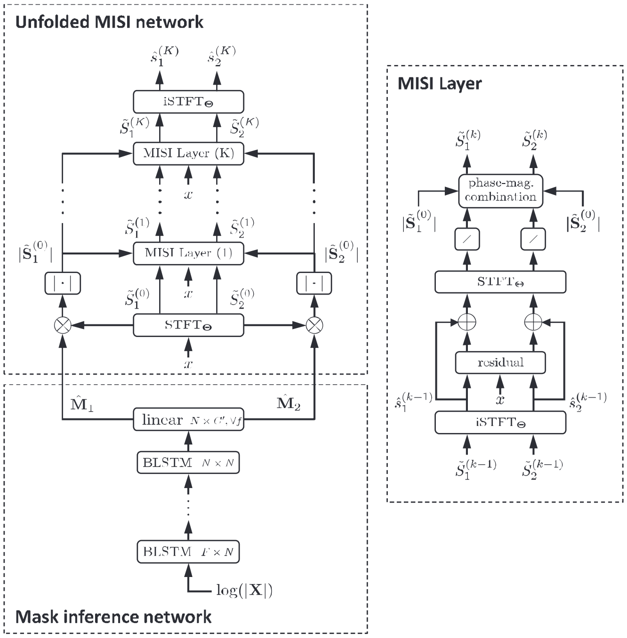

Fig. 4 Unfolded MISI network.

5.2. Complex Mask Estimation

Another way to improve the phase is through direct estimation of complex masks. By using complex values, complex ratio masks can modify both the magnitude and the phase of the mixture to obtain an estimate of a source. For example, complex ratio masks using a continuous real-imaginary representation were proposed in [28]. We here focus mainly on discrete representations involving a magnitude-phase factorization (phasebook) or a direct modeling of the complex value (combook) [24].

Consider a scalar codebook of phase values, or phasebook, denoted by $\mathcal{F}_P = \{\theta^{(1)}, \dots, \theta^{(P)}\}$. A network can estimate a softmax probability vector $p_{\boldsymbol{\phi}}(\theta_{t,f}|\boldsymbol{O}) \in \Delta^{P-1}$ at each time-frequency bin t, f, where $\boldsymbol{O}$ denotes the input features, $\boldsymbol{\phi}$ the network parameters, and $\Delta^n = \{(t_0, \dots, t_n) \in \mathbb{R}^{n+1} | \sum_{j=0}^{n} t_j = 1, t_i \geq 0, \forall i\}$ is the unit n-simplex. In [24], 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 as in [29] or sample a value, or as weights within some interpolation scheme as follows, which leads to the best results:

$$\theta_{t,f}^{\mathrm{out}} = \angle \sum_j p_{\boldsymbol{\phi}}(\theta_{t,f} = \theta^{(j)}|\boldsymbol{O})\, \mathrm{e}^{\mathrm{j}\theta^{(j)}}. \tag{12}$$

Note that the interpolation in (12) is performed in the complex domain and that taking the angle implies a renormalization step; this interpolation is illustrated in Fig. 5. 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, there is no need to introduce a notion of proximity between values; with (12), the phase is defined by its location around the unit circle, varies continuously with the softmax probabilities,

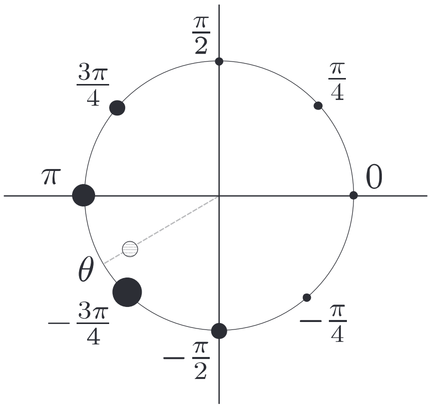

Fig. 5 Illustration of the phase interpolation scheme for a uniform phasebook with 8 elements. Softmax probabilities are displayed via the surface of each circle.

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 need to optimize the parameters $\boldsymbol{\phi}$ of the model under some objective function. We note that the codebooks themselves can be considered fixed (to uniform or pre-trained values), or optimized jointly with the rest of the network. 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 $\{0, 1\}$ and an extension of the convex softmax considered in [22] for the case of a fixed magbook of size 3 with elements $\{0, 1, 2\}$. 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 using (10). The latter approach obtains the better result.

To build a complex mask estimation network architecture based on combook, we replace the MI layer of chimera network, which is shown in Fig. 2(b), with a combook layer. We train the network using the WA objective (10) on the time-domain signal reconstructed by inverse STFT from the masked mixture, where the mask is obtained as the output $\hat{c}_{t,f}$ of the combook layer. In [24], the jointly trained combook 12 system obtains almost the same performance as the system which learns replacements for the STFT/iSTFT transforms presented in Sect. 5.1.

6. CONCLUSIONS

This paper reviewed deep clustering, one of the breakthrough techniques for single-channel speech separation, and introduced various follow up methods involving chimera networks, low-latency speech separation, phase reconstruction, and complex mask estimation. Several other extensions and related works have been recently investigated by our team. Multi-channel deep clustering was proposed in [30], where the phase difference of channels is added to the input. Embeddings are estimated from the combinations of multiple two-channel deep clustering networks. Our speech separation is also evaluated with automatic speech recognition (ASR). In [31], chimera network-based time-frequency mask estimation, and ASR are concatenated in an end-to-end manner. Direct text sequence recognition without mask estimation from the overlapped speech has also been proposed [32]. Deep clustering and chimera network have also been applied to musical audio separation [33] and noisy speech separation [34]. Evaluation measures for single-channel speech separation are discussed in [35].

ACKNOWLEDGMENT

The authors would like to thank Dr. Yohei Okato, Mr. Shu Murayama, and Mr. Toshiyuki Hanazawa (Mitsubishi Electric) for fruitful discussions.

REFERENCES

[1] K. Farrell, R. Mammone and J. L. Flanagan, "Beamforming microphone arrays for speech enhancement," *Proc. ICASSP*, pp. 285–288 (1992).

[2] F. Weninger, F. Eyben and B. Schuller, "Single-channel speech separation with memory-enhanced recurrent neural networks," *Proc. ICASSP*, pp. 3709–3713 (2014).

[3] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation," *IEEE Trans. Audio Speech Lang. Process.*, **21**, 1381–1390 (2013).

[4] J. R. Hershey, S. J. Rennie, P. A. Olsen and T. T. Kristjansson, "Super-human multi-talker speech recognition: A graphical modeling approach," *Comput. Speech Lang.*, **24**, 45–66 (2010).

[5] T. Virtanen, "Speech recognition using factorial hidden Markov models for separation in the feature space," *Proc. Interspeech*, pp. 89–92 (2006).

[6] J. R. Hershey, Z. Chen, J. Le Roux and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," *Proc. ICASSP*, pp. 31–35 (2016).

[7] Y. Isik, J. Le Roux, Z. Chen, S. Watanabe and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," *Proc. Interspeech*, pp. 545–549 (2016).

[8] D. Yu, X. Chang and Y. Qian, "Recognizing multi-talker speech with permutation invariant training," *arXiv preprint* (2017).

[9] Y. Luo and N. Mesgarani, "TasNet: Surpassing ideal time-frequency masking for speech separation," *arXiv preprint* (2018).

[10] Z. Chen, "Single channel auditory source separation with neural network," *Ph.D. dissertation, Columbia University* (2017).

[11] Z. Chen, Y. Luo and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," *Proc. ICASSP*, pp. 246–250 (2017).

[12] Z.-Q. Wang, J. Le Roux and J. R. Hershey, "Alternative objective functions for deep clustering," *Proc. ICASSP*, pp. 686–690 (2018).

[13] H. Erdogan, J. R. Hershey, S. Watanabe and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," *Proc. ICASSP*, pp. 708–712 (2015).

[14] S. Xue and Z. Yan, "Improving latency-controlled BLSTM acoustic models for online speech recognition," *Proc. ICASSP*, pp. 5340–5344 (2017).

[15] G. Wichern and A. Lukin, "Low-latency approximation of bidirectional recurrent networks for speech denoising," *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 66–70 (2017).

[16] R. Aihara, T. Hanazawa, Y. Okato, G. Wichern and J. Le Roux, "Teacher-student deep clustering for low-delay single channel speech separation," *Proc. ICASSP*, pp. 690–694 (2019).

[17] G. Hinton, O. Vinyals and J. Dean, "Distilling the knowledge in a neural network," *Proc. NIPS Deep Learning Workshop*, 9 pages (2014).

[18] Y. Zhang, G. Chen, D. Yu, K. Yaco, S. Khudanpur and J. Glass, "Highway long short-term memory RNNs for distant speech recognition," *Proc. ICASSP*, pp. 5755–5759 (2016).

[19] Y. Wakabayashi, "Speech enhancement using harmonic-structure-based phase reconstruction," *Acoust. Sci. & Tech.*, **40**, 162–169 (2019).

[20] J. Le Roux, N. Ono and S. Sagayama, "Explicit consistency constraints for STFT spectrograms and their application to phase reconstruction," *Proc. ISCA Workshop on Statistical and Perceptual Audition* (*SAPA*), pp. 23–28 (2008).

[21] T. Gerkmann, M. Krawczyk-Becker and J. Le Roux, "Phase processing for single-channel speech enhancement: History and recent advances," *IEEE Signal Process. Mag.*, **32**, 55–66 (2015).

[22] Z.-Q. Wang, J. Le Roux, D. Wang and J. R. Hershey, "End-to-end speech separation with unfolded iterative phase reconstruction," *Proc. Interspeech*, pp. 2708–2712 (2018).

[23] G. Wichern and J. Le Roux, "Phase reconstruction with learned time-frequency representations for single-channel speech separation," *Proc. IEEE International Workshop on Acoustic Signal Enhancement* (*IWAENC*), pp. 396–400 (2018).

[24] J. Le Roux, G. Wichern, S. Watanabe, A. Sarroff and J. R. Hershey, "Phasebook and friends: Leveraging discrete representations for source separation," *IEEE J. Sel. Top. Signal Process.*, **13**, 370–382 (2019).

[25] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoust. Speech Signal Process.*, 236–243 (1984).

[26] D. Gunawan and D. Sen, "Iterative phase estimation for the synthesis of separated sources from single-channel mixtures," *IEEE Signal Process. Lett.*, 421–424 (2010).

[27] J. R. Hershey, J. Le Roux and F. Weninger, "Deep unfolding: Model-based inspiration of novel deep architectures," *arXiv preprint* (2014).

[28] D. S. Williamson, Y. Wang and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, **24**, 483–492 (2016).

[29] N. Takahashi, P. Agrawal, N. Goswami and Y. Mitsufuji, "Phasenet: Discretized phase modeling with deep neural networks for audio source separation," *Proc. Interspeech*, pp. 2713–2717 (2018).

[30] Z.-Q. Wang, J. Le Roux and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," *Proc. ICASSP*, pp. 1–5 (2018).

[31] S. Settle, J. Le Roux, T. Hori, S. Watanabe and J. R. Hershey, "End-to-end multi-speaker speech recognition," *Proc. ICASSP*, pp. 4819–4823 (2018).

[32] H. Seki, T. Hori, S. Watanabe, J. Le Roux and J. Hershey, "A purely end-to-end system for multi-speaker speech recognition," *Proc. Annu. Meet. Association for Computational Linguistics* (*ACL*), pp. 2620–2630 (2018).

[33] P. Seetharaman, G. Wichern, S. Venkataramani and J. Le Roux, "Class-conditional embeddings for music source separation," *Proc. ICASSP*, pp. 301–305 (2019).

[34] G. Wichern, E. McQuinn, J. Antognini, M. Flynn, R. Zhu, D. Crow, E. Manilow and J. Le Roux, "WHAM!: Extending speech separation to noisy environments," *Proc. Interspeech*, pp. 1368–1372 (2019).

[35] J. Le Roux, S. Wisdom, H. Erdogan and J. R. Hershey, "SDR — Half-baked or well done?" *Proc. ICASSP*, pp. 626–630 (2019).

Ryo Aihara is a Researcher at Information Technology R & D Center of Mitsubishi Electric Corporation in Kamakura, Japan. He received his B.Eng., M.Eng., and Ph.D. in 2012, 2014, and 2017, respectively all from Kobe University, Japan. He was a recipient of the Japan Society for the Promotion of Science Research Fellowship for Young Scientists (DC1) from 2014 to 2017. His research interests are in signal processing and machine learning applied to speech, audio, and image. He is a member of the IEEE and the Acoustic Society of Japan.

Gordon Wichern is a Principal Research Scientist at Mitsubishi Electric Research Laboratories (MERL) in Cambridge, Massachusetts. He received his B.Sc. and M.Sc. degrees from Colorado State University in electrical engineering and his Ph.D. from Arizona State University in electrical engineering with a concentration in arts, media and engineering, where he was supported by a National Science Foundation (NSF) Integrative Graduate Education and Research Traineeship (IGERT) for his work on environmental sound recognition. He was previously a member of the research team at iZotope, inc. where he focused on applying novel signal processing and machine learning techniques to music and post production software, and a member of the Technical Staff at MIT Lincoln Laboratory where he worked on radar signal processing. His research interests include audio, music, and speech signal processing, machine learning, and psychoacoustics.

Jonathan Le Roux is a Senior Principal Research Scientist and the Speech and Audio Team Leader at Mitsubishi Electric Research Laboratories (MERL) in Cambridge, Massachusetts. He completed his B.Sc. and M.Sc. degrees in Mathematics at the École Normale Supérieure (Paris, France), his Ph.D. degree at the University of Tokyo (Japan) and the Université Pierre et Marie Curie (Paris, France), and worked as a postdoctoral researcher at NTT's Communication Science Laboratories from 2009 to 2011. His research interests are in signal processing and machine learning applied to speech and audio. He has contributed to more than 90 peer-reviewed papers and 20 patents in these fields. He is a founder and chair of the Speech and Audio in the Northeast (SANE) series of workshops, a Senior Member of the IEEE and a member of the IEEE Audio and Acoustic Signal Processing Technical Committee (AASP).