Listening and grouping: an online autoregressive approach for monaural speech separation

Zeng-Xi Li, Yan Song, Li-Rong Dai and Ian McLoughlin

***Abstract*—This paper proposes an autoregressive approach to harness the power of deep learning for multi-speaker monaural speech separation. It exploits a causal temporal context in both mixture and past estimated separated signals and performs online separation that is compatible with real-time applications. The approach adopts a learned listening and grouping architecture motivated by computational auditory scene analysis, with a grouping stage that effectively addresses the label permutation problem at both frame and segment levels. Experimental results on the WSJ0-2mix benchmark show that the new approach can achieve better signal-to-distortion ratio (SDR) and perceptual evaluation of speech quality (PESQ) scores than most state-of-the-art methods for both closed-set and open-set evaluations; even methods that exploit whole-utterance statistics for separation. It achieves this while requiring fewer model parameters.**

***Index Terms*—Speech separation, deep learning, label permutation problem, computational auditory scene analysis**

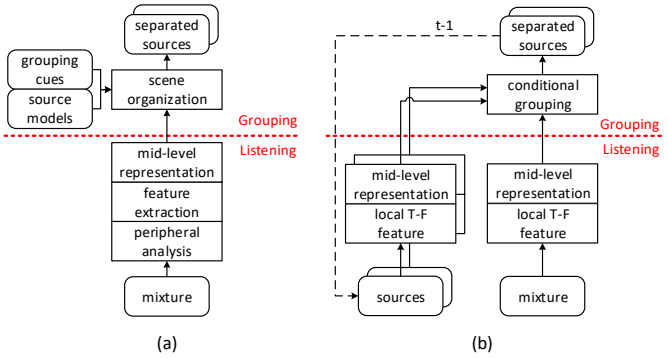

Fig. 1. The architecture of (a) a traditional CASA system from [4] and (b) the proposed listening and grouping method. The dotted line with label $t-1$ feeds back previously separated sources for the next time step. More details on the listening and grouping stages are described in Section IV.

I. Introduction

Despite recent progress in robust Automatic Speech Recognition [1], performance is still far from satisfactory for real-world applications like multi-speaker meeting transcription, audio/video captioning and hearing impairment assistants. The presence of multi-speaker interference is widely recognized as one of the main constraints. By contrast, humans can follow speech of interest in the presence of overlapping sources using innate listening and grouping [2] capabilities. These abilities have inspired research into computational auditory scene analysis (CASA) [3]–[6] for over half a century.

Prior to the emergence of deep learning, traditional CASA-based approaches, as shown in Fig. 1(a), followed listening and grouping rules that were typically hand-engineered or heuristic in nature, and utilized to group Time-Frequency (T-F) units belonging to the same speaker [3], [4]. Meanwhile, in [7]–[9], different grouping rules that utilize non-negative matrix factorization (NMF) and factorial Gaussian mixture model-hidden Markov models (GMM-HMM) were proposed. While approaches differ greatly, these techniques tend to suffer from similar issues relating to performance with unseen speakers, limitations on the exploitation of temporal or spectral dynamics, and high complexity – particularly when scaling to additional sources [10], [11].

Many recent methods have exploited the power of deep learning to formulate separation as a multi-class regression problem and learn an effective mapping from mixture to source T-F masks [12]–[14]. The improved listening ability performs well for dissimilar sources, but overlapping unseen speakers with similar characteristics are extremely difficult to separate. This is exacerbated by the *label permutation problem* [4], [11], [15], which will be detailed in Section III-A.

More recently, different grouping methods based on deep learning such as deep clustering (DPCL) [16]–[19], deep attractor network (DANet) [20], [21] and permutation invariant training (PIT) [15], [22], [23], were proposed to address the label permutation problem. The main idea of such methods is to determine source assignment based on a similarity measurement in embedding space or in original spectral space (*e.g.*, distance of embeddings in DPCL and DANet, mean square error (MSE) between estimated and target magnitude spectra in PIT).

Thanks to powerful listening network structures and effective grouping strategies, DPCL, DANet and PIT have achieved significant progress in speech separation [11], [19]. State-of-the-art methods usually operate in an offline manner; a long segment or a whole utterance mixture is fed into a network and processed together to yield a separation result. However in online scenarios, where current separated sources are generated without reference to future mixture inputs, state-of-the-art performance has a significant gap compared to offline methods which exploit both past and future context [15], [24]. But these are unlike the human auditory system – we can follow target

Manuscript received July 31, 2018; revised November 2, 2018; This work was supported in part by National Key R& D Program (Grant No. 2017YFB1002200 and 2017YFB1002202), National Natural Science Foundation of China (Grant No. U1613211), and by Key Science and Technology Project of Anhui Province (Grant No. 17030901005).

Z.-X. Li, Y. Song and L.-R. Dai, National Engineering Laboratory for Speech and Language Information Processing, University of Science and Technology of China, Hefei, China., lzx2010@mail.ustc.edu.cn; songy@ustc.edu.cn; lrdai@ustc.edu.cn

Ian McLoughlin, School of Computing, University of Kent, Medway, UK. ivm@kent.ac.uk

speech in babble with low latency, using past context only.

In this paper, we propose an online autoregressive approach in an explicit listening and grouping architecture, which can address the label permutation problem and meet online requirements, as shown in Fig. 1(b). Working in the spectral domain, the listening stage simultaneously and independently extracts mid-level representations [3], [4] of past estimated source frames and current mixture frames. The grouping stage then jointly consumes those representations to estimate current separated sources by modeling the dependency and interaction of mixture and sources. In strict observance of causality, the grouping outputs are fed back as input sources to the listening stage only for the following time step. Compared with traditional CASA systems shown in Fig. 1(a), the listening and grouping blocks are themselves neural networks trained jointly to exploit not only the temporal context of the mixture signal, but also enforce a temporal context constraint on the estimated sources, which is very effective in reducing mid-utterance speaker switching.

The proposed approach represents another class of deep learning based monaural speaker-independent speech separation. The main novelties are (i) its autoregressive nature, allowing output source order to be determined without additional operations (it is just the same as the input sources) and (ii) online processing of current and past frames, making it inherently suitable for low-latency online applications.

Listening and grouping could be implemented with recurrent neural networks (RNN) like long-short term memory (LSTM) [25], but in Section IV-A and IV-B we will introduce a specific structure to take full advantage of temporal context information in mixture and sources, which can exploit dependency and continuity of the same source. We evaluate the online approach on the WSJ0-2mix [16] dataset, showing that this approach can outperform state-of-the-art online methods and even achieve comparable or higher separation performance than the majority of state-of-the-art offline methods.

The remainder of this paper is organized as follows: after Section II discusses related work, Section III presents more detail on monaural speech separation and the label permutation problem. Section IV introduces listening and grouping and details the proposed network structure. Section V reports experimental results and Section VI concludes the paper.

II. Related Work

The CASA-based monaural speech separation approaches [3], [5], [26] are inspired by auditory scene analysis (ASA) [2], which perform listening and grouping in different ways. A traditional CASA system [3] is shown in Fig.1(a). In the listening stage, peripheral analysis is applied on the mixture signal for acquiring acoustic features such as periodicity and onsets/offsets. The grouping stage then uses these features to form mid-level representations for scene organization and speech separation based on source models and grouping cues. However, the listening and grouping rules are generally heuristically designed, leading to limited success for complex monaural speech separation tasks.

Recently, with the advance of deep learning techniques, the performance of speech separation has been significantly improved. For most deep learning-based methods, a listening stage uses neural networks to extract mid-level representations. These are used in a grouping stage along with additional information [27], [28] or operations to generate ordered estimated sources. In DPCL [16]–[19], the T-F bin similarity is measured in an embedding space. The key to DPCL is a deep network to generate embeddings for T-F bins in the mixture spectrogram. During grouping, a clustering algorithm for all embeddings is used to build segments of each source. DANet [20], [21] extends DPCL by creating an attractor point for each source in the embedding space. Unlike DPCL and DANet, PIT [15], [22], [23] measures similarity in the original spectral space, and determines the best label assignment by comparing separation errors of all possible orders.

In the latest research, several methods have been proposed to improve or extend DPCL, DANet and PIT frameworks. One main aspect is to focus on improved listening network structures, such as grid-LSTM [23] and gated convolutional networks [29]. Another is to explore better objective functions and training schemes. For example, in [30], speaker identification loss is added to the final loss function to reduce separation and permutation error. And in [31] adversarial training was introduced along with a sophisticated network to improve separation performance, while [18] takes a different approach where alternative objective functions such as whitened k-means loss are explored for DPCL.

Some other works further combine DPCL, DANet and PIT to acquire better separation results. For instance, Liu and Wang [32] decomposed the separation task into simultaneous and sequential grouping stages from a CASA perspective. The two grouping stages were implemented with individual bi-directional LSTM (BLSTM) networks, which are trained following PIT and DPCL frameworks respectively. Among the methods discussed above, the best separation performance is currently achieved by Wang et al. [19], using an unfolded iterative phase reconstruction algorithm, originating from multiple input spectrogram inverse (MISI) [33], applied in an end-to-end training structure.

Few of the recent separation architectures are compatible with online processing, but one example is TasNet [24], [34], [35], a network able to directly model a mixture waveform using an encoder-decoder framework based on PIT.

Unlike the methods mentioned above, this paper proposes an online autoregressive approach, which is an extension of our previous source-aware context network [36]. As shown in Fig.1(b), our approach first inputs the mixture and previously separated source frames, then directly outputs estimated sources, which are in turn fed back as inputs during the next time step. Moreover, a MISI-inspired (but online-compatible) algorithm is incorporated for waveform reconstruction.

III. Monaural Speech Separation

The task of monaural speech separation is to estimate S individual source signals $x_{s,n}, s = 1, \ldots, S$ from a single-channel mixture of speech y_n, given only the observed input y_n. In real-world situations, sources may be degraded by reverberation, but in this paper we only focus on the condition that y_n is linearly mixed, i.e., $y_n = \sum_{s=1}^{S} x_{s,n}$.

Apart from a few systems that perform separation directly in the waveform domain, waveforms are usually first transformed into time-frequency domain spectra by short-time Fourier transformation (STFT), using an analysis window w_n with FFT length N and frame shift R. The relationship between mixture and source spectra can then be formulated as,

$$Y_{t,f} = \sum_{s=1}^{S} X_{s,t,f} \quad (1)$$

$$Y_{t,f} = \sum_{n=-\infty}^{\infty} w_{n-tR} y_n e^{-j2\pi fn/N} \quad (2)$$

$$X_{s,t,f} = \sum_{n=-\infty}^{\infty} w_{n-tR} x_{s,n} e^{-j2\pi fn/N} \quad (3)$$

where t and f are frame and frequency indices respectively. When estimated sources spectra $\hat{X}_{s,t,f}$ are obtained, separated waveforms can be reconstructed by inverse STFT [37]:

$$\hat{x}_{s,n} = \frac{\sum_{t=-\infty}^{\infty} w_{n-tR} \frac{1}{N} \sum_{f=0}^{N-1} \hat{X}_{s,t,f} e^{j2\pi fn/N}}{\sum_{t=-\infty}^{\infty} w_{n-tR}^2} \quad (4)$$

There are several ways to acquire $\hat{X}_{s,t,f}$ in deep learning based techniques. One idea is to focus on the complex domain, for example estimating a complex ideal ratio mask [38] that jointly enhances both real and imaginary components. Another typical way is to only estimate magnitude spectra $|X_{s,t,f}|$, while the phase of $\hat{X}_{s,t,f}$ is either obtained directly from mixture phase $\angle Y_{t,f}$ or from a phase retrieval algorithm given $|\hat{X}_{s,t,f}|$ and $Y_{t,f}$, such as the Griffin-Lim algorithm [37] or MISI [33]. An online version of MISI is developed for the experiments in this paper.

A. Label Permutation Problem

Most deep learning approaches cast speech separation as a multi-class regression problem, i.e., source magnitude spectra $|X_{s,t,f}|$ are recovered by a neural network, given mixture magnitude spectra $|Y_{t,f}|$. For ease of description, we will focus on two-source notation. Generally, the separation model H can be formulated as,

$$\hat{\mathbf{x}}_{1,t}, \hat{\mathbf{x}}_{2,t} = H(\mathbf{y}_{t+Q}, \ldots, \mathbf{y}_{t-P}) \quad (5)$$

where $\hat{\mathbf{x}}_{s,t} = [|\hat{X}_{s,t,1}|, \ldots, |\hat{X}_{s,t,F}|]$ and $s = 1, 2$ are the positive frequency parts of the estimated source magnitude spectra. $\mathbf{y}_t = [|Y_{t,1}|, \ldots, |Y_{t,F}|]$ is the corresponding mixture magnitude spectra, and $F = \lfloor N/2 \rfloor + 1$, Q and P are receptive field length of future and past spectra respectively. In order to estimate target source spectra $\mathbf{x}_{s,t}$, the model H has to learn interaction and dependency between mixture and corresponding sources from a representative training data set.

During training, at each time step t, the error between targets $[\mathbf{x}_{1,t}, \mathbf{x}_{2,t}]$ and outputs $[\hat{\mathbf{x}}_{1,t}, \hat{\mathbf{x}}_{2,t}]$ needs to be computed for back-propagation. When $\mathbf{x}_{1,t}$ and $\mathbf{x}_{2,t}$ have very different time and frequency domain characteristics, e.g., $\mathbf{x}_{1,t}$ is the spectra of a speech signal and $\mathbf{x}_{2,t}$ is from background noise or music, then the ordering of corresponding output sources usually remains unchanged. However, for multi-speaker separation using only input $\mathbf{y}$, it is unknown in advance whether the correct output ordering should be $[\hat{\mathbf{x}}_{1,t}, \hat{\mathbf{x}}_{2,t}]$ or $[\hat{\mathbf{x}}_{2,t}, \hat{\mathbf{x}}_{1,t}]$. As a result, conflicting gradients produced by incorrect ordering will prevent the network from converging, especially when sources come from the same gender speakers. This is referred to as the *label permutation problem* [4], [11], [15]. DPCL, DANet and PIT can also be represented by Eqn. (5). As described in Section II, the final output ordering (and label permutation) is determined by additional similarity measures.

IV. Listening and Grouping

As mentioned, unlike most deep learning approaches formulated as Eqn. (5), the proposed approach aims to implicitly model the conditional distribution of current source spectra, given past source and mixture spectra, i.e.,

$$\hat{\mathbf{x}}_{1,t} \sim p(\mathbf{x}_{1,t}|\mathbf{x}_{1,t-1}, \ldots, \mathbf{x}_{1,t-P}; \mathbf{x}_{2,t-1}, \ldots, \mathbf{x}_{2,t-P}; \mathbf{y}_t, \ldots, \mathbf{y}_{t-P-1}) \quad (6)$$

$$\hat{\mathbf{x}}_{2,t} \sim p(\mathbf{x}_{2,t}|\mathbf{x}_{2,t-1}, \ldots, \mathbf{x}_{2,t-P}; \mathbf{x}_{1,t-1}, \ldots, \mathbf{x}_{1,t-P}; \mathbf{y}_t, \ldots, \mathbf{y}_{t-P-1}) \quad (7)$$

To implement Eqns. (6-7), our approach consists of two main stages of listening and grouping, as indicated in Fig. 1(b). In the listening stage, sequences of source and mixture magnitude spectra are individually and simultaneously transformed into mid-level representations, which can be formulated as

$$\mathbf{u}_t = \mathcal{L}(\tilde{\mathbf{x}}_{1,t-1}, \ldots, \tilde{\mathbf{x}}_{1,t-P_1}) \quad (8)$$

$$\mathbf{w}_t = \mathcal{L}(\tilde{\mathbf{x}}_{2,t-1}, \ldots, \tilde{\mathbf{x}}_{2,t-P_1}) \quad (9)$$

$$\mathbf{v}_t = \mathcal{L}(\mathbf{y}_t, \ldots, \mathbf{y}_{t-P_1-1}) \quad (10)$$

where $\tilde{\mathbf{x}}_{s,t}$ is the input spectrum of source s, during inference $\tilde{\mathbf{x}}_{s,t'} = \hat{\mathbf{x}}_{s,t'}$, $\forall t' = t-1, \ldots, t-P_1$, P_1 is the receptive field length of past spectra in the listening stage, $\mathbf{u}$, $\mathbf{w}$ and $\mathbf{v}$ are mid-level representations of sources and mixture respectively, and $\mathcal{L}(\cdot)$ is the operator performing the listening stage. Considering that all positions of each speaker are equivalent and exchangeable for multi-speaker speech separation, in our proposed structure the parameters of $\mathcal{L}(\cdot)$ in Eqns. (8-10) are shared between all sources and the mixture. However, it is worth noting that using independent parameters for each source is also feasible, especially for tasks where sources are dissimilar and have different characteristics, e.g. speech enhancement (clean speech vs. noise). Conceptually, Eqns. (8-10) share some similarities with a summary vector [40].

In the grouping stage, estimated source spectra $\hat{\mathbf{x}}_{1,t}$ and $\hat{\mathbf{x}}_{2,t}$ are generated simultaneously given sequences of mid-level representations $\mathbf{u}$, $\mathbf{w}$, $\mathbf{v}$ from the listening stage, i.e.,

$$\hat{\mathbf{x}}_{1,t}, \hat{\mathbf{x}}_{2,t} = \mathcal{G}(\mathbf{u}_t, \ldots, \mathbf{u}_{t-P_2-1}; \mathbf{w}_t, \ldots, \mathbf{w}_{t-P_2-1}; \mathbf{v}_t, \ldots, \mathbf{v}_{t-P_2-1}) \quad (11)$$

where P_2 is the receptive field length of past spectra in the grouping stage, $\mathbf{u}$ and $\mathbf{w}$ can be considered as CASA-like grouping cues or source models [3], [4], and $\mathcal{G}(\cdot)$ is the operator performing the grouping stage. After this stage, $\hat{\mathbf{x}}_{1,t}$ and $\hat{\mathbf{x}}_{2,t}$ will be fed back as inputs to the listening stage for the next time step.

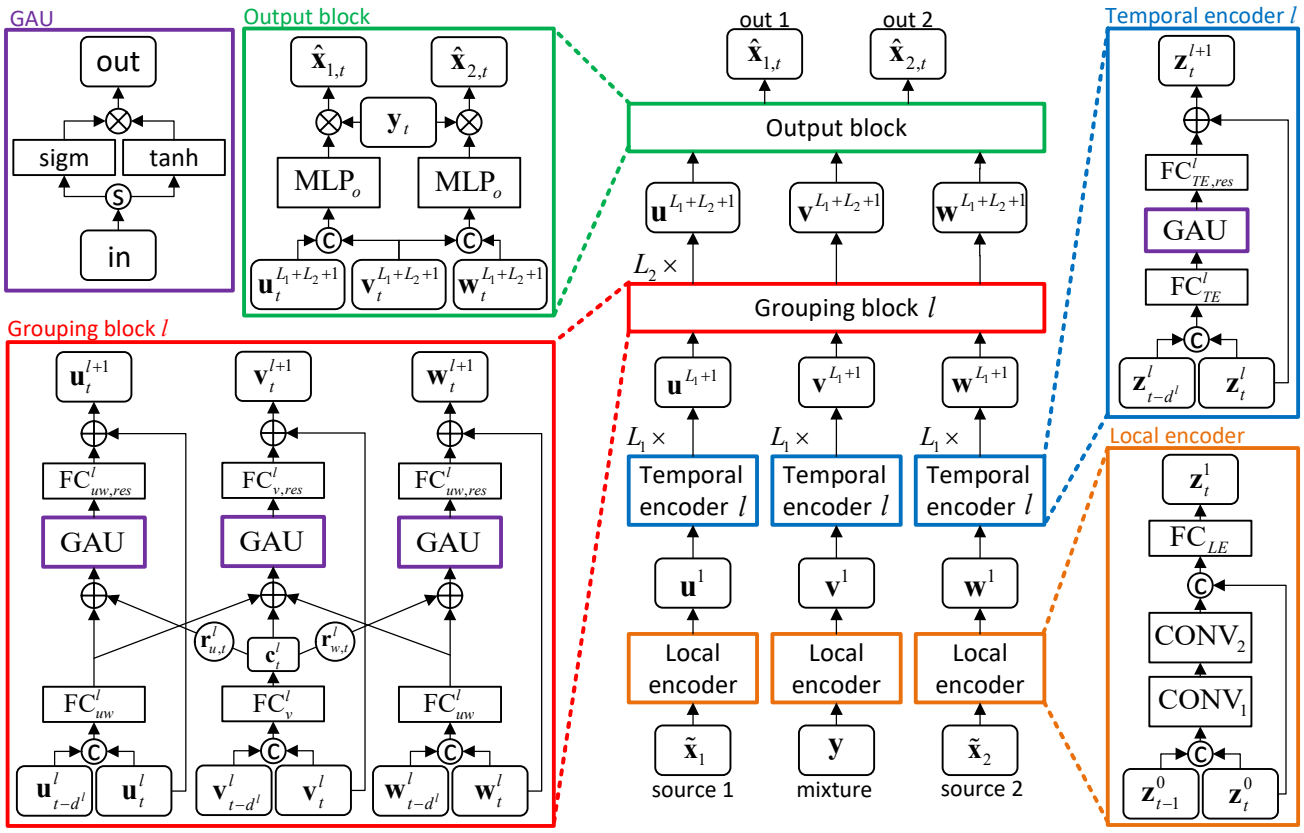

Fig. 2. The architecture of the proposed network for a two-speaker separation task. The listening stage is implemented with three local encoders and three stacks of L_1 temporal encoders, while grouping is performed by a cascade of L_2 grouping blocks and an output block. GAU, FC, MLP, sigm, Ⓒ, Ⓢ, $\otimes$ and $\oplus$ represent Gated Activation Unit [39], full connection, multi layer perceptron, sigmoid activation, concatenation, equally slicing operation, element-wise multiplication and addition respectively. $\hat{\mathbf{r}}$ denotes element-wise masking with routing vector $\mathbf{r}$. $\mathbf{z}_t^0$ represents $\tilde{\mathbf{x}}_{1,t}$, $\tilde{\mathbf{x}}_{2,t}$ or $\mathbf{y}_t$, while $\mathbf{z}_t^l, l = 1, \ldots, L_1 + L_2 + 1$ denotes $\mathbf{u}_t$, $\mathbf{v}_t$ or $\mathbf{w}_t$ respectively, l denotes the layer index of mid-level representations.

Considering the autoregressive nature of our approach, when implementing Eqns. (8-11) with non-causal structures, all previously generated mid-level representations and other intermediate products can be reused during training and inference to greatly reduce the amount of computation while also adapting to online processing conditions.

Conventional RNN structures like LSTM could also be employed as listening and grouping operators, however in Sections IV-A and IV-B we will introduce a novel network structure for our approach which has been designed directly with a CASA motivation in mind. The effectiveness of those structures will be evaluated in Section V-C.

A. Listening

In the experiments presented in this paper, the sampling rate for all waveforms is 8 kHz, from which magnitude spectra of dimension 129 are computed over 32 ms frames with an 8 ms shift between overlapping frames. The network inputs and outputs are μ-law companded [41] magnitude spectra[1] of mixture and estimated source speech.

As described in Eqns. (8-10), in the listening stage, source and mixture spectra $\tilde{\mathbf{x}}_1$, $\tilde{\mathbf{x}}_2$ and $\mathbf{y}$ are individually transformed into mid-level representations $\mathbf{u}$, $\mathbf{w}$ and $\mathbf{v}$ respectively. This paper proposes a structure for an effective listening stage, which consists of two types of module: local encoder and temporal encoder.

1) Local encoder: Local encoders extract T-F features as shown in Fig. 1(b). Represented towards the bottom of Fig. 2, they are designed to capture ASA acoustic cues, which function similarly to feature extraction in a CASA system [3]. The local encoder consists of 2D convolutional layers followed by PReLUs [42] and a fully connected layer[2], which are detailed in Table I. Specifically, two convolutional layers CONV_1 and CONV_2 focus on local temporal-spectral features, while the concatenation operation and the following fully connected layer FC_{LE} enable the local encoder to pay attention to full band spectral features.

2) Temporal encoder: As shown in the centre of Fig. 2, mid-level representations $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{w}$ are respectively extracted by three stacks of L_1 temporal encoders, i.e.,

$$\mathbf{z}_t^{l+1} = f(\mathbf{z}_t^l, \mathbf{z}_{t-d^l}^l), \ l = 1, \ldots, L_1 \tag{12}$$

[1]We performed a number of initial experiments with smaller models which demonstrated that this setting slightly improved performance; perhaps because μ-law companded magnitude spectra, unlike log magnitude spectra, lie in the range of [0, 1], which may be more benign for a feedback-structure model.

[2]In our initial experiments, this was found to perform better than other local encoder structures.

TABLE I
DETAILS OF LOCAL ENCODER. FEATURE MAP SHAPES ARE DENOTED AS (CHANNEL, HEIGHT, WIDTH), D IS THE MID-LEVEL REPRESENTATION DIMENSION. CONVOLUTIONS ARE KERNEL–STRIDE–PAD–CHANNEL.

Operator	Setting	Output feature shape
Input	inputs $\mathbf{z}_{t-1}^0$ and $\mathbf{z}_t^0$ in Fig. 2	(1,2,129)
CONV_1	(2,2)–(1,1)–(0,0)–24	(24,1,128)
CONV_2	(1,5)–(1,3)–(0,0)–48	(48,1,42)
©	reshape and concatenate	(2145,1)
FC_{LE}	fully connected layer	(D,1)

where $\mathbf{z}_t^l$ represents $\mathbf{u}_t^l$, $\mathbf{v}_t^l$ or $\mathbf{w}_t^l$, l denotes the layer index of mid-level representations, L_1 represents the number of temporal encoders in each stack, and d^l is the dilation factor for temporal encoder l. Inspired by WaveNet [43], the temporal encoder comprises dilated convolution and gated activation unit (GAU) [39] to model the temporal context of each source and mixture. As long as L_1 and d^l of all temporal encoders are known, the receptive field length in the listening stage can be determined by $P_1 = \sum_{l=1}^{L_1} d^l + 2$.

B. Grouping

To implement Eqn. (11) and perform the conditional grouping shown in Fig. 1(b), the proposed structure for the grouping stage includes an output block and a stack of L_2 grouping blocks, which are illustrated in Fig. 2.

1) Grouping block: As shown towards the top of the network in Fig. 2, with details provided in the bottom left of the figure, each grouping block l, imports mid-level representations $\mathbf{u}^l$, $\mathbf{v}^l$ and $\mathbf{w}^l$ from the previous layer and generates $\mathbf{u}^{l+1}$, $\mathbf{v}^{l+1}$ and $\mathbf{w}^{l+1}$ for the next layer. Grouping blocks are designed with the main consideration that the separation of mixtures will benefit from the estimation of sources, and vice versa. Using a conditioning method similar to [43], $\mathbf{u}_t^{l+1}$, $\mathbf{v}_t^{l+1}$ and $\mathbf{w}_t^{l+1}$ can be considered as outputs given conditions $\{\mathbf{v}_t^l, \mathbf{v}_{t-d^l}^l\}$, $\{\mathbf{u}_t^l, \mathbf{u}_{t-d^l}^l, \mathbf{w}_t^l, \mathbf{w}_{t-d^l}^l\}$ and $\{\mathbf{v}_t^l, \mathbf{v}_{t-d^l}^l\}$ respectively, i.e.,

$$\mathbf{u}_t^{l+1} = g(\mathbf{u}_t^l, \mathbf{u}_{t-d^l}^l | \mathbf{v}_t^l, \mathbf{v}_{t-d^l}^l) \tag{13}$$

$$\mathbf{w}_t^{l+1} = g(\mathbf{w}_t^l, \mathbf{w}_{t-d^l}^l | \mathbf{v}_t^l, \mathbf{v}_{t-d^l}^l) \tag{14}$$

$$\mathbf{v}_t^{l+1} = h(\mathbf{v}_t^l, \mathbf{v}_{t-d^l}^l | \mathbf{u}_t^l, \mathbf{u}_{t-d^l}^l, \mathbf{w}_t^l, \mathbf{w}_{t-d^l}^l) \tag{15}$$

where d^l is the temporal dilation factor for grouping block l. As with the temporal encoder, the receptive field length in the grouping stage can be determined by $P_2 = \sum_{l=L_1+1}^{L_2} d^l + 1$, where L_2 denotes the number of grouping blocks. Therefore, the total receptive field length in listening and grouping stages is $P = P_1 + P_2 - 1 = \sum_{l=1}^{L_2} d^l + 2$.

As we can see in Fig. 2, the structure of grouping blocks can be considered as the combination of three parallel temporal encoders with two cross conditioning connections. In this paper, a routing strategy is adopted to control the conditioning effect in Eqns. (13-14). The original condition vector $\mathbf{c}_t^l$ for $\mathbf{u}$ and $\mathbf{w}$ activations, which are generated by FC_v^l, are masked by two routing vectors $\mathbf{r}_{u,t}^l$ and $\mathbf{r}_{w,t}^l$ respectively. Specifically, masked condition vectors $\mathbf{c}_{u,t}^l$ and $\mathbf{c}_{w,t}^l$ are obtained according to Eqns. (16-23). At each time step t in grouping block l,

$$\bar{\mathbf{u}}_t = \mathrm{GRU}(\mathbf{u}_t^{L_1+1}, \bar{\mathbf{u}}_{t-1}) \tag{16}$$

$$\bar{\mathbf{w}}_t = \mathrm{GRU}(\mathbf{w}_t^{L_1+1}, \bar{\mathbf{w}}_{t-1}) \tag{17}$$

$$\mathbf{s}_{u,t}^l = \alpha \cdot \tanh(W_r^l[\mathbf{c}_t^l, \bar{\mathbf{u}}_t]/\alpha) \tag{18}$$

$$\mathbf{s}_{w,t}^l = \alpha \cdot \tanh(W_r^l[\mathbf{c}_t^l, \bar{\mathbf{w}}_t]/\alpha) \tag{19}$$

$$\mathbf{r}_{u,t}^l = \frac{\exp(\mathbf{s}_{u,t}^l)}{\exp(\mathbf{s}_{u,t}^l) + \exp(\mathbf{s}_{w,t}^l)} \tag{20}$$

$$\mathbf{r}_{w,t}^l = \frac{\exp(\mathbf{s}_{w,t}^l)}{\exp(\mathbf{s}_{u,t}^l) + \exp(\mathbf{s}_{w,t}^l)} \tag{21}$$

$$\mathbf{c}_{u,t}^l = \mathbf{c}_t^l \otimes \mathbf{r}_{u,t}^l \tag{22}$$

$$\mathbf{c}_{w,t}^l = \mathbf{c}_t^l \otimes \mathbf{r}_{w,t}^l \tag{23}$$

where $\bar{\mathbf{u}}_t$ and $\bar{\mathbf{w}}_t$ are outputs of gated recurrent unit (GRU) [44] for mid-level representations $\mathbf{u}^{L_1+1}$ and $\mathbf{w}^{L_1+1}$ in the listening stage. $\mathbf{s}_{u,t}^l$ and $\mathbf{s}_{w,t}^l$ are corresponding routing scores (or energy) computed by weight matrix W_r^l, $[\cdot]$ and $\otimes$ represent concatenation and element-wise multiplication respectively, α is a scalar defining the range of elements in $\mathbf{s}_{u,t}^l$ and $\mathbf{s}_{w,t}^l$ as $(-\alpha, \alpha)$, empirically set to 5 in our experiments.

From the attention mechanism perspective, this routing strategy performs additive attention [44] to every dimension of $\mathbf{c}_t^l$ at each time step t, with $\mathbf{c}_t^l$ as value, and $\bar{\mathbf{u}}_t$, $\bar{\mathbf{w}}_t$ as queries. Considering the fact that the corresponding elements of $\mathbf{r}_{u,t}^l$ and $\mathbf{r}_{w,t}^l$ lie in the range of $(0, 1)$ and they sum up to 1, routing vector $\mathbf{r}_{u,t}^l$ can be regarded as a dimension-wise probability distribution of the $\mathbf{u}$ component in $\mathbf{c}_t^l$, while $\mathbf{r}_{w,t}^l$ corresponds to the $\mathbf{w}$ component.

2) Output block: Given outputs of the final grouping block $\mathbf{u}_t^{L_1+L_2+1}$, $\mathbf{v}_t^{L_1+L_2+1}$ and $\mathbf{w}_t^{L_1+L_2+1}$, the output block generates estimated source spectra $\hat{\mathbf{x}}_{1,t}$ and $\hat{\mathbf{x}}_{2,t}$ using a multi layer perceptron (MLP) structure equipped with PReLUs, as shown near the top of Fig. 2. In experiments, the MLP structure is used to generate amplitude masks for estimation. The computation of $\hat{\mathbf{x}}_{1,t}$ and $\hat{\mathbf{x}}_{2,t}$ is defined as follows;

$$\hat{\mathbf{x}}_{1,t} = \mathbf{y}_t \otimes \mathrm{MLP_o}([\mathbf{u}_t^{L_1+L_2+1}, \mathbf{v}_t^{L_1+L_2+1}]) \tag{24}$$

$$\hat{\mathbf{x}}_{2,t} = \mathbf{y}_t \otimes \mathrm{MLP_o}([\mathbf{w}_t^{L_1+L_2+1}, \mathbf{v}_t^{L_1+L_2+1}]) \tag{25}$$

where $\mathrm{MLP_o}(\cdot)$ denotes the function of the MLP structure.

It is worth mentioning that, with respect to sources $\mathbf{x}_1$ and $\mathbf{x}_2$, the structure is completely symmetric and network parameters are all shared. This characteristic conforms to the common sense that all positions of each source are equivalent and exchangeable. Moreover, this design avoids model size growth when source number increases.

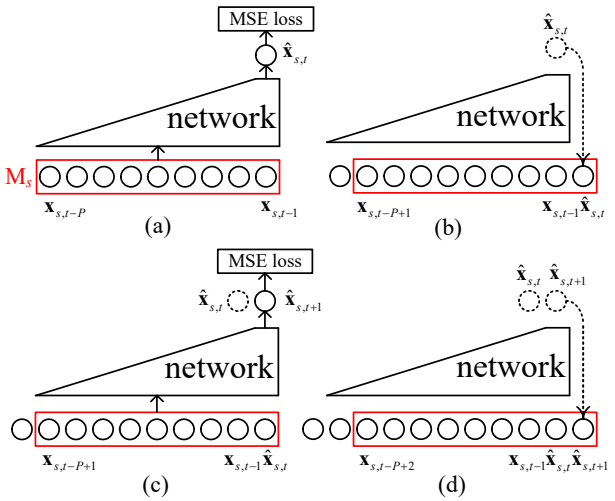

Fig. 3. The MPT strategy, showing only source s for clarity. Each circle represents one target or estimated source spectrum frame. $\mathbf{M}_s$ is a queue of fixed length P that stores source s inputs for each time step. (a) to (d) show two time steps t and $t+1$ during training. (a) $\mathbf{M}_s$ is initialized with all target spectra, and estimated spectrum $\hat{\mathbf{x}}_{s,t}$ is generated by the network. (b) The last item $\mathbf{x}_{s,t-P}$ is popped out from $\mathbf{M}_s$, and $\hat{\mathbf{x}}_{s,t}$ is pushed to the front of $\mathbf{M}_s$. (c) $\hat{\mathbf{x}}_{s,t+1}$ is generated from the updated $\mathbf{M}_s$. (d) $\mathbf{x}_{s,t-P+1}$ is popped out from $\mathbf{M}_s$, and $\hat{\mathbf{x}}_{s,t+1}$ is pushed to the front of $\mathbf{M}_s$, and the process continues. The mixture input and other source are handled similarly.

C. Multi-time-step prediction training

Using a conventional training method such as [43], a mismatch problem would arise between training and inference stages. During training, source inputs $\tilde{\mathbf{x}}_{s,<t}$ in Eqns. (6-7) are target spectra $\mathbf{x}_{s,<t}$, whereas in the inference stage, they change to estimated spectra $\hat{\mathbf{x}}_{s,<t}$. The error between the two spectra leads to a mismatch.

To alleviate this, we develop a multi-time-step prediction training (MPT) strategy. During training, in each mini batch the network does look-ahead prediction over T sequential time steps. After each spectrum is predicted, it is fed back as the next estimated source input. Fig. 3 illustrates the procedure at time steps t and $t+1$. At the first time step t, the input source spectra are initialized with all target spectra. These will gradually be replaced by estimated spectra as prediction proceeds. When multi-time-step prediction has finished, and T estimated spectra of each source have been generated, the loss gradients across all time steps are back-propagated as a batch. This enforces the network to exploit temporal context constraints on the estimated sources. The loss function is defined as the averaged MSE between targets and corresponding estimated source spectra:

$$O_t = \frac{1}{TF}\sum_{t'=0}^{T-1}\sum_{s=1}^{S}\|\mathbf{x}_{s,t+t'} - \hat{\mathbf{x}}_{s,t+t'}\|_2^2 \tag{26}$$

where S and F are source number and the dimension of spectrum, $\|\cdot\|_2$ is the L_2 norm.

V. Experiments

We evaluate separation performance with various settings in terms of average signal-to-distortion ratio (SDR) [45] improvement between separated speech and mixture – a widely adopted metric in multi-speaker speech separation research.

A. Experimental Settings

The WSJ0-2mix dataset, introduced in [16] and derived from the WSJ0 corpus [46], is adopted for our evaluations. It comprises a 30-hour training set and a 10-hour validation set of two-speaker mixtures generated by utterances randomly selected from the WSJ0 training set `si_tr_s`, mixed at various signal-to-noise ratios (SNR) between 0 dB and 10 dB. A 5-hour test set is similarly generated using utterances from 16 unseen speakers in the WSJ0 development set `si_dt_05` and evaluation set `si_et_05`. The validation set and the test set are used to evaluate separation performance for closed condition (CC) and open condition (OC) tests respectively, which is similar to [15], [16], [21].

Both conventional LSTM and the proposed network structures for listening and grouping (LG) stages are evaluated in our experiments. The structure details are as follows:

- **LG-Listen** is the proposed listening structure introduced in Section IV-A, comprising three local encoders and three stacks of $L_1 = 5$ temporal encoders. The dimension D of mid-level representations $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{w}$ is 256, and dilation factors for temporal encoders are $[d^1, \ldots, d^5] = [1, 2, 4, 8, 16]$, giving $P_1 = 33$.
- **LSTM-Listen** is composed of three stacks of LSTM layers, which are used to implement the listening stage described in Eqns. (8-10) respectively, i.e. $\mathcal{L}(\cdot) = \mathrm{LSTM}(\cdot)$. Each stack has 2 LSTM layers with 352 hidden units in each layer, and the output linearly transformed to $D = 256$ dimensions.
- **LG-Group** is the proposed grouping structure described in Section IV-B, comprising $L_2 = 5$ stacked grouping blocks and an output block. The dimension D of mid-level representations and the dilation factors are the same as those in **LG-Listen** and $P_2 = 32$.
- **LSTM-Group** comprises 2 LSTM layers with 480 hidden units in each layer, and a fully-connected layer to generate estimated source spectra $\hat{\mathbf{x}}_1$ and $\hat{\mathbf{x}}_2$. It is designed to implement the grouping stage described in Eqn. (11), given the concatenation of current mid-level representations $\mathbf{u}_t$, $\mathbf{v}_t$ and $\mathbf{w}_t$ from the listening stage as input, i.e. $\mathcal{G}(\mathbf{u};\mathbf{w};\mathbf{v}) = \mathrm{LSTM}([\mathbf{u}_t, \mathbf{w}_t, \mathbf{v}_t])$, where $[\cdot]$ represents concatenation.

For fair comparison, the number of parameters in **LG-Listen** and **LSTM-Listen** are matched at approximately 2.6 million each, while **LG-Group** and **LSTM-Group** are matched with about 5.6 million parameters each.

All networks are implemented using MXNet [47] and are optimized within 100 epochs using the Adam algorithm [48] with fixed batch size 256 and initial learning rate 0.001. Learning rate adjustment and early stopping strategies are adopted by observing SDR results on the validation set. No further regularization or training strategies are used.

Separated waveforms can be reconstructed from estimated sources spectra, using either original mixture phase, or the

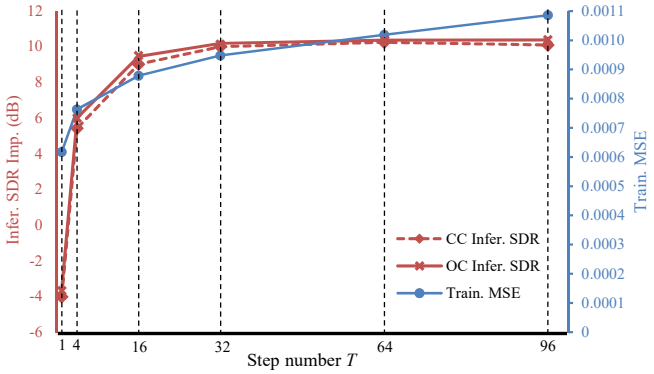

Fig. 4. Training stage MSE and inference stage SDR improvements (dB) in CC and OC conditions for the proposed structures with different step number T in MPT strategy. T=1 denotes a conventional training method.

TABLE II
SDR IMPROVEMENT (dB) IN CC AND OC CONDITIONS AND APPROXIMATE MODEL SIZES (NUMBER OF PARAMETERS) FOR VARIOUS METHODS, INCLUDING CONVENTIONAL LSTM, THE PROPOSED APPROACH AND OTHER STATE-OF-THE-ART ONLINE METHODS.

Network structure		Model size (million)	SDR Imp. CC	SDR Imp. OC
LSTM-Listen	+ LSTM-Group	8.2	7.9	8.0
LG-Listen	+ LSTM-Group	8.2	8.0	8.2
LSTM-Listen	+ LG-Group	8.2	9.6	9.8
LG-Listen	+ LG-Group	8.2	**10.3**	**10.4**
uPIT-LSTM-PSM [15]		65.7	7.0	7.0
TasNet-LSTM [24]		31.0	–	8.0
TasNet-LSTM-50% [34][3]		32.0	–	11.2
Conv-TasNet-BN [35][3]		8.8	–	11.2

phase retrieved using the MISI algorithm [33] for better performance. However, MISI requires the whole mixture utterance to be used as input for reconstruction, which is incompatible with low-latency or online implementation. We therefore reformulate MISI into a real-time (RT) algorithm which we denote RTMISI. This is inspired by RTISI [49], and retrieves source phase in time sequential order (frame-by-frame) without the need for future mixture information. However the use of some future information is beneficial to performance, and so we enable RTMISI to trade-off between latency and separation performance by allowing a limited number look-ahead frames. With a short frame overlap, processing is still online, but benefits from the increased temporal scope.

In the following sections, first we investigate the effect of MPT introduced in Section IV-C. We next compare between LSTM and the proposed structures before separately investigating the effects of listening to sources and grouping. Finally, our proposed approach with different phase retrieval algorithms is compared to other state-of-the-art approaches.

B. Effect of MPT strategy

We first investigate the effect of MPT strategy for the proposed structure **LG-Listen+LG-Group** in terms of training stage MSE and inference stage SDR improvements (dB) for both CC and OC. Separate waveforms are reconstructed with estimated sources spectra and original mixture phase. From Section IV-B, we can see that the total receptive field length of **LG-Listen+LG-Group** is P=64. The results of conventional training and MPT with different step numbers T in Section IV-C are compared in Fig. 4.

We can see immediately that although conventional training (T=1) obtains the lowest training stage MSE (about 0.0006), the inference stage SDR improvements in both CC and OC conditions are poor (i.e. approximately -4 dB in both CC and OC conditions), indicating a large mismatch between training and inference stages. When T increases, due to the MPT strategy, the training MSE rises moderately. This is reasonable since estimating a sequence of source spectra requires the network to learn the continuities and dependencies within estimated sources, which is more difficult than estimating an individual frame. However, inference SDRs in both CC and OC conditions are significantly improved and gradually converge to approximately 10 dB, suggesting that MPT mitigates the mismatch between training and inference stages. Moreover, the best SDR results can be observed at T=64, which is equal to P. Under this condition, input source spectra consist of 50% of target spectra and 50% of estimated spectra. This result suggests that an appropriate ratio of target and estimated source inputs may be essential for training. Finally, comparable or higher SDRs for OC compared to CC indicate the approach generalizes well for unseen speakers.

In summary, the MPT strategy is successful at alleviating the mismatch between training and inference stages. We set T=64 for the following experiments.

C. Comparison of Different Structures

In this section, conventional LSTM (**LSTM-Listen**, **LSTM-Group**) and the proposed structures (**LG-Listen**, **LG-Group**) are evaluated. By combining different structures in listening and grouping stages separately, there are four settings in total for comparison. The SDR improvements of all settings in CC and OC conditions are shown in Table II, where separate waveforms are reconstructed with estimated source spectra and original mixture phase. Other state-of-the-art online methods[3] are also included in Table II.

Firstly, it can be seen that even with conventional LSTM structures, our approach achieves comparable or higher SDRs (e.g., 8.0 dB SDR in OC conditions for **LSTM-Listen** + **LSTM-Group**) than other previously reported state-of-the-art networks uPIT-LSTM-PSM [15] and TasNet-LSTM [24], demonstrating the effectiveness of the approach. In addition, we can compare results for different grouping structures with the same listening structure, e.g., **LG-Listen**. The SDR gaps between **LSTM-Group** and **LG-Group**, e.g., from 8.2 dB to 10.4 dB in OC conditions with **LG-Listen**, indicate that our proposed grouping structure performs the grouping task better than **LSTM-Group**. Meanwhile, a similar trend can be observed by comparing different listening structures with the same grouping structure. It is worth noting that the proposed grouping structure yields more significant improvements

[3]TasNet-LSTM-50% [34] and Conv-TasNet-BN [35] are post-submission revisions of [24]. We include these results for fair comparison.

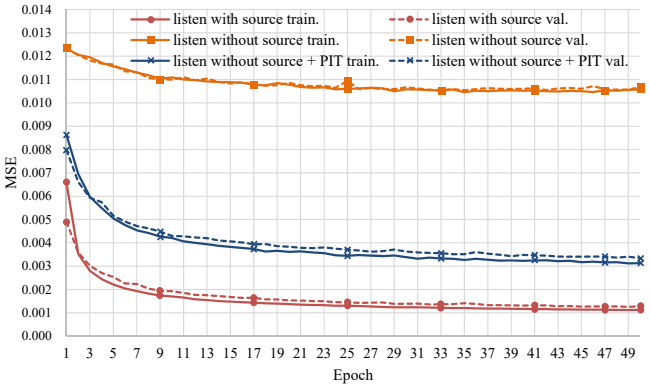

Fig. 5. MSE over epochs on training (solid lines) and validation (dotted lines) sets for three different arrangements described in Section V-D: Listening with sources, Listening without sources and Listening without sources + PIT.

compared to LSTM than the proposed listening structure. For example, in OC conditions, the averaged SDR difference between **LG-Group** and **LSTM-Group** is 2.0 dB, higher than that between **LG-Listen** and **LSTM-Listen**, which is 0.4 dB. This may be due to the fact that the grouping stage is more important to overall performance than the listening stage, so it demands more powerful structures. Finally, it can be seen that using both the proposed structures in listening and grouping stages achieves the highest SDR improvements among other previously reported online methods.

In summary, these results demonstrate the effectiveness of our approach compared with other state-of-the-art online methods. Meanwhile, results demonstrate the proposed structures in listening and grouping stages may both outperform conventional LSTMs for a limited number of parameters, particularly when operating in conjunction with each other. Therefore, in the following experiments we focus on the proposed listening and grouping structures, denoted as **LG**.

D. Effect of Listening to Sources

In our approach, listening to the mixture is obviously necessary for separation, but the effect of listening to sources still needs further justification. Therefore, we construct an experiment on three different arrangements for comparison:

1) **Listening with sources** represents the proposed listening and grouping approach, where the network **LG** listens to not only the mixture, but also separated sources.
2) **Listening without sources** sets the former network to listen only to the mixture but not sources; no separate source feedback is provided in the grouping stage. The loss function for training is conventional MSE.
3) **Listening without sources + PIT** originates from **Listening without sources**, but the difference is that a PIT modified MSE following utterance level PIT [15] is adopted in place of conventional MSE.

Each arrangement has the same batch size and initial learning rate, with step numbers $T = 64$ for MPT. The training progress of each setting, measured by MSE (or PIT modified MSE) on training and validation sets, is presented in Fig. 5.

We can clearly see that **Listening without sources** barely reduces either training or validation MSE (by about 0.106), which is mainly due to the label permutation problem mentioned in Section III-A. By contrast, incorporating the PIT technique, **Listening without sources + PIT** enables training and validation MSE to converge to a relatively low level, around 0.0030 and 0.0033 respectively. This observation is consistent with [15]. Meanwhile, by listening to separated sources, both the training and validation MSE in **Listening with sources** steadily decrease through training epochs, with final results of about 0.0011 for the training set and 0.0013 for the validation set. This implies that listening to sources can effectively address the label permutation problem. Moreover, by comparing **Listening without sources + PIT** and **Listening with sources**, we can see that listening to sources converges faster and finally achieves considerably lower training and validation MSEs than the PIT technique. This may derive from the fact that PIT attempts to obtain a constant output permutation based on separation error [15], but listening to sources enforces output permutation to be the same as the input sources, which encourages the network to exploit the temporal context of the source signal in an autoregressive manner. This allows listening to sources to provide a more effective constraint than the PIT technique. Moreover, as training gradually converges, listening to sources will provide more precise and useful information from sources for the grouping stage, in addition to that from the original mixture signal, which is beneficial for grouping to model the interaction of mixture and source signals. This is an important advantage of our listening and grouping approach, compared to most existing deep learning approaches formulated as Eqn. (5).

In summary, listening to sources is essential for our approach to address the label permutation problem and exploit the temporal context of source signals.

E. Effect of Grouping

In our approach, the grouping stage is designed to generate estimated sources spectra given mid-level representations from the listening stage. In this section, to investigate the effect of grouping, we visualize the corresponding spectra and mid-level representations from network **LG** using t-distributed stochastic neighbor embedding (t-SNE) [50]. Fig. 6 shows the corresponding two-dimensional t-SNE results of mixture spectra $\mathbf{y}$, sources spectra $\mathbf{x}_1$, $\mathbf{x}_2$ and mid-level representations $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$ in listening and grouping stages from one male-female mixture utterance in the test set. The perplexity for t-SNE is set to 30, and in Fig. 6 each point represents one frame of spectrum or mid-level representation respectively.

Firstly, from Fig. 6(a), we can see that the t-SNE distributions of mixture $\mathbf{y}$ and two target sources spectra $\mathbf{x}_1$, $\mathbf{x}_2$ substantially overlap, suggesting similarity between mixture and sources in original spectral space. Meanwhile, in Fig. 6(b) the final output mid-level representations $\mathbf{v}$, $\mathbf{u}$ and $\mathbf{w}$ in the listening stage, which correspond to mixture $\mathbf{y}$ and two sources $\tilde{\mathbf{x}}_1$, $\tilde{\mathbf{x}}_2$ respectively and are formulated in Eqns. (8-10), are still mixed together, indicating that the listening stage does not perform a separation operation on the mixture. However, as shown

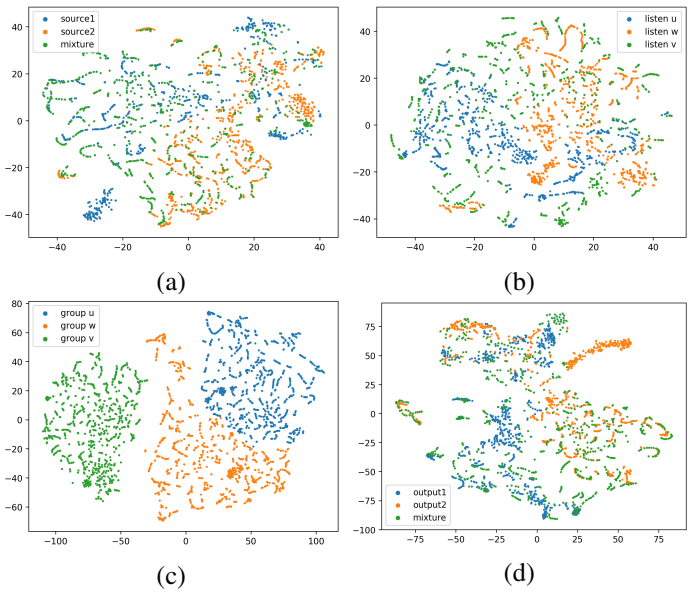

Fig. 6. Visualizations of two-dimensional t-SNE [50] results of mixture spectra $\mathbf{y}$, sources spectra $\mathbf{x}_1$, $\mathbf{x}_2$ and intermediate representations $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$ in listening and grouping stages from one utterance in the test set. Each point represents one frame. (a) Mixture and target sources spectra. (b) Final output mid-level representations from the listening stage. (c) Intermediate representations from the middle of the grouping stage. (d) Mixture and output sources spectra.

in Fig. 6(c), the mid-level representations $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{w}$ from the third grouping block have significantly different distributions – three well-separated and well-grouped sets. Moreover, it can be clearly seen that the interval between mixture and sources representations are much greater than that between two sources, suggesting that the mixture representations are more dissimilar than the source representations. Finally, estimated sources spectra are reconstructed given well separated and grouped representations from grouping blocks, but in Fig. 6(d), it can be observed that mixture and estimated sources spectra overlap again – a similar distribution to that of mixture and target source spectra.

In addition to the autoregressive and online processing nature of our approach described in Section I, in Fig. 6 we can find another important difference between our approach and other state-of-the-art deep learning methods, in the grouping stage. DPCL and DANet perform grouping by clustering T-F bin embeddings for each speaker category according to the distance in embedding space, which encourages embeddings to focus on speaker difference. On the other hand, PIT usually focuses on the spectral structure difference between target and estimated sources. However, compared to those methods, our approach has two characteristics – it not only pays attention to the differences between mixture and sources implicitly, it also preserves spectral information in mixture and source signals. The first characteristic can be observed from the distribution of three well-separated sets of representations in Fig. 6(c). Meanwhile, the second characteristic allows the network to successfully reconstruct estimated source spectra, indicated from the short curves formed by several $\mathbf{u}$ or $\mathbf{w}$ points, corresponding to continuous spectral structures in the signals. In summary, Fig. 6 reveals that the grouping stage firstly separates and groups mixture and sources representations respectively, then reconstructs sources spectra using well separated and grouped representations, which is similar to simultaneous and sequential grouping in CASA systems [3].

TABLE III
SDR IMPROVEMENTS (dB) AND APPROXIMATE MODEL SIZES (IN TERMS OF PARAMETER NUMBER ESTIMATED ACCORDING TO THE PAPERS) OF VARIOUS SYSTEMS IN OC AND CC CONDITIONS ON WSJ0-2MIX DATASET.

Method	Model Size (million)	SDR Imp. CC	SDR Imp. OC	Comments
Oracle NMF [16]	–	5.1	–	Conventional approaches
CASA [16]	–	2.9	3.1	
DPCL+ [17]	10.6	–	9.4	Offline approaches
DPCL++ [17]	16.9	–	10.8	
DANet-6 anchor-BLSTM [21]	8.3	–	10.8	
PIT-CNN-51\51 [15]	–	7.6	7.5	
uPIT-BLSTM-PSM [15]	46.4	9.4	9.4	
uPIT-BLSTM-PSM-ST [15]	94.6	10.0	10.0	
CASA-E2E [32]	54.3	–	11.0	
DC+MI+MISI [18]	29.6	11.4	11.5	
WA-MISI-5 [19]	29.6	13.2	13.1	
TasNet-BLSTM [24]	22.5	–	11.1	
TasNet-BLSTM-50% [34]??	23.6	–	13.6	
Conv-TasNet-gLN [35]??	8.8	–	15.0	
LG MISI[4]	8.2	13.1	13.0	
uPIT-LSTM-PSM [15]	65.7	7.0	7.0	Online approaches
TasNet-LSTM [24]	31.0	–	8.0	
TasNet-LSTM-50% [34]??	32.0	–	11.2	
Conv-TasNet-BN [35]??	8.8	–	11.2	
Source-Aware network [36]	7.2	9.3	9.5	
LG mixture phase	8.2	10.3	10.4	
LG RTMISI look-ahead 0ms	8.2	11.1	11.0	
LG RTMISI look-ahead 8ms	8.2	12.5	12.4	
LG RTMISI look-ahead 16ms	8.2	13.0	12.9	
LG RTMISI look-ahead 24ms	8.2	**13.1**	**13.0**	

F. Performance Comparison of Various Approaches

Table III summarizes SDR improvements (dB) in CC and OC conditions and approximate model size (in terms of estimated number of parameters according to the papers) for different approaches with similar or comparable experimental settings on WSJ0-2mix dataset. As described in Section II, we can divide approaches into three categories for comparison: conventional, offline and online deep learning approaches.

Firstly, we can see that offline deep learning approaches outperform conventional approaches in terms of SDR improvements. For instance, WA-MISI-5 [19] combines DPCL, PIT and MISI techniques and achieves 13.2 and 13.1 dB SDR in CC and OC conditions with an end-to-end training structure. However, the performance of online deep learning approaches still have a large gap compared to offline approaches. For example, TasNet-LSTM-50% [34] directly models the waveform domain using LSTM and PIT techniques and achieves 11.2 dB SDR in the OC condition, which is 2.4 dB lower than TasNet-BLSTM-50% [34]. As described in Section V-A, mixture phase, MISI and RTMISI algorithms are applied respectively for our approach. Using original mixture

[4]LG MISI uses the same magnitude spectra estimated online from LG mixture phase, and the whole mixture utterance is only used for phase retrieval, which is different from those offline approaches listed above.

TABLE IV
SDR IMPROVEMENTS (dB), PESQS AND STOIS WITH RESPECT TO DIFFERENT GENDER COMBINATIONS AND OVERALL PERFORMANCE IN OC CONDITIONS FOR LG RTMISI LOOK-AHEAD 24MS ON WSJ0-2MIX.

Gender info.	Male-male	Male-female	Female-female	Overall
SDR Imp.	12.889	14.840	12.753	13.003
PESQ	3.282	3.457	3.212	3.324
STOI	0.954	0.968	0.938	0.956

TABLE V
SDR IMPROVEMENTS (dB), PESQS AND STOIS WITH RESPECT TO DIFFERENT INPUT MIXTURE SNR LEVELS (dB) IN OC CONDITIONS FOR LG RTMISI LOOK-AHEAD 24MS ON WSJ0-2MIX.

SNR levels	0∼2.5	2.5∼5.0	5.0∼7.5	7.5∼ ∞
SDR Imp.	12.633	12.969	13.404	14.970
PESQ	3.238	3.291	3.328	3.411
STOI	0.945	0.953	0.956	0.960

phase, our network **LG** obtains 10.3 and 10.4 dB SDR in CC and OC conditions. Furthermore, when using various MISI and RTMISI settings, **LG** achieves up to 13.1 and 13.0 dB SDR in CC and OC conditions given the same output spectra, which is a significant boost (about 2.6 dB) compared with mixture phase. Meanwhile, these results reveal that our approach has achieved the highest SDRs over state-of-the-art online approaches, even outperforming the majority of offline deep learning approaches, and is only beaten by WA-MISI-5 [19] and the most recently reported results from TasNet-BLSTM-50% [34] and Conv-TasNet-gLN [35]. Finally, due to the symmetric structure and shared parameters settings, our network has the fewest parameters (about 8.2 million) among the compared models, apart from our previously proposed source-aware context network [36].

To further investigate the separation performance of our approach, we report SDR improvement (dB), perceptual evaluation of speech quality (PESQ) [51] and short-time objective intelligibility (STOI) [52] with respect to different gender combinations and input SNR levels respectively in OC conditions for model LG RTMISI look-ahead 24ms. Firstly, separation performance with respect to different gender combination and overall performance across all combinations are reported in Table IV. From this table, we can clearly see that our approach achieves much better SDR, PESQ and STOI on male-female combinations than same gender conditions. For example, the SDR of male-female speech is approximately 2 dB higher than male-male or female-female combinations. These results agree with the observation from some other works [15], [16], [28], [32], and indicate that same gender mixed speech separation is often a harder task. Secondly, Table V reports these metrics with respect to different input mixture SNR levels (dB), which are divided into four categories: $0 \leq \text{SNR} < 2.5$, $2.5 \leq \text{SNR} < 5.0$, $5.0 \leq \text{SNR} < 7.5$ and $7.5 \leq \text{SNR}$. From Table V, we can clearly observe that SDR improvements, PESQ and STOI all increase steadily with the increase in SNR. These results indicate that, for our approach, input mixtures with higher SNR levels may be easier to separate than low SNR levels, which is similar to some other works [28], [53], [54].

TABLE VI
SDR IMPROVEMENTS (dB) AND PESQ IN OC CONDITIONS FOR VARIOUS SYSTEMS EVALUATED ON THE WSJ0-2MIX DATASET.

Methods	SDR Imp.	PESQ	Comments
uPIT-BLSTM-PSM [15]	9.4	2.63	Offline approaches
DANet-6 anchor-BLSTM [21]	10.8	2.82	
WA-MISI-5 [19]	13.1	–	
TasNet-BLSTM-50% [34], [35]	13.6	3.04	
Conv-TasNet-gLN [35]	15.0	3.25	
TasNet-LSTM-50% [34], [35]	11.2	2.84	Online approaches
Conv-TasNet-BN [35]	11.2	2.86	
LG RTMISI look-ahead 24ms	**13.0**	**3.32**	
Mixture	0.0	2.01	

Finally, reported SDR improvements (dB) and PESQs in OC conditions for various systems are summarized in Table VI. From this table, we can see that the proposed approach achieves highest SDR improvement among state-of-the-art online approaches, only lower than WA-MISI-5 [19] and most recently reported offline models TasNet-BLSTM-50% [34] and Conv-TasNet-gLN [35]. Moreover, our model outperforms reported state-of-the-art offline model Conv-TasNet-gLN in terms of PESQ (about 0.07 absolute improvement). Compared to TasNet models [24], [34], [35], our approach focuses on the magnitude spectral domain, in which PESQ is measured, while TasNet models use SI-SNR or sample-level MSE as training objectives in the sample domain, both closely related to the SDR metric. This may explain the different SDR and PESQ trends in Table VI. In the future, it would be interesting to explore sample domain modeling, which is likely to provide better performance in terms of SDR.

VI. CONCLUSION AND FUTURE WORK

In this paper, we have proposed an online autoregressive approach for monaural multi-speaker speech separation in an explicit listening and grouping architecture. Our approach jointly exploits causal temporal context information in both mixture and past estimated sources signals, which can address the label permutation problem and meet online requirements. Meanwhile, we have proposed a specific network structure to take full advantage of dependency and interaction of mixture and sources. An MPT strategy is also developed to alleviate mismatch between training and inference stages, and the RTMISI algorithm is implemented for phase retrieval to improve waveform reconstruction. Experimental results on the benchmark WSJ0-2mix dataset reveal that the MPT strategy and the RTMISI algorithm enable the proposed approach to outperform the majority of online and offline state-of-the-art methods in terms of SDR improvement and PESQ in both CC and OC conditions, while having relatively fewer model parameters.

This approach can be extended to non-causal configuration, where future mixture information is utilized to improve separation performance. One possibility is to make the "listening to mixture" stage (formulated in Eq. (10)) non-causal, while keeping other stages unchanged. To make use of future mixture information, in this "listening to mixture" stage, mixture mid-level representations can be extracted from a sequence of past,

current and future mixture spectra using a non-causal structure, e.g., BLSTM, non-causal convolutional layer or other CNN-RNN hybrids. Our approach can also be generalized to more than two sources; more local encoders and temporal encoders could be employed to extract mid-level representations independently for additional sources. Meanwhile, since source representations are only directly connected to mixture representations in the grouping block, more connections between source and mixture representations could be established according to the number of sources. The output block could be extended similarly. Finally, the performance boost by using MISI and RTMISI indicates a potential improvement from more powerful network structures that can directly model the complex relationship between mixture and source waveforms.

REFERENCES

[1] D. Yu and J. Li, "Recent progresses in deep learning based acoustic models," *IEEE/CAA Journal of Automatica Sinica*, vol. 4, no. 3, pp. 396–409, 2017.

[2] A. S. Bregman, *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1990.

[3] G. J. Brown and M. Cooke, "Computational auditory scene analysis," *Computer Speech & Language*, vol. 8, no. 4, pp. 297–336, 1994.

[4] D. Wang and G. J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications*. Wiley-IEEE press, 2006.

[5] G. Hu and D. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 2067–2079, 2010.

[6] K. Hu and D. Wang, "An unsupervised approach to cochannel speech separation," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 21, no. 1, pp. 122–131, 2013.

[7] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *Ninth Int. Conf. on Spoken Language Processing*, 2006.

[8] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 1–12, 2007.

[9] M. Stark, M. Wohlmayr, and F. Pernkopf, "Source–filter-based single-channel speech separation using pitch information," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 2, pp. 242–255, 2011.

[10] M. Cooke, J. R. Hershey, and S. J. Rennie., "Monaural speech separation and recognition challenge," *Computer Speech & Language*, vol. 24, no. 1, pp. 1–15, 2010.

[11] Y.-M. Qian, C. Weng, X.-K. Chang, S. Wang, and D. Yu, "Past review, current progress, and challenges ahead on the cocktail party problem," *Frontiers of Information Technology & Electronic Engineering*, vol. 19, no. 1, pp. 40–63, 2018.

[12] J. Du, Y. Tu, Y. Xu, L. Dai, and C.-H. Lee, "Speech separation of a target speaker based on deep neural networks," in *Int. Conf. on Signal Processing*. IEEE, 2014, pp. 473–477.

[13] Y. Tu, J. Du, Y. Xu, L. Dai, and C.-H. Lee, "Speech separation based on improved deep neural networks with dual outputs of speech features for both target and interfering speakers," in *Int. Symposium on Chinese Spoken Language Processing*. IEEE, 2014, pp. 250–254.

[14] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2136–2147, 2015.

[15] 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 Trans. on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

[16] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*. IEEE, 2016, pp. 31–35.

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

[18] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Alternative objective functions for deep clustering," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, April 2018, pp. 686–690.

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

[20] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, March 2017, pp. 246–250.

[21] Y. Luo, Z. Chen, and N. Mesgarani, "Speaker-independent speech separation with deep attractor network," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 26, no. 4, pp. 787–796, 2018.

[22] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*. IEEE, 2017, pp. 241–245.

[23] C. Xu, W. Rao, X. Xiao, E. S. Chng, and H. Li, "Single channel speech separation with constrained utterance level permutation invariant training using grid LSTM," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, April 2018, pp. 6–10.

[24] Y. Luo and N. Mesgarani, "TasNet: Time-domain audio separation network for real-time, single-channel speech separation," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, April 2018, pp. 696–700.

[25] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[26] Y. Shao and D. Wang, "Model-based sequential organization in cochannel speech," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 289–298, 2006.

[27] C. Weng, D. Yu, M. L. Seltzer, and J. Droppo, "Deep neural networks for single-channel multi-talker speech recognition," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 23, no. 10, pp. 1670–1679, 2015.

[28] Y. Wang, J. Du, L.-R. Dai, and C.-H. Lee, "A gender mixture detection approach to unsupervised single-channel speech separation based on deep neural networks," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 25, no. 7, pp. 1535–1546, 2017.

[29] K. Tan, J. Chen, and D. Wang, "Gated residual networks with dilated convolutions for supervised speech separation," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2018.

[30] L. Drude, T. von Neumann, and R. Haeb-Umbach, "Deep attractor networks for speaker re-identification and blind source separation," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, April 2018, pp. 11–15.

[31] C. Li, L. Zhu, S. Xu, P. Gao, and B. Xu, "CBLDNN-based speaker-independent speech separation via generative adversarial training," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, April 2018, pp. 711–715.

[32] Y. Liu and D. Wang, "A CASA approach to deep learning based speaker-independent co-channel speech separation," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, April 2018, pp. 5399–5403.

[33] D. Gunawan and D. Sen, "Iterative phase estimation for the synthesis of separated sources from single-channel mixtures," *IEEE Signal Processing Letters*, vol. 17, no. 5, pp. 421–424, May 2010.

[34] Y. Luo and N. Mesgarani, "Real-time single-channel dereverberation and separation with time-domain audio separation network," *Interspeech*, pp. 342–346, 2018.

[35] ——, "TasNet: Surpassing ideal time-frequency masking for speech separation," *arXiv preprint arXiv:1809.07454*, 2018.

[36] Z.-X. Li, Y. Song, L.-R. Dai, and I. McLoughlin, "Source-aware context network for single-channel multi-speaker speech separation," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, April 2018, pp. 681–685.

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

[38] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for joint enhancement of magnitude and phase," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, March 2016, pp. 5220–5224.

[39] A. van den Oord, N. Kalchbrenner, L. Espeholt, k. kavukcuoglu, O. Vinyals, and A. Graves, "Conditional image generation with PixelCNN decoders," in *Advances in Neural Information Processing Systems 29*, 2016, pp. 4790–4798.

[40] K. Vesel, S. Watanabe, K. molkov, M. Karafit, L. Burget, and J. H. ernock, "Sequence summarizing neural network for speaker adaptation," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, March 2016, pp. 5315–5319.

[41] C. Recommendation, "Pulse code modulation (PCM) of voice frequencies," *ITU*, 1988.
[42] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *IEEE Int. Conf. on Computer Vision*, Dec 2015, pp. 1026–1034.
[43] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.
[44] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.
[45] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.
[46] J. Garofolo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) complete LDC93S6A," *Philadelphia, USA: Linguistic Data Consortium*, 1993.
[47] T. Chen, M. Li, Y. Li, M. Lin, N. Wang, M. Wang, T. Xiao, B. Xu, C. Zhang, and Z. Zhang, "MXNet: A flexible and efficient machine learning library for heterogeneous distributed systems," *Neural Information Processing Systems, Workshop on Machine Learning Systems*, 2015.
[48] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[49] X. Zhu, G. T. Beauregard, and L. L. Wyse, "Real-time signal estimation from modified short-time Fourier transform magnitude spectra," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1645–1653, 2007.
[50] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.
[51] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," *ITU-T Recommendation*, vol. 862, 2001.
[52] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.
[53] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, Jan 2015.
[54] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.