Multi-talker Speech Separation and Tracing with Permutation Invariant Training of Deep Recurrent Neural Networks

Morten Kolbæk*, *Student Member, IEEE*, Dong Yu†, *Senior Member, IEEE*,
Zheng-Hua Tan*, *Senior Member, IEEE*, and Jesper Jensen*, *Senior Member, IEEE*
*Department of Electronic Systems, Aalborg University, Denmark. †Tencent AI Lab, USA.

***Abstract*—Despite the significant progress made in the recent years in dictating single-talker speech, the progress made in speaker independent multi-talker mixed speech separation and tracing, often referred to as the cocktail-party problem, has been less impressive. In this paper we propose a novel technique for attacking this problem. The core of our technique is permutation invariant training (PIT), which aims at minimizing the source stream reconstruction error no matter how labels are ordered. This is achieved by aligning labels to the output streams automatically during the training time. This strategy effectively solves the label permutation problem observed in deep learning based techniques for speech separation. More interestingly, our approach can integrate speaker tracing in the PIT framework so that separation and tracing can be carried out in one step and trained end-to-end. This is achieved using recurrent neural networks (RNNs) by forcing separated frames belonging to the same speaker to be aligned to the same output layer during training. Furthermore, the computational cost introduced by PIT is very small compared to the RNN computation during training and is zero during separation. We evaluated PIT on the WSJ0 and Danish two- and three-talker mixed-speech separation tasks and found that it compares favorably to non-negative matrix factorization (NMF), computational auditory scene analysis (CASA), deep clustering (DPCL) and deep attractor network (DANet), and generalizes well over unseen speakers and languages.**

***Index Terms*—Permutation Invariant Training, Speech Separation, Cocktail Party Problem, Deep Learning, DNN, CNN, RNN, LSTM.**

I. Introduction

DESPITE the significant progress made in the recent years in dictating single-speaker speech [2]–[5], the progress made in multi-talker mixed speech separation and recognition, often referred to as the cocktail-party problem [6], [7], has been less impressive. Although human listeners can easily perceive separate sources in an acoustic mixture, the same task seems to be extremely difficult for automatic computing systems, especially when only a single microphone recording of the mixed-speech is available [8], [9].

Nevertheless, solving the cocktail-party problem is critical to enable scenarios such as automatic meeting transcription, automatic captioning for audio/video recordings (e.g., YouTube), multi-party human-machine interactions (e.g., in the world of Internet of things (IoT)), and advanced hearing aids, where speech overlapping is commonly observed.

Over the decades, many attempts have been made to attack this problem. Before the deep learning era, the most popular technique was computational auditory scene analysis (CASA) [10], [11]. In this approach, certain segmentation rules based on perceptual grouping cues [12] are (often semi-manually) designed to operate on low-level features to estimate a time-frequency mask that isolates the signal components belonging to different speakers. This mask is then used to reconstruct the signal. Non-negative matrix factorization (NMF) [13]–[15] is another popular technique, which aims to learn a set of non-negative bases that can be used to estimate mixing factors during evaluation. Both CASA and NMF have led to very limited success in separating sources in multi-talker mixed speech [8]. The most successful technique before the deep learning era is the model based approach [16]–[18], such as factorial GMM-HMM [19], that models the interaction between the target and competing speech signals and their temporal dynamics. Unfortunately this model assumes and only works under the closed-set speaker condition.

Motivated by the success of deep learning techniques in single-talker ASR [2]–[5], researchers have developed many deep learning techniques for speech separation in recent years. Typically, networks are trained based on parallel sets of mixtures and their constituent target sources [20]–[23]. The networks are optimized to predict the source belonging to the target class, usually for each time-frequency bin. Unfortunately, these works often focus on, and only work for, separating speech from (often challenging) background noise (or music) because speech has very different characteristics than noise/music. Note that there are indeed works that are aiming at separating multi-talker mixed speech (e.g., [23]). However, these works rely on speaker-dependent models by assuming that the (often few) target speakers are known during training.

The difficulty in speaker-independent multi-talker speech separation comes from the label ambiguity or permutation problem (which will be described in Section IV). Three deep learning based works [9], [24]–[26] have tried to address and solve this harder problem. In Weng et al. [9], which achieved the best result on the dataset used in 2006 monaural speech separation and recognition challenge [8], the instantaneous energy was used to solve the label ambiguity problem and

Manuscript received Jan 25, 2017. This paper significantly extends the work published in [1]. Corresponding author: D. Yu (email: dongyu@ieee.org). Part of work was done while at Microsoft Research.

a two-speaker joint-decoder with a speaker switching penalty was used to separate and trace speakers. This approach tightly couples with the decoder and is difficult to scale up to more than two speakers due to the way labels are determined. Hershey et al. [24], [25] made significant progress with their deep clustering (DPCL) technique. In their work, they trained an embedding for each time-frequency bin to optimize a segmentation (clustering) criterion. During evaluation, each time-frequency bin was first mapped into the embedding space upon which a clustering algorithm was used to generate a partition of the time-frequency bins. However, in their approach it is assumed that each time-frequency bin belongs to only one speaker (i.e., a partition) due to the clustering step. Although this is often a good approximation, it is known to be sub-optimal. To further improve the performance, they need to stack yet another network to estimate real masks for each source stream given the results from the deep clustering. In addition, DPCL is inefficient in performing end-to-end mapping, because the objective function is the affinity between the sources in the embedded space, instead of the separated signals themselves. Chen et al. [26] proposed a technique called deep attractor network (DANet). Following DPCL, their approach also learns a high-dimensional embedding of the acoustic signals. Different from DPCL, however, it creates attractor points (cluster centers) in the embedding space which pull together the time-frequency bins corresponding to each source. The training is conducted in a way similar to the expectation maximization (EM) principle. The main limitation of DANet is the requirement to estimate attractor points during evaluation time.

In this paper, we propose a new technique for attacking the speaker independent multi-talker speech separation and tracing problem. The main ingredient of our technique is a novel training criterion named permutation invariant training (PIT). Most prior arts treat speech separation as either a multi-class regression problem with given labels for each output layer, or a segmentation (or clustering) problem as in DPCL. PIT, however, considers it a *separation* problem by minimizing the separation error. More specifically, PIT first determines the best output-label assignment automatically and then minimizes the error given the assignment. This strategy, which is directly implemented inside the network structure, elegantly solves the long-lasting label permutation problem that has prevented progress on deep learning based techniques for speech separation. Moreover, unlike other techniques such as DPCL and DANet that require a separate clustering step to trace speech streams during evaluation, our approach can integrate speaker tracing in the PIT framework directly so that separation and tracing can be carried out in one step and trained end-to-end. This is implemented using recurrent neural networks (RNNs) by forcing separated frames belonging to the same speaker to be aligned to the same output layer during training. Note that in PIT the computational cost associated with label assignment is negligible compared to the RNN computation during training and is zero during evaluation.

We evaluated PIT on the WSJ0 and Danish two- and three-talker mixed-speech separation tasks. Experimental results indicate that PIT compares favorably to NMF, CASA, DPCL and DANet, and generalizes well over unseen speakers and languages. In other words, through the training process PIT learns acoustic cues for source separation, which are both speaker and language independent, similarly to humans. Since PIT is simple to implement and can be easily integrated and combined with other advanced techniques, we believe improvements built upon PIT have great potential to solve the cocktail-party problem.

The rest of the paper is organized as follows. In Section II we describe the monaural speech separation problem and the basic speech separation and reconstruction technique. In Section III we extend effective masks and optimization criteria used in separating single-talker speech from noises to multi-talker speech separation tasks. In Section IV we discuss the label ambiguity problem and introduce the permutation invariant training framework. In Section V we show how we can modify PIT so that speech tracing can be integrated into the framework naturally. We report series of experimental results in Section VI and conclude the paper in Section VII.

II. Monaural speech separation

The goal of monaural speech separation is to estimate the individual source signals $x_s[n], s = 1, \cdots, S$ in a linearly mixed single-microphone signal $y[n] = \sum_{s=1}^{S} x_s[n]$, based on the observed signal $y[n]$ only. In real situations, the source signals may be reverberated, i.e., the underlying clean signals are filtered before observed in the mixture. In this condition, we aim at recovering the reverberated source signals $x_s[n]$, i.e., we are not targeting the dereverberated signals.

The separation is usually carried out in the time-frequency domain, in which the task can be cast as recovering the short-time Fourier transformation (STFT) of the source signals $X_s(t, f)$ for each time frame t and frequency bin f, given the mixed speech

$$\begin{aligned} Y(t,f) &= \sum_{s=1}^{S} X_s(t,f) \\ &= \sum_{n=0}^{N-1} y[n+tL] w_a[n] \exp(-j2\pi nf/N), \end{aligned} \tag{1}$$

where $w_a[n]$ is the analysis window of length N, the signal is shifted by an amount of L samples for each time frame $t = 0, \cdots, T-1$, and each frequency bin $f = 0, \cdots, N-1$ is corresponding to a frequency of $(f/N)f_s$ [Hz] when the sampling rate is f_s [Hz]. From the estimated STFT $\hat{X}_s(t, f)$ of each source signal, the inverse STFT

$$\hat{x}_t[n] = \frac{1}{N} \sum_{f=0}^{N-1} \hat{X}(t,f) \exp(j2\pi nf/N) \tag{2}$$

can be used to compute the estimated time-domain signal in each frame, and the overlap-add operation

$$\hat{x}[n] = \sum_{t=0}^{T-1} w_s[n - tL] \hat{x}_t[n - tL] \tag{3}$$

can be exploited to reconstruct the estimate of the original signal, where $w_s[n]$ is the synthesis window.

In a typical setup, however, only the STFT magnitude spectra $|X_s(t,f)|$ is estimated from the mixture during the separation process. The phase of the mixed speech is used directly when recovering the time domain waveforms of the sources. This is because phase prediction is extremely hard in the speech separation setup [27].

Obviously, given only the magnitude of the mixed spectrum $|Y(t,f)|$, the problem of recovering $|X_s(t,f)|$ is underdetermined, as there are an infinite number of possible $|X_s(t,f)|$ combinations that lead to the same $|Y(t,f)|$. To overcome this problem, the system has to learn from some training set $\mathbb{S}$ that contains corresponding observations of $|Y(t,f)|$ and $|X_s(t,f)|$, $s = 1, ..., S$. More specifically, we train a deep learning model $g(\cdot)$ such that $g\left(v\left(\mathbf{Y}\right);\mathbf{\Phi}\right) = |\hat{\mathbf{X}}_s|, s = 1, \cdots, S$, where $\mathbf{\Phi}$ are the model parameters, and $v(\mathbf{Y})$ is some feature representation of the spectra $\mathbf{Y}$: In a particularly simple situation, $v(\mathbf{Y}) = |\mathbf{Y}|$, i.e., the feature representation is simply identical to the magnitude spectrum of the observed signal $\mathbf{Y}$.

It is possible to directly estimate the magnitude spectra of each source using a deep learning model. However, it is well-known (e.g., [20], [28]) that better results can be achieved if, instead of estimating $|X_s(t,f)|$ directly, we first estimate a set of masks $M_s(t,f)$ using a deep learning model $h\left(v(\mathbf{Y});\mathbf{\Phi}\right) = \hat{\mathbf{M}}_s$ and reconstruct the magnitude spectra $|\mathbf{X}_s|$ as $|\hat{\mathbf{X}}_s| = \hat{\mathbf{M}}_s \circ |\mathbf{Y}|$, where $\circ$ is the element-wise product of two operands. This is because masks are well constrained and are invariant to input variabilities caused by, e.g., energy differences. This strategy is adopted in this study.

III. Masks and training criteria

Since masks are to be estimated as an intermediate step towards estimating magnitude spectra of source signals, it is important to adopt the correct masks and effective training objectives. In this study, we extend the three most effective and popular masks defined for separating single-talker speech from noises to the multi-talker speech separation task at hand.

In our study, the ideal ratio mask (IRM) for each source is defined as

$$M_s^{irm}(t,f) = \frac{|X_s(t,f)|}{\sum_{s=1}^{S}|X_s(t,f)|}. \tag{4}$$

When the phase of $\mathbf{Y}$ is used for reconstruction, the IRM maximize the signal to distortion ratio (SDR) [29], when all sources have the same phase, which is an invalid assumption in most cases. IRMs have the constraint that $0 \leq M_s^{irm}(t,f) \leq 1$ and $\sum_{s=1}^{S} M_s^{irm}(t,f) = 1$ for all time-frequency bins (t,f). This constraint can be easily satisfied with the softmax activation function. If one or more sources are noise and those sources are not estimated (modeled) by the network, then the sum-to-one constraint is no longer valid, under which condition a sigmoid is a better activation function for estimating those masks. Since $\sum_{s=1}^{S}|X_s(t,f)|$ is unknown in the mixed speech, the IRM cannot be practically used to reconstruct the source streams, but instead can be used to compute an estimated upper bound of performance.

Another applicable mask is the ideal amplitude mask (IAM), which for each source is defined as

$$M_s^{iam}(t,f) = \frac{|X_s(t,f)|}{|Y_s(t,f)|}. \tag{5}$$

Through IAMs we can construct the exact $\mathbf{X}_s$ given the magnitude spectra of the mixed speech. If the phase of each source equals the phase of the mixed speech, we can achieve the highest SDR. Unfortunately, this assumption is also not satisfied in most cases. IAMs have the constraint that $0 \leq M_s^{iam}(t,f) \leq \infty$, although we found empirically that the majority of the T-F units are in the range of $0 \leq M_s^{iam}(t,f) \leq 1$. For this reason, softmax, sigmoid and ReLU are possible activation functions for estimating IAMs.

Both IRM and IAM do not consider phase differences between source signals and the mixture. This leads to sub-optimal results. The phase sensitive mask (PSM) [28], [30]

$$M_s^{psm}(t,f) = \frac{|X_s(t,f)|cos(\theta_y(t,f) - \theta_s(t,f))}{|Y_s(t,f)|}, \tag{6}$$

however, takes phase difference into consideration, where θ_y and θ_s are the phases of mixed speech y and source x_s, respectively. Due to the phase-correcting term, the PSM sums to one, i.e. $\sum_{s=1}^{S} M_s^{psm}(t,f) = 1$. Note that since $|cos(\cdot)| \leq 1$ the PSM is smaller than the IAM especially when the phase difference between the mixed speech and the source is large.

Even though the PSM in theory is unbounded, we found empirically that the majority of the PSM is in the range of $0 \leq M_s^{psm}(t,f) \leq 1$. Actually, in our study we have found that approximately 20% of PSMs are negative. However, those negative PSMs usually are very close to zero. To account for this observation, we propose the non-negative PSM (NPSM) which is defined as

$$M_s^{npsm}(t,f) = max(0, M_s^{psm}(t,f)). \tag{7}$$

Softmax, Sigmoid, and ReLU are possible activation functions to estimate PSMs and NPSMs.

Since we first estimate masks, through which the magnitude of each source spectrum can then be estimated, the model parameters can be optimized to minimize the mean square error (MSE) between the estimated mask $\hat{M}_s$ and one of the target masks defined above as

$$J_m = \frac{1}{B}\sum_{s=1}^{S}\|\hat{\mathbf{M}}_s - \mathbf{M}_s\|_F^2, \tag{8}$$

where $B = T \times N \times S$ is the total number of time-frequency bins over all sources and $\|\cdot\|_F$ is the Frobenius norm. This approach comes with two problems. First, in silence segments, $|X_s(t,f)| = 0$ and $|Y(t,f)| = 0$, so that the target masks $M_s(t,f)$ are not well defined. Second, what we really care about is the error between the reconstructed source signal and the true source signal, while a smaller error on masks may not lead to a smaller reconstruction error.

To overcome these limitations, recent works [20] directly minimize the mean squared error (MSE)

$$J_x = \frac{1}{B}\sum_{s=1}^{S} \||\hat{\mathbf{X}}_s| - |\mathbf{X}_s|\|_F^2 = \frac{1}{B}\sum_{s=1}^{S} \|\hat{\mathbf{M}}_s \circ |\mathbf{Y}| - |\mathbf{X}_s|\|_F^2 \quad (9)$$

between the estimated magnitude and the true magnitude. Note that in silence segments $|\mathbf{X}_s| = 0$ and $|\mathbf{Y}| = 0$, the accuracy of mask estimation does not affect the training criterion for those segments.

When the phase sensitive mask is used, we optimize

$$J_p = \frac{1}{B}\sum_{s=1}^{S} \|\hat{\mathbf{M}}_s \circ |\mathbf{Y}| - |\tilde{\mathbf{X}}_s|\|_F^2 = \frac{1}{B}\sum_{s=1}^{S} \|\hat{\mathbf{M}}_s \circ |\mathbf{Y}| - |\mathbf{X}_s| \circ cos(\theta_y - \theta_s)\|_F^2 \quad (10)$$

following [28], where $|\tilde{\mathbf{X}}_s| = |\mathbf{X}_s| \circ cos(\theta_y - \theta_s)$ is the phase discounted magnitude target. In other words, using phase sensitive masks is as easy as replacing the original training targets with the phase discounted targets.

IV. Permutation invariant training

As far as we know, except DPCL [24], [25] and DANets [26], all other recent speech separation works treat the separation problem as a multi-class regression problem as depicted in Figure 1.

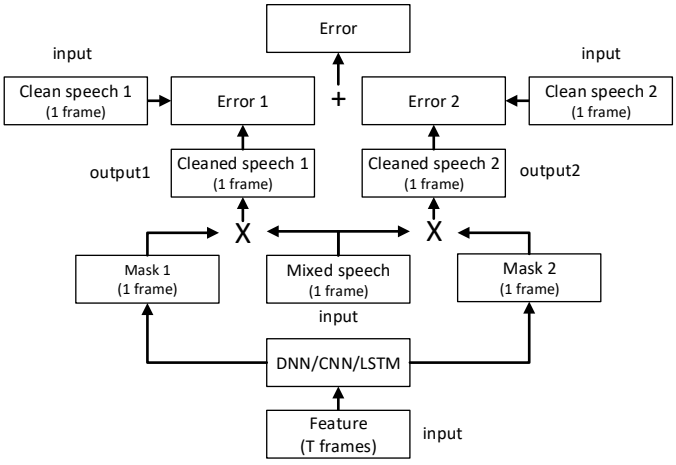

Fig. 1. The conventional two-talker speech separation model.

For this particular architecture, T frames of feature vectors of the mixed signal $|\mathbf{Y}|$ are used as the input to some deep learning models, such as deep neural networks (DNNs), convolutional neural networks (CNNs), and long short-term memory (LSTM) recurrent neural networks (RNNs), to generate one (often the center) frame of masks for each talker. These masks are then used to construct one frame of single-source speech $|\hat{\mathbf{X}}_1(t)|$ and $|\hat{\mathbf{X}}_2(t)|$, for sources 1 and 2, respectively.

During training we need to provide the correct reference (or target) magnitude $|\mathbf{X}_1|$ and $|\mathbf{X}_2|$ to the corresponding

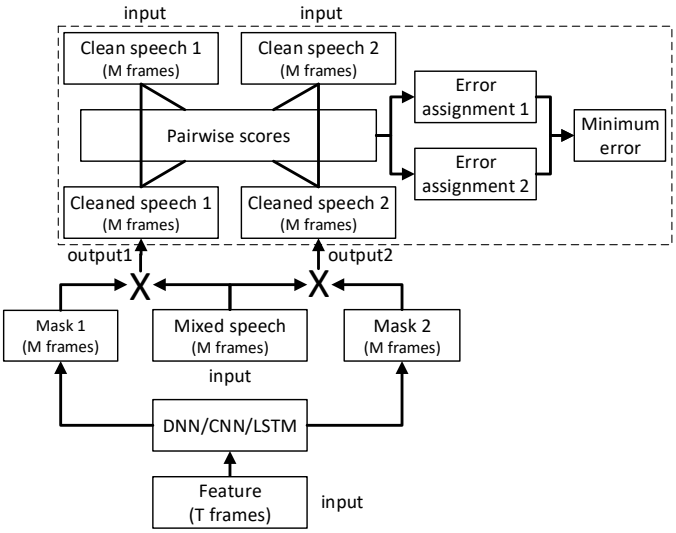

Fig. 2. The two-talker speech separation model with permutation invariant training.

output layers for supervision. Since the model has multiple output layers, one for each mixing source, and they depend on the same input mixture, reference assigning can be difficult especially if the training set contains many utterances spoken by many speakers of both genders. This problem is referred to as the label ambiguity (or permutation) problem in [9], [24]. Due to this problem, prior arts perform poorly on speaker-independent multi-talker speech separation. It was believed that speaker-independent multi-talker speech separation is not feasible [1]

Our solution to the label permutation problem is illustrated in Figure 2 and the core in the solution is permutation invariant training (PIT).

In the model depicted in Figure 2 the reference source streams are given as a set instead of an ordered list. In other words, the same training result is obtained, no matter in which order these sources are listed. This behavior is achieved with PIT highlighted inside the dashed rectangular in Figure 2. In order to associate references to the output layers, we first compute the (total of S^2) pairwise MSE between each reference $|\mathbf{X}_s|$ and each estimated source $|\hat{\mathbf{X}}_s|$. We then determine the (total of $S!$) possible assignments between the references and the estimated sources, and compute the total MSE for each assignment. The assignment with the least MSE is chosen and the model is optimized to reduce this least MSE. In other words, we simultaneously conduct label assignment and error evaluation. Similar to that in prior arts, we can use as input T successive frames (i.e., an input *meta-frame*) of features to exploit the contextual information. Different from the prior arts, the output of the PIT is also a window of frames. With PIT, we directly minimize the separation error at the meta-frame level.

During evaluation, the only information available is the mixed speech. Speech separation can be directly carried out for each input meta-frame, for which an output meta-frame with

[1]D. Wang. "Tutorial: Supervised speech separation". ICASSP 2016.

M frames of speech is estimated for each stream. The input meta-frame (and thus the output meta-frame) is then shifted by one or more frames. Due to the PIT training criterion, output-to-speaker assignment will stay the same for frames inside the same output meta-frame but may change across output meta-frames. In the simplest setup, we can just assume they do not change when reconstructing sources. However, this usually leads to unsatisfactory results as we will report in Section VI. To achieve better performance speaker-tracing algorithms need to be developed and applied on top of the output of the network.

V. Integrated separation and tracing

There are many possible ways to trace the source streams, i.e. identifying what network output a given speaker is represented at, or assigned to. For example, since there are several overlapping frames in adjacent output meta-frames, we can determine whether assignment has been switched or not by comparing MSEs of different permutations measured on the overlapping frames of adjacent output meta-frames. However, this approach has two major problems. First, it requires a separate tracing step which may complicate the model. Second, since the assignment for later frames depends on that for earlier frames, one incorrect assignment at earlier frame would completely switch the assignment for all frames after it, even if the rest of the assignment decisions are all correct.

In this work we propose a simpler yet more effective approach to solve the tracing problem. More specifically, we integrate speech separation and tracing into the same PIT framework. Our solution is based on the observation that when speech streams are perfectly traced, all frames from the same speaker should be aligned to the same output layer. For example, in a two-talker mixed speech case, if frame 1 of the speaker 1's speech is aligned to output 1 (or 2) then the rest of the frames of this speaker should also be aligned to output 1 (or 2). This suggests that we can learn to trace the speech streams, if we force speech frames from the same speaker be aligned to the same output layer during training. Fortunately, this can be easily implemented in the PIT framework with the following changes.

First, since we need to enforce assignment of speech streams for the whole utterance, we should provide the whole utterance, instead of a T (M)-frame segment, as the input (and output) to the deep learning model during training. In the segment-based PIT (which we will denote as PIT-S from now on) depicted in Figure 2, the label-output assignment decision is made based on the M-frame output meta-frame. In the tracing integrated model, however, the decision is made on the whole utterance. In other words, the pair-wise scores in Figure 2 are computed on the whole utterance assuming all frames from the same output layer belongs to the same speaker. To distinguish the new model from PIT-S we denote it as PIT-T (PIT with tracing).

Second, since utterances have variable length, and effective tracing requires exploitation of long-range dependency, models such as DNNs and CNNs are no longer good fits. Instead, we use recurrent neural networks (RNNs) such as deep long short-term memory (LSTMs) and bi-directional (BLSTMs) to learn the masks. Different from PIT-S, in which the input layer and each output layer has $N \times T$ and $N \times M$ units, respectively, in PIT-T, both input and output layers have N units (adding contextual frames in the input does not help for LSTMs). With deep LSTMs, the utterance is evaluated frame-by-frame exploiting the whole past or future history information at each layer. When BLSTMs are used, the information from the past and future (i.e., across the whole utterance) is stacked at each layer and used as the input to the next layer for better decisions at higher layers. With PIT-T, the model learns to separate and trace speech streams at the same time. During separation, we don't need to compute pairwise MSEs and errors of each possible permutation and no separate tracing step is needed. We simply treat all frames from the same output layer to be from the same speaker. This makes it a clean and attractive solution.

The separation and tracing accuracy can be further improved with a 2-stage (or stacking) system, in which the first stage's separation results are used to inform the second-stage model to make better decisions. More specifically the input to the second stage includes the mixed speech and the separated speech streams from the first stage. In addition, the final mask is the average of that from the first stage using only mixed speech as the input and that from the second stage that uses augmented features as the input. More specifically, in the 2-stage system, we have

$$\begin{aligned} \hat{\mathbf{M}}_s^{(1)} &= LSTM^{(1)}(|\mathbf{Y}|) \\ \hat{\mathbf{M}}_s^{(2)} &= LSTM^{(2)}(|\mathbf{Y}|, \hat{\mathbf{M}}_s^{(1)} \circ |\mathbf{Y}|) \\ \hat{\mathbf{M}}_s &= \frac{\hat{\mathbf{M}}_s^{(1)} + \hat{\mathbf{M}}_s^{(2)}}{2}. \end{aligned} \tag{11}$$

VI. Experimental results

We evaluated our technique on various setups. All our models were implemented using the Microsoft Cognitive Toolkit (CNTK) [31][2]. The models were evaluated on their potential to improve the signal-to-distortion ratio (SDR) [29] and the perceptual evaluation of speech quality (PESQ) [32] score, both of which are metrics widely used to evaluate speech enhancement performance for multi-talker speech separation tasks.

A. Datasets

We evaluated PIT on the WSJ0-2mix, WSj0-3mix[3] and Danish-2mix datasets using 129-dimensional STFT magnitude spectra computed with a sampling frequency of 8 kHz, a frame size of 32 ms and a 16 ms frame shift.

The WSJ0-2mix dataset was introduced in [24] and was derived from the WSJ0 corpus [33]. The 30h training set and the 10h validation set contain two-speaker mixtures generated by randomly selecting from 49 male and 51 female speakers and utterances from the WSJ0 training set si_tr_s, and mixing

[2]Available at: https://www.cntk.ai/

[3]Available at: http://www.merl.com/demos/deep-clustering

them at various signal-to-noise ratios (SNRs) uniformly chosen between 0 dB and 5 dB. The 5h test set was similarly generated using utterances from 16 speakers from the WSJ0 validation set si_dt_05 and evaluation set si_et_05. The WSJ0-3mix dataset was generated using a similar approach but contains mixtures of speech from three talkers.

The Danish-2mix dataset[4] consists of approximately 560 speakers each speaking 312 utterances with average utterance duration of approximately 5 sec. The dataset was constructed by randomly selecting a set of 45 male and 45 female speakers from the corpus, and then allocating 232 and 40 utterances from each speaker to generate mixed speech in the training, and validation set, respectively. A number of 40 utterances from each of another 45 male and 45 female speakers were randomly selected to construct the open-condition (OC) (unseen speaker) test set. Speech mixtures were constructed similarly to the WSJ0-2mix with SNRs selected uniformly between 0 dB and 5 dB. Similarly to the WSJ0-2mix dataset we constructed 20k and 5k mixtures in total in the training and validation set, respectively, and 3k mixtures for the OC test set.

In our study, the validation set is used to evaluate closed-condition (CC) (seen speaker) performance, control the learning rate and to choose the model for open-condition evaluation.

B. Segment-based Permutation Invariant Training

We first evaluated segment-based PIT on the two-talker separation dataset WSJ0-2mix. In the segment-based PIT, the input window and output window sizes are fixed. For this reason, we can use feed-forward DNNs and CNNs. The DNN model has three hidden layers each with 1024 ReLU units. In (inChannel, outChannel)-(strideW, strideH) format, the CNN model has one $(1, 64) - (2, 2)$, four $(64, 64) - (1, 1)$, one $(64, 128) - (2, 2)$, two $(128, 128) - (1, 1)$, one $(128, 256) - (2, 2)$, and two $(256, 256) - (1, 1)$ convolution layers with 3×3 kernels, a pooling layer and a 1024-unit ReLU layer. The input to the models is the stack (over multiple frames) of the 129-dimensional STFT spectral magnitude of the speech mixture. There are S output streams for S-talker mixed speech. Each output stream has a dimension of $129 \times M$, where M is the number of frames in the output meta-frame.

In Figure 3 we plot the DNN training progress as measured by the MSE on the training and validation set with conventional training (CONV-DNN) and PIT (PIT-S-DNN) on the WSJ0-2mix datasets described in subsection VI-A. We also plotted training progress for another conventionally trained model but with a slightly modified version of the WSJ0-2mix dataset, where speaker labels have been randomzed (CONV-DNN-RAND).

The WSJ0-2mix dataset, used in [24], was designed such that speaker one was always assigned the most energy, and consequently speaker two the lowest, when scaling to a given SNR. Previous work [9] has shown that such speaker energy patterns are an effective discriminative feature, which is clearly seen in Figure 3, where the CONV-DNN model achives considerably lower training and validation MSE than

[4]Available at: http://www.nb.no/sbfil/dok/nst_taledat_dk.pdf

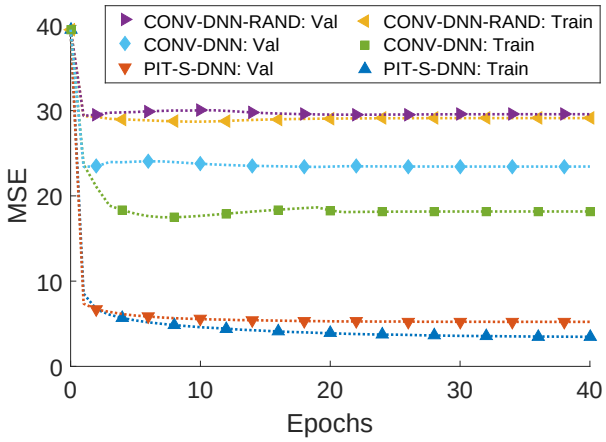

Fig. 3. MSE over epochs on the WSJ0-2mix training and validation sets with conventional training and PIT.

TABLE I
SDR IMPROVEMENTS (DB) FOR DIFFERENT SEPARATION METHODS ON THE WSJ0-2MIX DATASET USING PIT-S.

Method	Input\Output window	Opt. Assign. CC	Opt. Assign. OC	Def. Assign. CC	Def. Assign. OC
PIT-S-DNN	51\51	6.8	6.7	**5.2**	**5.2**
PIT-S-DNN	51\5	**10.3**	**10.2**	-0.8	-0.8
PIT-S-CNN	51\51	9.6	9.6	**7.6**	**7.5**
PIT-S-CNN	51\5	**10.9**	**11.0**	-1.0	-0.9
IRM	-	12.4	12.7	12.4	12.7

the CONV-DNN-RAND model, which hardly decreases in either training or validation MSE due to the label permutation problem discussed in [9], [24]. In contrast, training converges quickly to a very low MSE when PIT is used no matter how speakers are ordered.

In Table I we summarize the SDR improvement in dB from different segment-based PIT separation configurations for two-talker mixed speech in closed condition (CC) and open condition (OC). In these experiments each frame was reconstructed by averaging over all output meta-frames that contain the same frame. In the default assignment (def. assign.) setup it is assumed that there is no output-speaker switch across frames (which is not true). This is the achievable SDR improvement using PIT-S without any additional speaker tracing. In the optimal assignment (opt. assign.) setup, the output-speaker assignment for each output meta-frame is determined based on the true sources, i.e. oracle information. This reflects the separation performance within each segment (meta-frame) and is the improvement achievable when the speakers are correctly traced. The gap between these two values indicates the possible contribution from speaker tracing. As a reference, we also provided the IRM result which is an estimated upper bound achievable on this task using real masks.

From the table we can make several observations. First, without speaker tracing (def. assign.) PIT-S can already achieve 7.5 dB SDR improvement, even though the model is very simple. Second, as we reduce the output window size, we can improve the separation performance within each

TABLE II
SDR IMPROVEMENTS (DB) FOR DIFFERENT SEPARATION METHODS ON THE WSJ0-2MIX DATASET USING PIT-T.

Method	Mask Type	Activation Function	Opt. Assign. CC	Opt. Assign. OC	Def. Assign. CC	Def. Assign. OC
PIT-T-BLSTM	IAM	softmax	**10.4**	**10.3**	**9.0**	**8.7**
PIT-T-BLSTM	IAM	sigmoid	8.3	8.3	7.1	7.2
PIT-T-BLSTM	IAM	ReLU	9.9	9.9	8.7	8.6
PIT-T-BLSTM	IAM	Tanh	8.5	8.6	7.5	7.5
PIT-T-BLSTM	PSM	softmax	10.3	10.2	9.1	9.0
PIT-T-BLSTM	PSM	sigmoid	10.5	10.4	9.2	9.1
PIT-T-BLSTM	PSM	ReLU	**10.9**	**10.8**	**9.4**	**9.4**
PIT-T-BLSTM	PSM	Tanh	10.4	10.3	9.0	8.9
PIT-T-BLSTM	NPSM	softmax	8.7	8.6	7.5	7.3
PIT-T-BLSTM	NPSM	sigmoid	**10.6**	**10.6**	**9.4**	**9.3**
PIT-T-BLSTM	NPSM	ReLU	8.8	8.8	7.6	7.6
PIT-T-BLSTM	NPSM	Tanh	10.1	10.0	8.9	8.8
PIT-T-LSTM	PSM	ReLU	**9.8**	**9.8**	7.0	**7.0**
PIT-T-LSTM	PSM	sigmoid	**9.8**	9.6	**7.1**	6.9
PIT-T-LSTM	NPSM	ReLU	**9.8**	**9.8**	**7.1**	**7.0**
PIT-T-LSTM	NPSM	sigmoid	9.2	9.2	6.8	6.8
PIT-U-BLSTM	PSM	ReLU	**11.7**	**11.7**	**-1.7**	-1.9
PIT-U-BLSTM	PSM	sigmoid	**11.7**	**11.7**	**-1.7**	**-1.7**
PIT-U-BLSTM	NPSM	ReLU	**11.7**	**11.7**	**-1.7**	-1.8
PIT-U-BLSTM	NPSM	sigmoid	11.6	11.6	-1.6	**-1.7**
IRM	-	-	12.4	12.7	12.4	12.7

window and achieve better SDR improvement, if speakers are correctly traced (opt. assign.). However, when output window size is reduced, the output-speaker assignment changes more frequently as indicated by the poor def. assign. performance. Speaker tracing thus becomes more important given the larger gap between the opt. assign. and def. assign. Third, PIT-S generalizes well to unseen speakers, since the performances on the open and closed conditions are very close. Fourth, powerful models such as CNNs consistently outperform DNNs, but the gain diminishes when the output window size is small.

C. Permutation Invariant Training with Integrated Tracing

As indicated by Table I, speaker tracing is critical to further improve the separation quality. In this subsection we evaluate the permutation invariant training technique with integrated tracing discussed in Section V and the results are summarized in Table II. Because speaker tracing requires exploiting long-range context, recurrent neural networks (RNN) are the obvious choice. In this set of experiments, we used long short-term memory (LSTM) RNNs. All the uni-directional LSTMs (PIT-T-LSTM) evaluated have 3 LSTM layers each with 1792 units and all the bi-directional LSTMs (PIT-T-BLSTM) have 3 BLSTM layers each with 896 units, so that both models have similar number of parameters. All models contain random dropouts when fed from a lower layer to a higher layer and were trained with a dropout rate of 0.5. Note that, since we used Nvidia's cuDNN implementation of LSTMs to speed up training, we were unable to apply dropouts across time steps, which was adopted by the best DPCL model [25] and is known to be more effective, both theoretically and empirically, than the simple dropout strategy used in this work [34]. In all the experiments reported in Table II the maximum epoch is set to 200 although we noticed that further performance improvement is possible with additional training epochs. Note that the epoch size of 200 seems to be significantly larger than that in PIT-S as indicated in Figure 3. This is because in PIT-S each frame is used by T ($T = 51$) training samples (input meta-frames) while in PIT-T each frame is used just once in each epoch. The learning rates were set to 2×10^{-5} per sample initially and scaled down by 0.7 when the training objective function value increases on the training set. The training was terminated when the learning rate got below 10^{-10}. Each minibatch contains 8 randomly selected utterances without bucketing the utterances based on the length first as in some other works. As a related baseline, we also included PIT-U-BLSTM results in the table. These models were also trained using whole utterances (instead of segments) as unit. The only difference between them and PIT-T models is that in these models tracing is not enabled and output-target assignment was determined frame by frame during training.

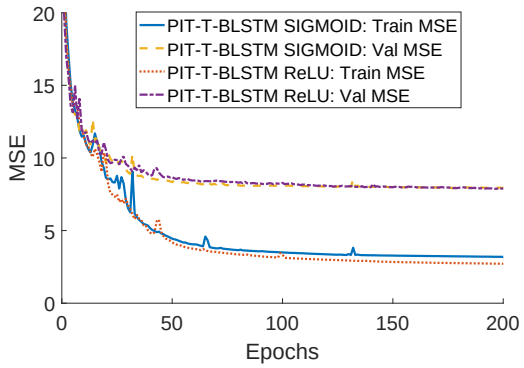

Fig. 4. MSE over epochs on the WSJ0-2mix PSM training and validation sets wit PIT-T.

In Figure 4 we plotted the BLSTM training progress, as measured by the MSE on the training and validation set, with PIT-T on the two-talker mixed speech using the phase sensitive masks. It is obvious that the training behaves desirable even though training models with tracing is assumed harder than training without, because assignments are fixed for the same outputs for a whole utterance.

We can notice several things from Table II. First, with integrated tracing, we can significantly improve the SDR with def. assign. over the segment-based PIT. In fact, we can achieve 9.4 dB SDR improvement on both CC and OC sets by directly treating estimated magnitude spectra from the same output layer as belonging to the same speaker (def. assign.), which compares favorably to 7.6 dB (CC) and 7.5 dB (OC) achieved with deep CNN and segment-based PIT. We want to emphasize that this is achieved by the integrated tracing, and not by the usage of BLSTM because the corresponding PIT-U-BLSTM def. assign. results are so much worse even though the opt. assign. result is the best among all models. Second, we can achieve better SDR improvement over ideal amplitude mask (IAM) using phase-sensitive magnitude (PSM) and non-negative phase-sensitive magnitude (NPSM) tracing criteria. This indicates that including phase information does improve performance even though it was used implicitly. Third, for

TABLE III
FURTHER IMPROVEMENT ON THE WSJ0-2MIX DATASET WITH ADDITIONAL TRAINING EPOCHS WITH REDUCED DROPOUT (-RD) OR STACKED MODELS (-ST)

Method	Mask Type	Activation Function	Opt. Assign. CC	Opt. Assign. OC	Def. Assign. CC	Def. Assign. OC
PIT-T-BLSTM-RD	PSM	ReLU	11.0	11.0	9.5	9.5
PIT-T-BLSTM-ST	PSM	ReLU	**11.7**	**11.7**	10.0	**10.0**
PIT-T-BLSTM-RD	NPSM	Sigmoid	10.7	10.7	9.5	9.4
PIT-T-BLSTM-ST	NPSM	Sigmoid	11.5	11.5	**10.1**	**10.0**

TABLE IV
SDR (DB) IMPROVEMENTS ON TEST SETS OF WSJ0-2MIX WITH DIFFERENT GENDER-COMBINATIONS

Method	Config	CC Same	CC Diff	OC Same	OC Diff
PIT-T-BLSTM-RD	PSM-ReLU	7.5	11.5	7.1	11.6
PIT-T-BLSTM-ST	PSM-ReLU	7.8	**12.1**	**7.5**	**12.2**
PIT-T-BLSTM-RD	NPSM-Sigmoid	7.5	11.5	7.0	11.5
PIT-T-BLSTM-ST	NPSM-Sigmoid	**8.0**	**12.1**	**7.5**	12.1
IRM	-	12.2	12.7	12.4	12.9

different training criteria, the optimal non-linear activation function varies. For IAM, PSM and NPSM, they are softmax, ReLU and Sigmoid, respectively, in this set of experiments. Fourth, we can observe that with the integrated tracing, the gap between opt. assign. and def. assign. is always less than 1.5 dB across different setups.

The performance can be further improved with additional epochs, better control of dropout rate and usage of the two-stage stacked model described in Section V. For the dropout rate control, we used a simple two-step rate with 0.5 first for 200 epochs and then 0.3 for another 200 epochs. The two-stage stacked model is built upon the corresponding best single-stage model. Table III shows that continued training with slightly lower dropout rate helps, and the two-stage stacked model always outperforms the single-stage model. Overall we can achieve a 10 dB SDR improvement on this task.

Table IV reports SDR (dB) improvements on test sets of WSJ0-2mix divided into different-gender and same-gender. From this table we can clearly see that our approach achieves much better SDR improvements on the different-gender mixed speech than the same-gender mixed speech, although the gender information is not explicitly used in our model and training procedure. In fact, for the different-gender condition, the SDR improvement is already very close to the IRM result. These results agree with breakdowns from other works [24], [25] and generally indicate that same-gender mixed speech separation is a harder task.

Table V summarizes SDR (dB) and PESQ improvements for different separation methods on the WSJ0-2mix dataset. As a reference, we also provide the IRM result, which can be considered as the upper bound on this task. From the table we can observe that the models trained with segment-based PIT can achieve similar or better performance than the original DPCL [24], respectively, with DNN and CNN, without any additional tracing component. However they underperform the more algorithmically complicated DPCL++ models [25], [26] and DANet models. Our best model performs better than all but the most complicated DPCL++ model[5] reported in [25] which used a multi-stage model, involving complicated training scheduling, and used cross-time dropout. Note that, PIT models are much simpler than even the original (simpler) DPCL, because PIT models do not require any clustering step during evaluation. In addition, PIT is much easier to integrate with other techniques such as complex-domain separation techniques than DPCL and DANet.

TABLE V
SDR (DB) AND PESQ IMPROVEMENTS FOR DIFFERENT SEPARATION METHODS ON THE WSJ0-2MIX DATASET WITHOUT ADDITIONAL TRACING (I.E., DEF. ASSIGN.).

Method	Config	PESQ Imp CC	PESQ Imp OC	SDR Imp CC	SDR Imp OC
Oracle NMF [24]	-	-	-	5.1	-
CASA [24]	-	-	-	2.9	3.1
DPCL [24]	-	-	-	5.9	5.8
DPCL+ [26]	-	-	-	-	9.1
DANet [26]	-	-	-	-	9.6
DPCL++ [25]	-	-	-	-	10.8
PIT-S-DNN	51\51	0.2	0.2	5.2	5.2
PIT-S-CNN	51\51	0.5	0.5	7.6	7.6
PIT-T-BLSTM	PSM-ReLU	0.7	0.6	9.4	9.4
PIT-T-BLSTM-ST	PSM-ReLU	**0.9**	**0.8**	**10.0**	**10.0**
IRM	-	2.1	2.1	12.4	12.7

TABLE VI
SDR (DB) AND PESQ IMPROVEMENTS ON WSJ0-2MIX AND DANISH-2MIX WITH PIT-T-BLSTM-PSM-RELU TRAINED ON WSJ0-2MIX AND A COMBINATION OF TWO LANGUAGES.

Trained on	WSJ0-2mix SDR	WSJ0-2mix PESQ	Danish-2mix SDR	Danish-2mix PESQ
WSJ0-2mix	9.3	0.7	8.1	0.4
+Danish-2mix	8.8	0.6	10.6	0.5
IRM	12.7	2.1	15.2	1.9

To further understand the properties of our approach, we evaluated the PIT-T-BLSTM-PSM-ReLU model trained on wsj0-2mix (English) on the Danish-2mix test set. The results of this is reported in Table VI. An interesting observation is that although the system has never seen Danish speech, it performs remarkably well on the Danish-2mix dataset, when compared to the IRM (oracle) values. These results indicate that the separation ability learned with PIT generalizes well not only across speakers but also across languages.

We also trained a model with the combination of English and Danish datasets and evaluated the models on both languages. The results of these experiments, as summarized in Table VI. Table VI, indicate that by including Danish data, we can achieve better performance on the Danish dataset, at the cost of slightly worse performance on the English dataset. Note that while doubling the training set we did not change the model size, which likely will improve performance on both languages.

[5]In [25] they did not use the SDR measure from [29] as they did in [24]. Instead they defined and used a slightly different measure called scale-invariant SNR.

TABLE VII
SDR IMPROVEMENTS (DB) FOR DIFFERENT SEPARATION METHODS ON THE WSJ0-3MIX DATASET USING PIT-T.

Method	Units/ layer	Activation function	Opt. Assign. CC	Opt. Assign. OC	Def. Assign. CC	Def. Assign. OC
Oracle NMF [24]	-	-	4.5	-	-	-
DPCL++ [25]	-	-	-	-	-	7.1
PIT-T-BLSTM	896	Sigmoid	10.0	9.9	7.4	7.2
PIT-T-BLSTM	1280	Sigmoid	10.1	10.0	7.5	7.4
PIT-T-BLSTM-RD	1280	Sigmoid	10.2	10.1	7.6	7.4
PIT-T-BLSTM-ST	1280	Sigmoid	**10.7**	**10.6**	**7.9**	**7.7**
IRM	-	-	12.6	12.8	12.6	12.8

D. Three-Talker Speech Separation

In Figure 5 we plotted the PIT-T training progress as measured by MSE on the three-talker mixed speech training and validation sets WSJ0-3mix. We observe that similar to the two-talker scenario in Figure 4, a low training MSE is achieved, although the validation MSE is slightly higher. A better balance between the training and validation MSEs may be achieved by controlling the learning schedule. We also observe that increasing the model size decreases both training and validation MSE, which is expected due to the more variability in the dataset.

In Table VII we summarized the SDR improvement in dB from different PIT-T separation configurations for three-talker mixed speech in closed condition (CC) and open condition (OC). We observe that the basic PIT-T-BLSTM model (896 units) compares favorably with DPCL++. With additional units, further training and stacked models (based on PIT-T-BLSTM), PIT-T outperform DPCL++ on this three-talker separation task. Our best model achieved 7.7 dB SDR improvement in open condition.

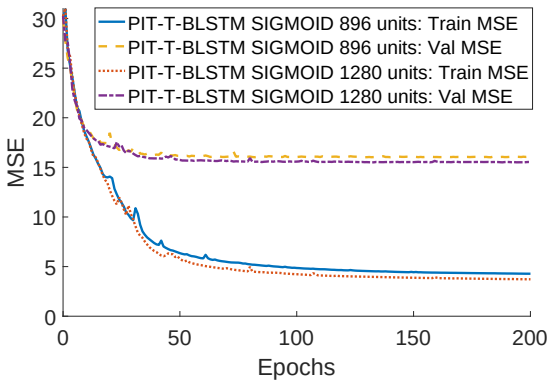

Fig. 5. MSE over epochs on the WSJ0-3mix NPSM training and validation sets wit PIT-T.

E. Combined Two and Three-Talker Speech Separation

In Table VIII we summarized the performance of the 3-speaker PIT-T-BLSTM and PIT-T-BLSTM-ST models from Table VII when trained and tested on both the wsj0-2mix and wsj0-3mix datasets, i.e. on both 2 and 3 speakers. To allow training the 3-speaker models with the 2-speaker wsj0-2mix dataset, we extended wsj0-2mix with a third "silent" channel. We see from Table VIII that PIT-T-BLSTM achieves good, but slightly worse, performance compared to the corresponding 2-speaker (Table V) and 3-speaker (Table VII) models. Surprisingly, the PIT-T-BLSTM-ST model outperforms both the 2-speaker (Table III) and 3-speaker PIT-T-BLSTM-ST (Table VII) models. These results indicate that a single model can handle a varying, and more importantly, an unknown number of speakers, without compromising performance. This is of great practical importance, since a priori knowledge about the number of speakers is not needed at test time, as required by competing methods such as DPCL++ [25] and DANet [26].

TABLE VIII
SDR IMPROVEMENTS (DB) FOR 3-SPEAKER MODELS TRAINED ON BOTH THE WSJ0-2MIX AND WSJ0-3MIX PSM DATASETS. BOTH MODELS HAVE 1280 UNITS PER LAYER AND RELU OUTPUTS.

Method	2 Spkr. Def. Assign. CC	2 Spkr. Def. Assign. OC	3 Spkr. Def. Assign. CC	3 Spkr. Def. Assign. OC
PIT-T-BLSTM	9.4	9.3	7.2	7.1
PIT-T-BLSTM-ST	**10.2**	**10.1**	**8.0**	**7.8**

VII. CONCLUSION AND DISCUSSION

In this paper, we have introduced the permutation invariant training (PIT) technique for speaker-independent multi-talker speech separation and tracing. We consider this an interesting step towards solving the important cocktail-party problem in a real-world setup, where the set of speakers are unknown during the training time.

Our experiments on two and three-talker mixed speech separation tasks indicate that PIT can indeed effectively deal with the label permutation problem. These experiments show that bi-directional long short-term memory (LSTM) recurrent neural networks perform better than uni-directional LSTMs and phase sensitive masks (PSM) are better training criteria than ideal amplitude masks (IAM). Our results also suggest that the acoustic cues learned by the model are largely speaker and language independent and the models generalize well to unseen speakers and languages. This indicates that it is possible to train a universal speech separation model using speech in various speaker, language and noise conditions.

The proposed model PIT-T is algorithmically simpler yet performs on par or better than DPCL [24], [25] and DANets [26], both of which involve separate embedding and clustering stages during evaluation. Since PIT, as a training technique, can be easily integrated and combined with other advanced techniques such as complex-domain separation and multi-channel techniques such as beam-forming, it has great potential for further improvement.

ACKNOWLEDGMENT

We would like to thank Dr. John Hershey at MERL and Zhuo Chen at Columbia University for sharing the WSJ0-2mix dataset and for valuable discussions. We also thank Dr. Hakan Erdogan at Microsoft Research for discussions on PSM.

REFERENCES

[1] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. ICASSP*, 2017.

[2] D. Yu, L. Deng, and G. E. Dahl, "Roles of pre-training and fine-tuning in context-dependent dbn-hmms for real-world speech recognition," in *NIPS 2010 Workshop on Deep Learning and Unsupervised Feature Learning*, 2010.

[3] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 20, no. 1, pp. 30–42, 2012.

[4] F. Seide, G. Li, and D. Yu, "Conversational speech transcription using context-dependent deep neural networks," in *Proc. INTERSPEECH*, 2011, pp. 437–440.

[5] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[6] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[7] A. S. Bregman, *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1994.

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

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

[10] M. Cooke, *Modelling auditory processing and organisation*. Cambridge University Press, 2005.

[11] D. P. Ellis, "Prediction-driven computational auditory scene analysis," Ph.D. dissertation, Massachusetts Institute of Technology, 1996.

[12] M. Wertheimer, *Laws of organization in perceptual forms*. Kegan Paul, Trench, Trubner & Company, 1938.

[13] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *INTERSPEECH*, 2006.

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

[15] J. Le Roux, F. Weninger, and J. Hershey, "Sparse nmf–half-baked or well done," *Mitsubishi Electric Research Labs (MERL), Cambridge, MA, USA, Tech. Rep., no. TR2015-023*, 2015.

[16] T. T. Kristjansson, J. R. Hershey, P. A. Olsen, S. J. Rennie, and R. A. Gopinath, "Super-human multi-talker speech recognition: the ibm 2006 speech separation challenge system," in *Proc. INTERSPEECH*, 2006.

[17] T. Virtanen, "Speech recognition using factorial hidden markov models for separation in the feature space," in *Proc. INTERSPEECH*, 2006.

[18] R. J. Weiss and D. P. W. Ellis, "Monaural speech separation using source-adapted models," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct 2007, pp. 114–117.

[19] Z. Ghahramani and M. I. Jordan, "Factorial hidden markov models," *Machine learning*, vol. 29, no. 2-3, pp. 245–273, 1997.

[20] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 12, pp. 1849–1858, 2014.

[21] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2014.

[22] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr," in *Latent Variable Analysis and Signal Separation*. Springer, 2015, pp. 91–99.

[23] 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. Audio, Speech, Language Process.*, vol. 23, no. 12, pp. 2136–2147, 2015.

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

[25] Y. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-Channel Multi-Speaker Separation Using Deep Clustering," in *Proc. INTERSPEECH*, 2016, pp. 545–549.

[26] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *Proc. ICASSP*, 2017.

[27] D. S. Williamson, Y. Wang, and D. Wang, "Complex Ratio Masking for Monaural Speech Separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 3, pp. 483–492, 2016.

[28] H. Erdogan, J. R. Hershey, S. Watanabe, and J. L. Roux, "Deep recurrent networks for separation and recognition of single channel speech in non-stationary background audio," *New Era for Robust Speech Recognition: Exploiting Deep Learning*, 2017.

[29] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 14, no. 4, pp. 1462–1469, 2006.

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

[31] A. et al., "An introduction to computational networks and the computational network toolkit," MSR-TR-2014-112, Tech. Rep., 2014.

[32] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, 2001, pp. 749–752.

[33] Garofolo, John, et al., "CSR-I (WSJ0) Complete LDC93S6A," philadelphia: Linguistic Data Consortium, 1993.

[34] Y. Gal, "A theoretically grounded application of dropout in recurrent neural networks," *arXiv:1512.05287*, 2015.

Morten Kolbæk received the B.Eng. degree in electronic design at Aarhus University, Business and Social Sciences, AU Herning, Denmark, in 2013 and the M.Sc. in signal processing and computing from Aalborg University, Denmark, in 2015. He is currently pursuing his PhD degree at the section for Signal and Information Processing at the Department of Electronic Systems, Aalborg University, Denmark. His research interests include speech enhancement, deep learning, and intelligibility improvement of noisy speech.

Dong Yu (M'97-SM'06) is a distinguished scientist and vice general manager at Tencent AI Lab. Before joining Tencent, he was a principal researcher at Microsoft Research where he joined in 1998. His pioneer works on deep learning based speech recognition have been recognized by the prestigious IEEE Signal Processing Society 2013 and 2016 best paper award. He has served in various technical committees, editorial boards, and conference organization committees.

Zheng-Hua Tan (M'00–SM'06) is an Associate Professor and a co-head of the Centre for Acoustic Signal Processing Research (CASPR) at Aalborg University, Denmark. He was a Visiting Scientist at MIT, USA, an Associate Professor at SJTU, China, and a postdoctoral fellow at KAIST, Korea. His research interests include speech and speaker recognition, noise-robust speech processing, multimodal signal processing, social robotics, and machine learning. He has served as an Associate/Guest Editor for several journals.

Jesper Jensen is a Senior Researcher with Oticon A/S, Denmark, where he is responsible for scouting and development of signal processing concepts for hearing instruments. He is also a Professor in Dept. Electronic Systems, Aalborg University. He is also a co-head of the Centre for Acoustic Signal Processing Research (CASPR) at Aalborg University. His work on speech intelligibility prediction received the 2017 IEEE Signal Processing Society's best paper award. His main interests are in the area of acoustic signal processing, including signal retrieval from noisy observations, intelligibility enhancement of speech signals, signal processing for hearing aid applications, and perceptual aspects of signal processing.