Deep Learning for Talker-Dependent Reverberant Speaker Separation: An Empirical Study

Masood Delfarah, *Student Member, IEEE*, and DeLiang Wang, *Fellow, IEEE*

***Abstract*—Speaker separation refers to the problem of separating speech signals from a mixture of simultaneous speakers. Previous studies are limited to addressing the speaker separation problem in anechoic conditions. This paper addresses the problem of talker-dependent speaker separation in reverberant conditions, which are characteristic of real-world environments. We employ recurrent neural networks with bidirectional long short-term memory (BLSTM) to separate and dereverberate the target speech signal. We propose two-stage networks to effectively deal with both speaker separation and speech dereverberation. In the two-stage model, the first stage separates and dereverberates two-talker mixtures and the second stage further enhances the separated target signal. We have extensively evaluated the two-stage architecture, and our empirical results demonstrate large improvements over unprocessed mixtures and clear performance gain over single-stage networks in a wide range of target-to-interferer ratios and reverberation times in simulated as well as recorded rooms. Moreover, we show that time-frequency masking yields better performance than spectral mapping for reverberant speaker separation.**

***Index Terms*—Cochannel speech separation, two-stage network, deep neural networks, speech dereverberation.**

I. Introduction

SOUNDS recorded in real acoustic scenes are usually distorted by room reverberation. These distortions, which are a result of the sound reflections from surrounding walls and objects, cause a challenge to human listeners and speech processing systems alike. A more severe kind of distortion occurs when target speech signal is also corrupted by the presence of other sound sources. Perceptual studies on speech intelligibility report that human listeners, particularly those with hearing impairment, have trouble understanding speech in noisy and reverberant conditions [3], [8], [13].

Monaural speech separation aims at separating a target speech signal from a single-microphone recording that contains additive and convolutive interference. Due to its wide applicability, monaural speech separation has been studied for decades. Traditional methods include speech enhancement [24], such as spectral subtraction, and computational auditory scene analysis [32], such as pitch-based separation of voiced speech. In recent years, supervised learning techniques, particularly deep learning algorithms, have elevated speech separation performance by large margins [33]. In these studies, deep neural networks (DNNs) are typically used to learn a mapping from a mixture signal to the clean signal or its ideal time-frequency (T-F) mask. For instance, the first such study by Wang and Wang [36] used a deep feedforward network (DFN) to estimate the ideal binary mask for speech separation. Subsequent studies demonstrated that DNN based monaural separation improves human speech intelligibility in noisy environments [9], [12].

One kind of speech separation is speaker separation, where the interference is one or multiple competing talkers. Deep learning methods have also been employed to address the speaker separation problem. Previous studies [6], [16], [17], [41] trained DNN models to separate two-talker mixtures in anechoic environments. Recently, we showed that a DNN produces significant speech intelligibility benefits for human listeners [11]. These studies can be categorized as talker-dependent speaker separation as the speakers to be separated are the same as those used in training. Other kinds of speaker separation are target-dependent and talker-independent [33]. In target-dependent speaker separation [6], [41] the target speaker is assumed to be known and used during training, while interfering speakers can be unknown and untrained. In talker-independent separation, test speakers can be all untrained. Significant advances have been achieved recently on such a task [14], [23], [25], [30], [37]. Although talker-independent speaker separation is least constrained in terms of applicability, there are application scenarios where talker-dependent or target-dependent separation is a natural choice. One such scenario is when speaker separation is applied to a small number of registered speakers, as in the case of Alibaba's Tmall Genie, an Echo-like voice assistant, that features speaker recognition. Our study focuses on talker-dependent speaker separation. We will also compare with target-dependent and talker-independent models, demonstrating that broader speaker separation may come at the expense of performance loss.

Although DNNs have been used to enhance noisy and reverberant speech [42], no previous study, to our knowledge, has addressed the reverberant speaker separation problem in monaural recordings except for [38] where a single-channel scenario is evaluated as a baseline for multi-channel talker-independent speaker separation. In this paper, we investigate this problem by using recurrent neural networks (RNNs) with BLSTM [15]. As room reverberation exhibits strong temporal structure, RNNs should be more suited than DFNs for speech

Manuscript received December 18, 2018; revised April 19, 2019 and June 29, 2019; accepted August 4, 2019. Date of publication August 12, 2019; date of current version August 23, 2019. This work was supported in part by the NIDCD under Grant R01DC012048 and in part by the Ohio Supercomputer Center. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Carlos Busso. *(Corresponding author: Masood Delfarah.)*

The authors are with the Computer Science and Engineering, The Ohio State University, Columbus, OH 43210 USA (e-mail: delfarah.1@osu.edu; dwang@cse.ohio-state.edu).

Digital Object Identifier 10.1109/TASLP.2019.2934319

dereverberation. This is indeed what is found in our study. Motivated by a recent two-stage model for speech dereverberation and denoising [42], we propose two-stage networks to tackle the challenge of reverberant speaker separation. We find that two-stage networks outperform single-stage DNNs. In addition, our empirical investigation shows that T-F masking yields better results than spectral mapping. Other studies have also used two-stage networks for the separation problem. The study in [10] addresses speech-music separation where the first stage separates speech and music, and the second stage enhances each of the sources. Another study [34] performs speaker separation using a gender-mixture detection network followed by a separation network. Unlike [42], reverberation is not considered in these studies.

A preliminary version of this paper has appeared in [5]. Compared to the previous conference version, this paper introduces the second stage DNN for further enhancement of the separated target signal. In addition, more comprehensive experiments are conducted and new comparisons are made with speaker-independent and target-dependent methods.

The rest of the paper is organized as follows. In Section II we describe the baseline single-stage model and the proposed two-stage system. Section III presents experimental results. We conclude the paper in Section IV.

II. Proposed Method

Let us define the anechoic target speech signal as $s_1(t)$ and the anechoic interfering speech signal as $s_2(t)$. We assume that $s_1(t)$ and $s_2(t)$ are convolved with different room impulse responses (RIRs) $h_1(t)$ and $h_2(t)$, respectively. Then, the reverberant mixture signal $y(t)$ can be described as:

$$y(t) = h_1(t) * s_1(t) + h_2(t) * s_2(t) \tag{1}$$

where symbol * denotes convolution.

We study different DNN architectures to separate the direct sound $s_1(t)$ from $y(t)$. The goal of our separation is to improve the speech intelligibility and quality for human listeners. In other words, we intend to separate the target speaker from the interferer and, at the same time, dereverberate the target utterance since interfering speech and room reverberation both adversely affect speech perception [3], [28].

A. Feature Extraction

The mixture signal $y(t)$ is sampled at 16 kHz and windowed into 20-ms frames with 10-ms frame shift. In each time frame we extract 31-dimensional (31-D) power-normalized cepstral coefficients (PNCC) [21], 31-D gammatone frequency cepstral coefficients (GFCC) [29], and 40-D log-mel filterbank (LOG-MEL) features. These feature choices are made on the basis of our recent feature study for reverberant speech separation [4], where a detailed description for each of these features can be found. This feature study concludes that PNCC, GFCC, and LOG-MEL form a complementary feature set.

Let $\boldsymbol{F}(m)$ represent the 102-D input feature vector, where m is the time frame index. From the entire training set, mean ($\boldsymbol{\mu}_F$) and standard deviation ($\boldsymbol{\sigma}_F$) is calculated in each feature dimension. Then, the zero-mean and unit-variance normalized feature vector, $\bar{\boldsymbol{F}}(m)$, is obtained as follows:

$$\bar{\boldsymbol{F}}(m) = \frac{\boldsymbol{F}(m) - \boldsymbol{\mu}_F}{\boldsymbol{\sigma}_F} \tag{2}$$

The same $\boldsymbol{\mu}_F$ and $\boldsymbol{\sigma}_F$ values are used for feature normalization during the cross-validation and the test phase. To encode temporal information in the input signal we concatenate frames to form the following feature vector:

$$\bar{\boldsymbol{F}}_{a,b}(m) = \left[\bar{\boldsymbol{F}}(m-a), \ldots, \bar{\boldsymbol{F}}(m), \ldots, \bar{\boldsymbol{F}}(m+b)\right] \tag{3}$$

where a and b denote the number of the past and future frames with respect to the current frame.

B. Training Targets

Applying short-time Fourier transform (STFT) to $s_1(t)$, $s_2(t)$, and $y(t)$ results in the complex STFT representations as S_1, S_2, and Y, respectively. In this study, we aim at obtaining the magnitude spectrogram of the anechoic target signal $|S_1|$. Then the obtained magnitude spectrogram along with the mixture phase produces the separated target signal using the overlap-add method [1].

Our study considers two different training targets. The first is simply the log-magnitude spectrograms of the two sources $[log|S_1|, log|S_2|]$. Such a training target is commonly known as mapping-based [6]. An alternative target is the ideal ratio mask (IRM) [35]:

$$\text{IRM} = \left[\frac{|S_1|}{|S_1| + |Y - S_1|}, \frac{|S_2|}{|S_2| + |Y - S_2|}\right] \tag{4}$$

In this case, a DNN generates an estimated ratio mask, and the separated magnitude spectrograms, $|\hat{S}_1|$ and $|\hat{S}_2|$, are obtained by point-wise multiplication of the mixture magnitude spectrogram and each of the estimated ratio masks. This training target is masking-based as used in [16], [17], [39].

In [41], the mapping-based and masking-based DNNs were studied for speaker separation and it was reported that the two kinds of training targets have relative advantages in different conditions. We will compare those training objective functions for speaker separation in reverberant conditions.

C. Baseline One-Stage Networks

The one-stage system is illustrated in Fig. 1(a). In this study, RNNs with BLSTM are used due to their strong representational capacity, particularly for temporal patterns.

Our RNN consists of 4 BLSTM layers with 500 units in each layer (250 units per direction). An output layer is stacked on top of the BLSTM layers. The activation function for the units in the output layer is linear for mapping-based target. Since the IRM ranges in [0,1], we use the sigmoid function as the output layer activation for the masking-based target. During training, the network is unrolled for 100 time frames to perform the truncated backpropagation through time (BPTT) [40] and update the weights. To predict one output frame, the BLSTM network is fed by one feature frame, i.e. $\bar{\boldsymbol{F}}_{0,0}(.)$, without using neighboring

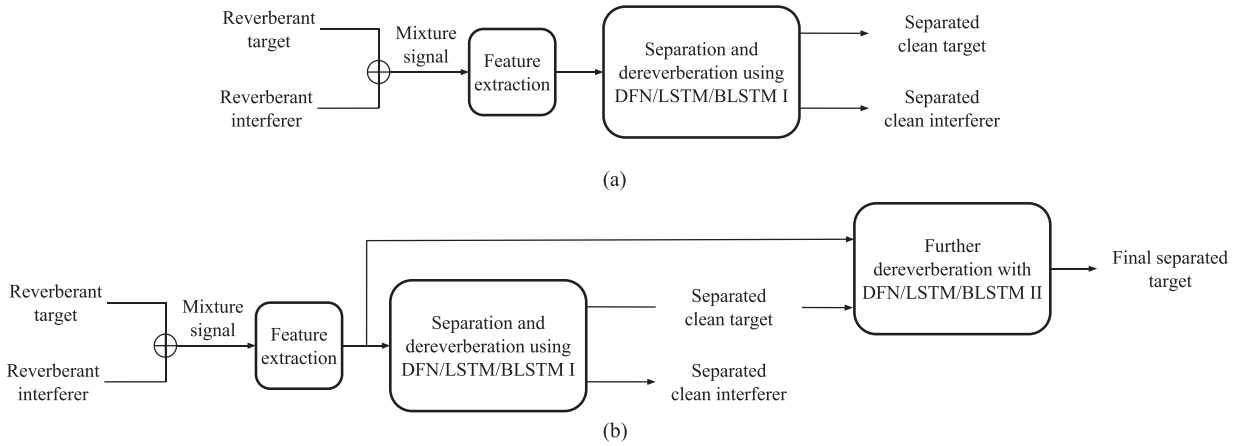

Fig. 1. Illustration of reverberant speaker separation. (a) Diagram of a baseline one-stage system and (b) diagram of the proposed two-stage system.

frames, because the memory cells in the RNN contain the past and future contextual information.

Our BLSTM makes a prediction after observing the whole utterance, which is non-casual. We also evaluate a casual RNN with unidirectional LSTMs which proceeds from the past to the current frame. To have a fair comparison with BLSTM, we provide the LSTM with the feature vector $\bar{\boldsymbol{F}}_{0,7}(.)$, which includes 7 future frames. The same network hyperparameters used in BLSTM are used in LSTM.

To contrast feedforward and recurrent networks, we also generate a baseline with DFNs. A DFN layer has less trainable parameters than an LSTM layer with the same number of units, and for a fair comparison a DFN with more units is required. We use DFNs with 4 hidden layers, each consisting of 2000 rectified linear units (ReLU) [26]. Input features used in this case are $\bar{\boldsymbol{F}}_{7,7}(.)$.

In each network, the IRM or log-spectrograms are predicted frame by frame. The mean square error (MSE) loss function $\mathcal{L}$ is:

$$\mathcal{L}(\boldsymbol{D}(m,:);\boldsymbol{\Theta}) = \frac{1}{C}\sum_{c=1}^{C}(D(m,c) - \mathcal{G}(\bar{\boldsymbol{F}}(m)))^2 \qquad (5)$$

where D is the desired target (i.e. the IRM or log-magnitude spectrogram), C is the number of frequency channels, $\boldsymbol{\Theta}$ represents the DNN parameters, and $\mathcal{G}(.)$ represents the neural network operation. The Adam optimizer [22] with the learning rate of 3×10^{-4} is used to minimize $\mathcal{L}$. In each network, the learning algorithm is run for 50 epochs, and $\boldsymbol{\Theta}$ with the least MSE on the validation set is chosen and used during the test phase.

D. Two-Stage Networks

Zhao *et al.* [42] recently proposed a two-stage network to address the problem of noisy-reverberant speech separation. Their first stage is a masking-based DFN that separates additive noise from reverberant speech signal. The denoised signal is fed into the second stage which is a mapping-based DFN to perform dereverberation. Each DFN is trained separately, and then the two networks are trained jointly. During the test phase, the network performs speech denoising and dereverberation given only a noisy-reverberant signal. One potential drawback of Zhao *et al.*'s architecture is that the second stage is provided with only the output from the first stage and does not directly operate on the input signal. The first-stage output is itself distorted and the discriminative power of the original acoustic features lost in the first stage would not be recovered by the second network. For speaker separation in reverberant conditions, we attempted to extend their approach so that the first stage is trained to separate the two reverberant speakers and the second is trained to dereverberate the target speaker. However, such an extension did not achieve satisfactory performance. Instead, we propose a different two-stage network for this problem.

The proposed two-stage system is depicted in Fig. 1(b). In the first stage, a DNN is trained to separate and dereverberate the target and interferer signals. This stage can be mapping-based or masking-based. The network output corresponding to the target speaker is converted to the log-magnitude spectrogram feature and normalized to zero mean and unit variance. This is concatenated to the mixture feature and used to train the second-stage network. The purpose of the second stage is further dereverberation of the initially separated and dereverberated target speaker. Because of the difficulty of combined separation and dereverberation, it is unlikely that a single DNN can achieve a high level of performance. With the output of the first stage, as well as the original mixture feature, the learning task of the second DNN is more focused. Hence the second stage is expected to attenuate or remove residual reverberation in the first-stage output. The second-stage network can also be mapping-based or masking-based DNN. The training targets for the target speaker are the same in both stages. Finally, the two networks are jointly trained for further fine tuning.

Four different two-stage networks can be constructed by combining the masking-based and mapping-based methods. Table I shows how the final output is calculated in each combination. We train each of four two-stage architectures using DFNs, LSTMs, and BLSTMs. Each stage network is basically the same as its corresponding single-stage network i.e. four LSTM, BLSTM, or DFN layers. Each stage network is first trained separately with the learning rate of 3×10^{-4} and then the two networks are jointly trained with the learning rate of 3×10^{-7}.

TABLE I
CALCULATION OF THE ESTIMATED TARGET SPECTROGRAM IN DIFFERENT TWO-STAGE NETWORKS. $\mathcal{G}^{(1)}(.)$ AND $\mathcal{G}^{(2)}(.)$ DENOTE THE FIRST AND SECOND STAGE DNN. $\boldsymbol{\mu}_{o_1}$ AND $\boldsymbol{\sigma}_{o_1}$ DENOTE NORMALIZATION PARAMETERS FOR THE OUTPUT OF THE FIRST STAGE, AND $\boldsymbol{\mu}_{o_2}$ AND $\boldsymbol{\sigma}_{o_2}$ THE PARAMETERS FOR THE SECOND STAGE

Combination	DNN formula
Mapping+Mapping	$\lvert\hat{S}_1\rvert = exp\left(\mathcal{G}^{(2)}\left(\left[\frac{\mathcal{G}^{(1)}\left(\bar{\boldsymbol{F}}(m)\right)-\boldsymbol{\mu}_{o_1}}{\boldsymbol{\sigma}_{o_1}}, \bar{\boldsymbol{F}}(m)\right]\right) \times \boldsymbol{\sigma}_{o_2} + \boldsymbol{\mu}_{o_2}\right)$
Mapping+Masking	$\lvert\hat{S}_1\rvert = \mathcal{G}^{(2)}\left(\left[\frac{\mathcal{G}^{(1)}\left(\bar{\boldsymbol{F}}(m)\right)-\boldsymbol{\mu}_{o_1}}{\boldsymbol{\sigma}_{o_1}}, \bar{\boldsymbol{F}}(m)\right]\right) \times \lvert Y\rvert$
Masking+Mapping	$\lvert\hat{S}_1\rvert = exp\left(\mathcal{G}^{(2)}\left(\left[\frac{log\left(\mathcal{G}^{(1)}\left(\bar{\boldsymbol{F}}(m)\right)\times\lvert Y\rvert\right)-\boldsymbol{\mu}_{o_1}}{\boldsymbol{\sigma}_{o_1}}, \bar{\boldsymbol{F}}(m)\right]\right) \times \boldsymbol{\sigma}_{o_2} + \boldsymbol{\mu}_{o_2}\right)$
Masking+Masking	$\lvert\hat{S}_1\rvert = \mathcal{G}^{(2)}\left(\left[\frac{log\left(\mathcal{G}^{(1)}\left(\bar{\boldsymbol{F}}(m)\right)\times\lvert Y\rvert\right)-\boldsymbol{\mu}_{o_1}}{\boldsymbol{\sigma}_{o_1}}, \bar{\boldsymbol{F}}(m)\right]\right) \times \lvert Y\rvert$

In the two-stage BLSTM no neighboring frames are used in the input or the output. To train the two-stage LSTMs, $\bar{\boldsymbol{F}}_{0,7}(.)$ is fed to the first-stage network to predict the output for the current and the 3 future frames. This output is concatenated with $\bar{\boldsymbol{F}}_{0,3}(.)$ and passed to the second-stage to predict a single output frame. On the other hand, the two-stage DFNs use $\bar{\boldsymbol{F}}_{7,7}(.)$ in the first stage to predict 7 consecutive output frames, centered at the current frame. Then the second-stage network concatenates $\bar{\boldsymbol{F}}_{3,3}(.)$ with the first-stage output to predict a single output frame.

In order to evaluate the potential speech intelligibility benefits of separated speech, we use the Extended Short-time Objective Intelligibility (ESTOI) [20] metric, which is shown to strongly correlate with human intelligibility scores. ESTOI is a number mainly between 0 and 1 and a higher score indicates better intelligibility. We use Perceptual Evaluation of Speech Quality (PESQ) [27] to evaluate the quality of the separated target signals. PESQ score is a number between -0.5 and 4.5 and higher score indicates better speech quality. We also use signal-to-distortion ratio improvement, or ΔSDR, which is another widely used speech separation evaluation metric [31]. The anechoic target is used as the reference signal in these evaluations.

III. EVALUATION RESULTS AND COMPARISONS

A. Experimental Setup

The speech corpus used in this empirical study consists of 1440 IEEE sentences [19] uttered by a male and a female speaker. In the experiments, we arbitrarily designate the male speaker as the target and the female as the interferer. From this set, we randomly choose and set aside 120 female and 120 male utterances for testing and the rest are used for training. To generate the training mixtures, one male utterance and one female utterance are randomly picked. In the case that the interfering signal is shorter, it is repeated until it covers the whole target sentence. We use the image method [2] to generate simulated RIRs in a room with the dimensions of (6.5, 8.5, 3) m, by placing a microphone at (3, 4, 1.5) m. The reverberation time (T_{60}) is sampled from the continuous range [0.3, 1.0] s. The male speaker is randomly placed at 1 m and the interferer at 2 m distance from the microphone at the same elevation. Then the reverberant male and female signals are mixed at a random target-to-interferer energy ratio (TIR) sampled from the continuous range [-12, 12] dB.

TABLE II
AVERAGE DRR VALUES (DB) IN DIFFERENT ROOM CONDITIONS

	Simulated room			Recorded room			
T_{60} (s)	0.3	0.6	0.9	0.32	0.47	0.68	0.89
Target DRR	3.3	-1.4	-3.7	6.1	5.3	8.8	6.1
Interferer DRR	-2.7	-7.4	-9.7				

In total, 100,000 mixtures are generated for training and 1000 mixtures for validation. To train the single-stage networks, the entire training set is used. In the two-stage cases, the first stage is trained with half of the training set. Then, the second half is passed through the first stage and used to train the second stage. Finally, the joint training is done using the whole training set.

To generate simulated reverberant test mixtures, we use a different simulated room with the dimensions of (6, 8, 3) m, where the microphone position is set to (3.5, 2.5, 1.2) m. Target speaker is randomly placed at a 1 m distance and the interferer at a 2 m distance from the microphone. Note that since the test room is different from the training room, no RIRs are common between the test and training sets. Test T_{60} is chosen from {0.3, 0.6, 0.9} s and test TIR from {-12, -6} dB. In each condition, the networks are tested using 2000 mixtures and average scores are reported.

In order to further evaluate the generalization of the systems to real room conditions, we also perform experiments using recorded RIRs. For this purpose, we use the recorded RIRs from [18], which consist of recordings from four rooms with $T_{60} =$ {0.32, 0.47, 0.68, 0.89} s. Aside from T_{60}, direct-to-reverberant energy ratio (DRR) is an important characteristic of a reverberant signal. In general, a lower DRR entails a more challenging processing condition. Table II shows DRRs for our simulated and recorded test rooms.

B. Single-Stage Reverberant Speaker Separation

To provide a baseline for reverberant speaker separation we first present results in anechoic conditions. Table III shows ESTOI, PESQ, and SDR scores with DNNs trained and tested in

TABLE III
ESTOI, PESQ, AND ΔSDR SCORES FOR SPEAKER SEPARATION IN ANECHOIC CONDITIONS. SINGLE-STAGE NETWORKS ARE TRAINED WITH ANECHOIC DATA. BOLDFACE HIGHLIGHTS THE BEST RESULT IN EACH CONDITION

		ESTOI (%)			PESQ			ΔSDR (dB)		
TIR (dB)		-12	-6	Average	-12	-6	Average	-12	-6	Average
Unprocessed		24.6	36.4	30.5	1.35	1.58	1.46			
DFN	Mapping	62.4	73.6	68.0	2.35	2.65	2.50	14.15	10.77	12.46
	Masking	63.7	76.2	69.9	2.40	2.75	2.57	15.83	12.94	14.28
LSTM	Mapping	67.9	77.3	72.6	2.47	2.76	2.61	14.60	11.14	12.87
	Masking	69.1	79.7	74.4	2.52	2.86	2.69	16.26	13.32	14.79
BLSTM	Mapping	71.6	79.9	75.7	2.61	2.87	2.74	15.24	11.70	13.47
	Masking	**72.0**	**81.5**	**76.7**	**2.62**	**2.94**	**2.78**	**16.77**	**13.58**	**15.17**

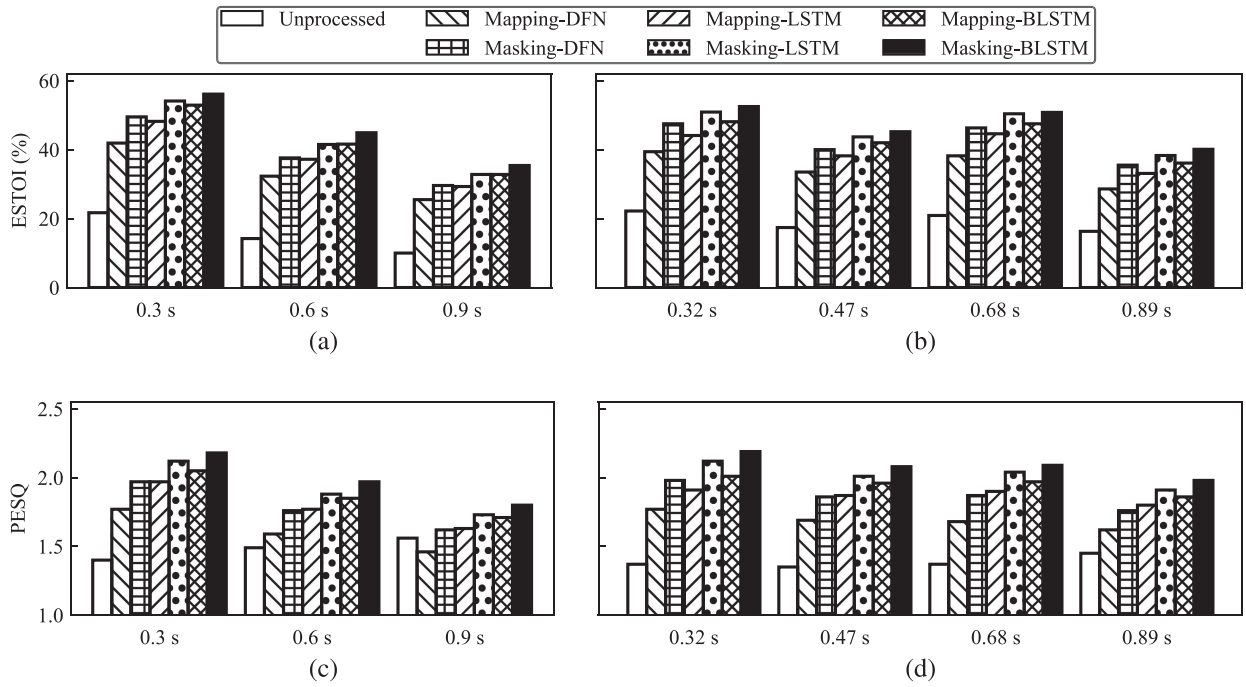

Fig. 2. Separation performance in reverberant environments for single-stage mapping-based and masking-based DFN, LSTM, and BLSTM. Test results are shown for different T_{60} values with $\mathrm{TIR} = -12$ dB. (a) ESTOI scores in simulated RIR conditions, (b) ESTOI scores in recorded RIR conditions, (c) PESQ scores in simulated RIR conditions, and (d) PESQ scores in recorded RIR conditions.

anechoic conditions. The results indicate that the masking-based systems perform better than the mapping-based systems. In addition, BLSTM and LSTM outperform DFN, showing that recurrent networks can better separate the speakers. A masking-based BLSTM achieves the best intelligibility and quality scores.

We train single-stage mapping and masking-based DFNs, LSTMs, and BLSTMs to perform speaker separation and speech dereverberation. Objective scores for simulated and recorded RIR conditions for TIR of -12 dB are presented in Figure 2. Similar to anechoic conditions, BLSTMs outperform LSTMs and DFNs, and T-F masking outperforms spectral mapping. These observations are consistent across simulated and real room conditions, and the amounts of improvement are also comparable between simulated and real room conditions. Note that the systems are trained using only simulated RIRs, and the substantial improvements in the real rooms suggest that the trained DNNs are capable of generalizing to different reverberant conditions.

From Fig. 2, we observe that the scores decrease with the increase of T_{60}. This trend is most evident when comparing with the performance in anechoic conditions in Table III. For example, the masking-based BLSTM achieves 47.4% ESTOI improvement in the anechoic condition with $\mathrm{TIR} = -12$ dB, and this improvement is reduced to 36.3% at $T_{60} = 0.3$ s with the same TIR. This indicates that separation in reverberant conditions is in general more challenging.

C. Two-Stage Reverberant Speaker Separation

We train the four combinations of masking-based and mapping-based networks shown in Table I. Table IV gives ESTOI scores using two-stage DNNs in simulated reverberant test conditions. The results from single-stage networks are also included in this table for reference. As seen in the table, two-stage networks in general outperform single-stage networks. Our experiments demonstrate that using two masking-based

TABLE IV
ESTOI (%) SCORES FOR DIFFERENT TWO-STAGE AND SINGLE-STAGE DFNS, LSTMS AND BLSTMS IN SIMULATED REVERBERANT CONDITIONS. IN EACH TWO-STAGE DNN, * INDICATES THAT THE SCORE IS SIGNIFICANTLY BETTER THAN THE MASKING-BASED SINGLE-STAGE DNN BASELINE SCORE (WITH THE SIGNIFICANCE LEVEL OF $p < 0.0005$)

T_{60} (s)		0.3		0.6		0.9		
TIR (dB)		-12	-6	-12	-6	-12	-6	Average
Unprocessed		21.7	33.3	14.2	23.5	10.0	17.8	20.1
DFN	Single-stage Mapping	41.9	54.9	32.3	46.4	25.5	38.5	39.9
	Single-stage Masking	49.5	62.8	37.6	51.5	29.6	43.1	45.7
	Mapping+Mapping	44.8	57.7	34.1	47.6	26.6	39.6	41.7
	Mapping+Masking	49.4	62.7	37.1	50.9	29.0	42.2	45.2
	Masking+Mapping	49.6	62.0	37.5	51.2	29.2	42.9	45.4
	Masking+Masking	53.1*	66.2*	40.0*	53.9*	31.3*	45.0*	48.2*
LSTM	Single-stage Mapping	48.2	60.0	37.2	50.5	29.3	42.5	44.6
	Single-stage Masking	54.1	66.3	41.5	55.2	32.8	46.7	49.4
	Mapping+Mapping	49.5	60.8	37.5	50.6	29.2	42.4	45.0
	Mapping+Masking	53.1	65.1	40.4	53.6	31.9	45.1	48.2
	Masking+Mapping	53.2	64.4	40.2	53.2	31.4	44.6	47.8
	Masking+Masking	55.5*	67.6*	42.2*	55.6*	33.3*	46.8*	50.2*
BLSTM	Single-stage Mapping	52.9	64.5	41.6	54.8	32.8	46.8	48.9
	Single-stage Masking	56.1	68.0	44.9	58.4	35.4	50.1	52.1
	Mapping+Mapping	53.7	64.3	40.5	54.6	30.4	45.4	48.1
	Mapping+Masking	54.2	65.4	41.8	55.4	32.2	46.9	49.3
	Masking+Mapping	57.4*	68.6*	44.8	58.5	34.5	49.4	52.2
	Masking+Masking	**58.0***	**69.8***	**45.5***	**59.3***	**36.0***	**50.8***	**53.2***

TABLE V
PESQ SCORES FOR DIFFERENT TWO-STAGE AND SINGLE-STAGE DFNS, LSTMS AND BLSTMS IN SIMULATED REVERBERANT CONDITIONS

T_{60} (s)		0.3		0.6		0.9		
TIR (dB)		-12	-6	-12	-6	-12	-6	Average
Unprocessed		1.40	1.51	1.49	1.51	1.56	1.56	1.50
DFN	Single-stage Mapping	1.77	2.07	1.59	1.91	1.46	1.77	1.76
	Single-stage Masking	1.97	2.29	1.76	2.07	1.62	1.91	1.94
	Mapping+Mapping	1.84	2.15	1.65	1.95	1.52	1.80	1.82
	Mapping+Masking	1.95	2.28	1.75	2.04	1.62	1.89	1.92
	Masking+Mapping	1.96	2.27	1.73	2.04	1.58	1.87	1.91
	Masking+Masking	2.06*	2.40*	1.82*	2.14*	1.67*	1.97*	2.01*
LSTM	Single-stage Mapping	1.97	2.23	1.77	2.04	1.63	1.90	1.92
	Single-stage Masking	2.12	2.42	1.88	2.18	1.73	2.02	2.06
	Mapping+Mapping	1.97	2.23	1.76	2.04	1.63	1.90	1.92
	Mapping+Masking	2.08	2.37	1.85	2.13	1.70	1.97	2.02
	Masking+Mapping	2.05	2.33	1.81	2.09	1.66	1.94	1.98
	Masking+Masking	2.14	2.46*	1.88	2.19	1.72	2.02	2.07
BLSTM	Single-stage Mapping	2.05	2.31	1.85	2.13	1.71	1.99	2.01
	Single-stage Masking	2.18	2.47	1.97	2.26	1.81	2.10	2.13
	Mapping+Mapping	2.14	2.42	1.89	2.18	1.68	1.99	2.05
	Mapping+Masking	2.17	2.48	1.93	2.23	1.74	2.06	2.10
	Masking+Mapping	2.22	2.48	1.94	2.24	1.76	2.06	2.12
	Masking+Masking	**2.26***	**2.57**	**2.00***	**2.31***	**1.82***	**2.14***	**2.18***

systems is the best combination. Among DFNs, LSTMs, and BLSTMs, we observe that the BLSTMs provide the smallest advantage by using a two-stage system. On the other hand, the combination of two masking-based BLSTMs achieves the best performance among all the DNNs evaluated. We also observe that using a mapping-based network either in the first stage or the second stage does not perform as well as a masking-based network. Table V shows the corresponding PESQ results, which exhibit a trend similar to that of ESTOI results. Again, the best results are achieved by the masking+masking BLSTM system.

As mentioned earlier, in a two-stage system, two single-stage networks are first trained separately and then jointly. The results indicate that the two-stage masking+mapping system outperforms single-stage mapping. To see how much improvement is due to joint training, Figure 3 compares a masking+mapping DFN with and without joint training as well as a single-stage DFN. A simulated room condition with $T_{60} = 0.9$ s and $\mathrm{TIR} = -6$ dB is used for this comparison. As seen in the figure, the performance gain of the two-stage network is mostly because of joint training. We have also observed this in other DNN architectures.

TABLE VI
ESTOI (%) SCORES FOR DIFFERENT TWO-STAGE AND SINGLE-STAGE DFNS, LSTMS AND BLSTMS IN RECORDED REVERBERANT CONDITIONS

T_{60} (s)		0.32		0.47		0.68		0.89		Average
TIR (dB)		-12	-6	-12	-6	-12	-6	-12	-6	
Unprocessed		22.2	33.0	17.4	27.2	20.9	32.1	16.3	26.7	24.5
DFN	Single-stage Mapping	39.4	50.3	33.5	44.7	38.2	49.7	28.6	39.7	40.5
	Single-stage Masking	47.5	59.8	40.0	52.2	46.3	59.1	35.5	48.2	48.6
	Mapping+Mapping	41.3	51.8	35.7	46.6	41.4	52.9	30.8	41.6	42.8
	Mapping+Masking	47.2	59.2	39.8	51.7	46.3	59.0	35.1	47.6	48.2
	Masking+Mapping	44.9	55.0	39.0	49.7	45.3	56.3	33.1	44.0	45.9
	Masking+Masking	50.1*	62.0*	42.3*	54.1*	49.4*	61.8*	36.8*	49.5*	50.7*
LSTM	Single-stage Mapping	44.1	53.8	38.2	49.0	44.6	55.3	33.1	43.8	45.2
	Single-stage Masking	50.9	62.6	43.7	55.6	50.4	62.7	38.3	51.1	51.9
	Mapping+Mapping	44.2	53.8	38.6	49.3	45.1	56.0	33.6	44.3	45.6
	Mapping+Masking	49.9	61.2	42.8	54.5	49.6	61.9	38.0	50.4	51.0
	Masking+Mapping	47.2	57.0	41.6	52.2	47.9	58.6	35.5	46.4	48.3
	Masking+Masking	51.7	63.3	44.5	56.3	51.3*	63.6*	39.0	51.5	52.6
BLSTM	Single-stage Mapping	48.1	57.8	42.0	52.7	47.5	58.6	36.1	47.3	48.8
	Single-stage Masking	52.5	64.0	45.2	57.0	50.8	63.1	**40.1**	52.9	53.2
	Mapping+Mapping	49.7	59.8	43.1	54.1	49.3	60.5	35.6	49.1	50.1
	Mapping+Masking	50.6	62.1	43.4	55.0	50.1	61.9	37.8	52.1	51.6
	Masking+Mapping	51.1	61.2	44.7	55.8	50.8	61.9	37.6	49.4	51.6
	Masking+Masking	**53.2**	**64.7***	**46.0**	**57.8***	**52.8***	**64.9***	39.9	**53.0**	**54.0***

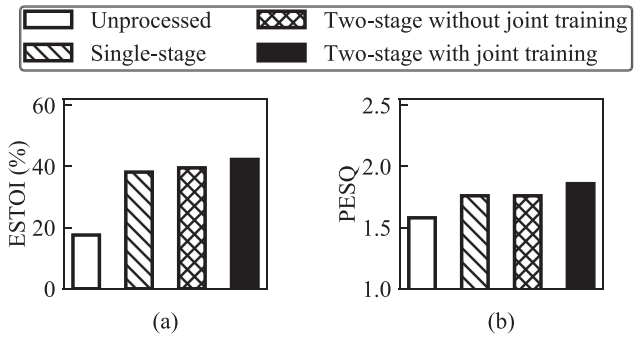

Fig. 3. A single-stage mapping-based DFN is compared with a two-stage masking+mapping DFN with and without joint training. The test condition is a simulated room with $T_{60} = 0.9$ s and TIR $= -6$ dB.

Fig. 4 illustrates IRM prediction from a mixture signal from a reverberant male and female utterance. Broadly speaking, both single-stage masking and two-stage masking+masking networks are able to estimate the IRM well. On the other hand, the two-stage network is better at recovering finer spectrogram structures in the IRM and this explains the superior performance of the two-stage approach.

Next, we present separation results with recorded RIRs. It is worth emphasizing that no recorded RIR is used in training. Table VI shows ESTOI results for real room conditions. Similar to simulated reverberant conditions, the two-stage BLSTM system achieves the best results. Again, the masking+masking BLSTM system outperforms other DNN architectures. In Table VII, we present PESQ results for these conditions. The trend of the PESQ scores is similar to that of ESTOI scores in Table VI.

D. Comparisons with Talker-Independent and Target-Dependent Separation

Our model is trained and tested using a fixed pair of speakers whose utterances have been used during training. Such speaker separation is talker-dependent. Deep learning models have been recently developed to perform talker-independent speaker separation, i.e. test speakers can be different from training speakers. One prominent method is utterance-level permutation invariant training (uPIT)[23]. This algorithm calculates the training loss by accounting for speaker permutations across time frames and then optimizes the network using the minimum loss of different permutations. To get an idea on whether our talker-dependent separation yields an expected improvement over talker-independent separation, we train a uPIT model and compare with our two-stage network. To this end, we use the WSJ0 corpus [7] to generate reverberant mixtures for uPIT. All signals are downsampled to 16 kHz and frame size is set to 20 ms with the frame shift of 10 m. Experiments are performed as described in Section III-A. The uPIT network with BLSTM is optimized to estimate the IRM for each of the two anechoic utterances via minimizing the utterance-level loss.

A target-dependent speaker separation model aims to separate a trained target speaker from an open set of interfering speakers [6], [41]. To train a target-dependent model we use the IEEE male speaker as the target and WSJ0 speakers as interferers. A BLSTM network is trained to predict the IRM for the anechoic target utterances via the loss function in (5).

The results are shown in Fig. 5. As expected, in both simulated and recorded RIR conditions, our two-stage talker-dependent model outperforms the target-dependent model, which in turn outperforms uPIT. Furthermore, uPIT improves both ESTOI and

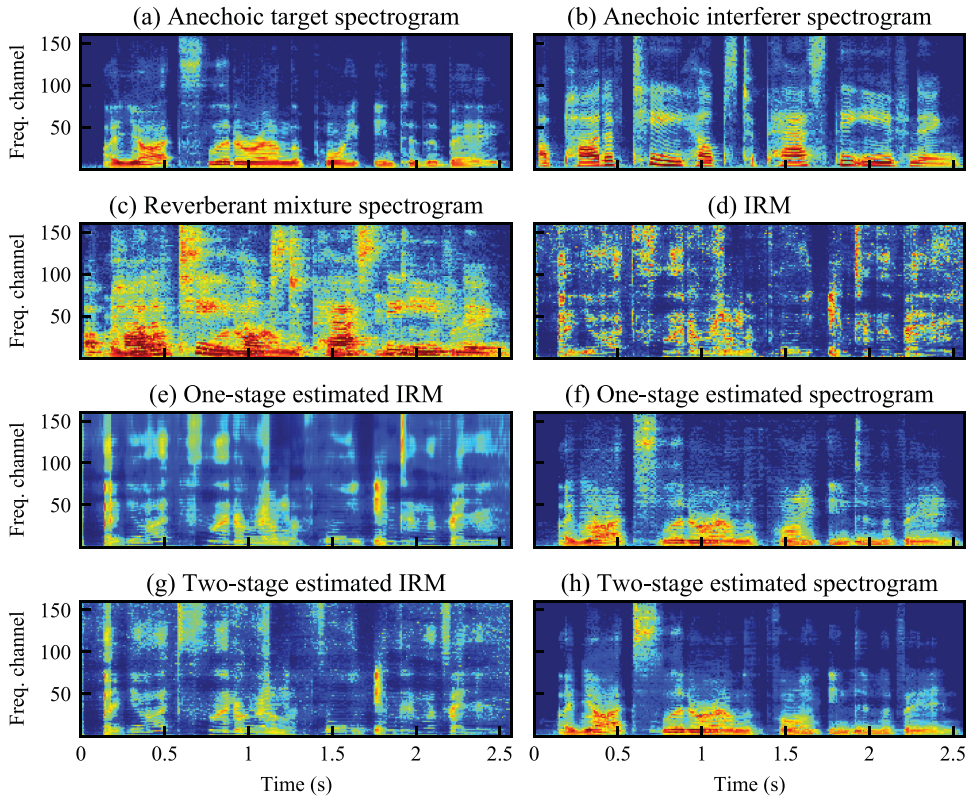

Fig. 4. IRM prediction for reverberant speaker separation using BLSTMs. The test condition is for a simulated room with $T_{60} = 0.3$ s and $\mathrm{TIR} = -6$ dB. Each panel is indicated by its corresponding label.

TABLE VII
PESQ SCORES FOR DIFFERENT TWO-STAGE AND SINGLE-STAGE DFNS, LSTMS AND BLSTMS IN RECORDED REVERBERANT CONDITIONS

T_{60} (s)		0.32		0.47		0.68		0.89		
TIR (dB)		-12	-6	-12	-6	-12	-6	-12	-6	Average
Unprocessed		1.37	1.55	1.35	1.46	1.37	1.52	1.45	1.54	1.45
DFN	Single-stage Mapping	1.77	2.04	1.69	1.98	1.68	1.95	1.62	1.87	1.82
	Single-stage Masking	1.98	2.31	1.86	2.18	1.87	2.20	1.76	2.05	2.03
	Mapping+Mapping	1.82	2.07	1.76	2.03	1.76	2.04	1.67	1.92	1.88
	Mapping+Masking	1.99	2.30	1.87	2.17	1.88	2.20	1.78	2.05	2.03
	Masking+Mapping	1.90	2.16	1.83	2.12	1.86	2.14	1.76	2.00	1.97
	Masking+Masking	2.06*	2.38*	1.93*	2.25*	1.96*	2.30*	1.82*	2.11*	2.10*
LSTM	Single-stage Mapping	1.91	2.14	1.87	2.12	1.90	2.14	1.80	2.03	1.99
	Single-stage Masking	2.12	2.42	2.01	2.32	2.04	2.35	1.91	2.19	2.17
	Mapping+Mapping	1.89	2.11	1.85	2.11	1.89	2.14	1.80	2.01	1.97
	Mapping+Masking	2.08	2.37	1.97	2.27	2.01	2.31	1.89	2.16	2.13
	Masking+Mapping	1.96	2.20	1.92	2.19	1.95	2.21	1.84	2.07	2.04
	Masking+Masking	2.14	2.45	2.03	2.35*	2.06	2.38*	1.93	2.22	2.19
BLSTM	Single-stage Mapping	2.01	2.22	1.96	2.21	1.97	2.21	1.86	2.09	2.07
	Single-stage Masking	2.19	2.48	2.08	2.37	2.09	2.38	1.98	2.26	2.23
	Mapping+Mapping	2.03	2.27	1.97	2.24	1.97	2.27	1.86	2.08	2.09
	Mapping+Masking	2.16	2.49*	2.05	2.39*	2.10	2.42*	1.95	2.24	2.22
	Masking+Mapping	2.13	2.36	2.05	2.32	2.07	2.34	1.94	2.18	2.17
	Masking+Masking	**2.23***	**2.53***	**2.12***	**2.43***	**2.15***	**2.46***	**2.01**	**2.29***	**2.28***

PESQ scores of unprocessed mixtures. All these improvements are statistically significant ($p < 0.0005$). These consistent results suggest that speaker-specific information, when available, contributes to speaker separation performance. The comparative results between the speaker-dependent and target-dependent models further suggest that interferer information also plays a role in separation performance.

Our evaluation so far is on a pair of male-female speakers. We expect a similar pattern of results for same-gender pairs, although the results are expected to be a little worse. To verify

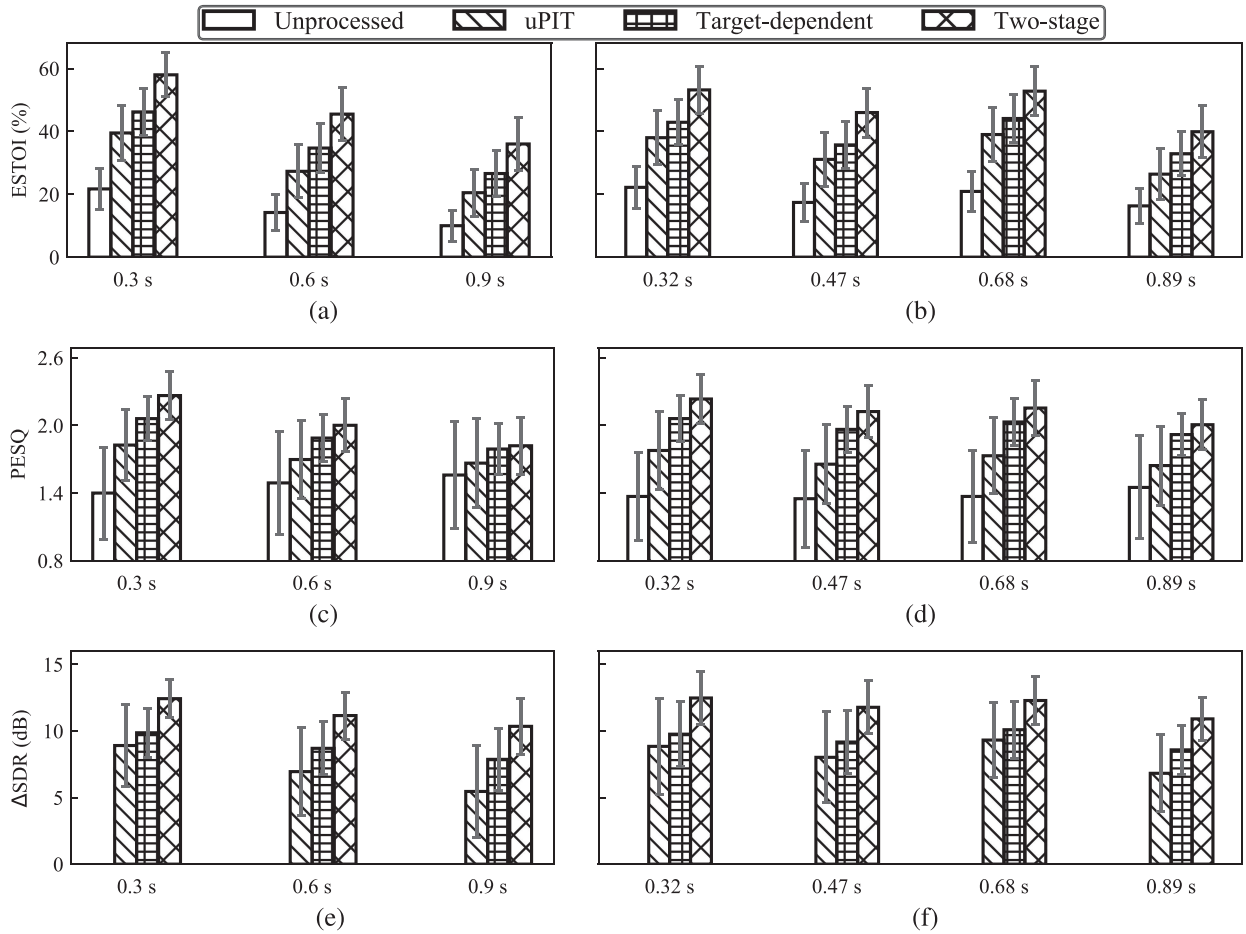

Fig. 5. Comparison of the proposed two-stage network with target-dependent and talker-independent speaker separation. All networks use BLSTM and test TIR is −12 dB. (a) ESTOI scores in simulated RIR conditions, (b) ESTOI scores in recorded RIR conditions, (c) PESQ scores in simulated RIR conditions, (d) PESQ scores in recorded RIR conditions, (e) ΔSDR scores in simulated RIR conditions, and (f) ΔSDR scores in recorded RIR conditions. Error bars depict the standard deviation.

TABLE VIII
SEPARATION RESULTS AND COMPARISONS FOR A PAIR OF MALE-MALE SPEAKERS. RESULTS ARE SHOWN AT TIR = −6 dB, AVERAGED OVER ALL T_{60} CONDITIONS

Metric	ESTOI	PESQ	ΔSDR
Unprocessed	23.9	1.21	0.00
Talker-independent	36.4	1.56	3.19
Target-dependent	39.2	1.77	3.41
One-stage talker-dependent	49.5	2.00	5.26
Two-stage talker-dependent	**50.1***	**2.02**	**5.40***

this, we choose a new pair of male speakers, both uttering the IEEE corpus, with one of them designated as the target speaker. This evaluation is conducted in simulated reverberant conditions at three T_{60}s (0.3, 0.6, 0.9 s) and recorded reverberant rooms at four T_{60}s (0.32, 0.47, 0.68, 0.89 s) with one TIR (−6 dB). The same-gender results and comparisons are presented in Table VIII. Like the male-female results, talker-dependent speaker separation outperforms target-dependent separation, which in turn yields better results than talker-independent separation. Furthermore, the two-stage network performs better than the single-stage network.

IV. CONCLUDING REMARKS

In this paper, we have proposed two-stage deep neural networks for the speaker separation in reverberant conditions. We have compared the performances of BLSTMs, LSTMs, and DFNs, and our experimental results show that recurrent networks outperform feedforward networks in a wide range of conditions, with BLSTMs performing the best. We have also shown that masking-based separation outperforms mapping-based separation. Our empirical study shows that talker-dependent speaker separation in reverberant conditions yields better results than target-dependent models, which in turn perform better than talker-independent separation. This observation is expected as talker-dependent models operate in more constrained conditions. To our knowledge, this is the first study to address monaural speaker separation in reverberant conditions using RNNs. In the future we plan to extend the current system to speech separation conditions with both background noise and interfering speakers.

ACKNOWLEDGMENT

The author would like to thank Y. Liu for providing the uPIT code and Eric Johnson for help with statistical analysis.

REFERENCES

[1] J. B. Allen, "Short term spectral analysis, synthesis, and modification by discrete fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 25, no. 3, pp. 235–238, Jun. 1977.

[2] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, pp. 943–950, 1979.

[3] J. F. Culling, K. I. Hodder, and C. Y. Toh, "Effects of reverberation on perceptual segregation of competing voices," *J. Acoust. Soc. Amer.*, vol. 114, pp. 2871–2876, 2003.

[4] M. Delfarah and D. L. Wang, "Features for masking-based monaural speech separation in reverberant conditions," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 5, pp. 1085–1094, May 2017.

[5] M. Delfarah and D. L. Wang, "Recurrent neural networks for cochannel speech separation in reverberant environments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 5404–5408.

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

[7] J. Garofolo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) complete LDC93S6A," Philadelphia, PA, USA: Linguistic Data Consortium, 1993.

[8] E. L. George, S. T. Goverts, J. M. Festen, and T. Houtgast, "Measuring the effects of reverberation and noise on sentence intelligibility for hearing-impaired listeners," *J. Speech Lang. Hear. Res.*, vol. 53, pp. 1429–1439, 2010.

[9] T. Goehring, F. Bolner, J. J. Monaghan, B. van Dijk, A. Zarowski, and S. Bleeck, "Speech enhancement based on neural networks improves speech intelligibility in noise for cochlear implant users," *Hearing Res.*, vol. 344, pp. 183–194, 2017.

[10] E. M. Grais *et al.*, "Two-stage single-channel audio source separation using deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 9, pp. 1773–1783, Sep. 2017.

[11] E. W. Healy, M. Delfarah, J. L. Vasko, B. L. Carter, and D. L. Wang, "An algorithm to increase intelligibility for hearing-impaired listeners in the presence of a competing talker," *J. Acoust. Soc. Amer.*, vol. 141, pp. 4230–4239, 2017.

[12] E. W. Healy, S. E. Yoho, Y. Wang, and D. L. Wang, "An algorithm to improve speech recognition in noise for hearing-impaired listeners," *J. Acoust. Soc. Amer.*, vol. 134, pp. 3029–3038, 2013.

[13] K. S. Helfer and L. A. Wilber, "Hearing loss, aging, and speech perception in reverberation and noise," *J. Speech Lang. Hear. Res.*, vol. 33, pp. 149–155, 1990.

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

[15] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, pp. 1735–1780, 1997.

[16] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 1562–1566.

[17] 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, Lang. Process.*, vol. 23, no. 12, pp. 2136–2147, Dec. 2015.

[18] C. Hummersone, R. Mason, and T. Brookes, "Dynamic precedence effect modeling for source separation in reverberant environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1867–1871, Sep. 2010.

[19] IEEE, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.*, vol. AE-17, no. 3, pp. 225–246, Sep. 1969.

[20] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 11, pp. 2009–2022, Nov. 2016.

[21] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 7, pp. 1315–1329, Jul. 2016.

[22] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICML*, 2015.

[23] M. Kolbæk *et al.*, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 10, pp. 1901–1913, Oct. 2017.

[24] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL, USA: CRC Press, 2013.

[25] Y. Luo and N. Mesgarani, "TasNet: Surpassing ideal time-frequency masking for speech separation," 2018, *arXiv:1809.07454*.

[26] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn.*, 2010, pp. 807–814.

[27] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2001, pp. 749–752.

[28] M. Sayles and I. M. Winter, "Reverberation challenges the temporal representation of the pitch of complex sounds," *Neuron*, vol. 58, pp. 789–801, 2008.

[29] Y. Shao and D. L. Wang, "Robust speaker identification using auditory features and computational auditory scene analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2008, pp. 1589–1592.

[30] Z. Shi, H. Lin, L. Liu, R. Liu, and J. Han, "FurcaNeXt: End-to-end monaural speech separation with dynamic gated dilated temporal convolutional networks," 2019, *arXiv:1902.04891*.

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

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

[33] D. L. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018.

[34] 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. Audio, Speech, Lang. Process.*, vol. 25, no. 7, pp. 1535–1546, Jul. 2017.

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

[36] Y. Wang and D. L. Wang, "Towards scaling up classification-based speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 7, pp. 1381–1390, Jul. 2013.

[37] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Alternative objective functions for deep clustering," in *Proc. IEEE Int. Conf. Acoust., Speech., Signal Process.*, 2018, pp. 686–690.

[38] Z.-Q. Wang and D. L. Wang, "Combining spectral and spatial features for deep learning based blind speaker separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 2, pp. 457–468, Feb. 2019.

[39] F. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *Proc. IEEE Global Conf. Signal Inf. Process.*, 2014, pp. 577–581.

[40] R. J. Williams and J. Peng, "An efficient gradient-based algorithm for on-line training of recurrent network trajectories," *Neural Comput.*, vol. 2, pp. 490–501, 1990.

[41] X.-L. Zhang and D. L. Wang, "A deep ensemble learning method for monaural speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 5, pp. 967–977, May 2016.

[42] Y. Zhao, Z.-Q. Wang, and D. L. Wang, "Two-stage deep learning for noisy-reverberant speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 1, pp. 53–62, Jan. 2019.

Authors' photographs and biographies not available at the time of publication.