JASA ARTICLE

A two-stage deep learning algorithm for talker-independent speaker separation in reverberant conditions

Masood Delfarah,[a] Yuzhou Liu,[b] and DeLiang Wang[c]
Department of Computer Science and Engineering, The Ohio State University, Columbus, Ohio 43210, USA

ABSTRACT:
Speaker separation is a special case of speech separation, in which the mixture signal comprises two or more speakers. Many talker-independent speaker separation methods have been introduced in recent years to address this problem in anechoic conditions. To consider more realistic environments, this paper investigates talker-independent speaker separation in reverberant conditions. To effectively deal with speaker separation and speech dereverberation, extending the deep computational auditory scene analysis (CASA) approach to a two-stage system is proposed. In this method, reverberant utterances are first separated and separated utterances are then dereverberated. The proposed two-stage deep CASA system significantly outperforms a baseline one-stage deep CASA method in real reverberant conditions. The proposed system has superior separation performance at the frame level and higher accuracy in assigning separated frames to individual speakers. The proposed system successfully generalizes to an unseen speech corpus and exhibits similar performance to a talker-dependent system. © *2020 Acoustical Society of America.*
https://doi.org/10.1121/10.0001779
(Received 27 February 2020; revised 29 July 2020; accepted 4 August 2020; published online 3 September 2020)
[Editor: Michael I. Mandel] Pages: 1157–1168

I. INTRODUCTION

Speech perception by human listeners is affected by the presence of competing speakers (Brungart, 2001; Miller, 1947). In realistic scenarios, room reverberation corrupts the spectro-temporal structure of the speech signal and degrades speech intelligibility (Helfer and Wilber, 1990). Competing speech and room reverberation have a confounding effect on speech perception (Culling *et al.*, 2003; Festen and Plomp, 1990; Moore, 2007). These effects can be mitigated by enhancing target speech and separating it from interfering sources. Effective speech separation especially benefits people with hearing impairment because these listeners have a particular difficulty dealing with competing talkers and room reverberation.

Traditional speaker separation is based on statistical models, such as hidden Markov models (HMMs; Weiss and Ellis, 2010) and nonnegative matrix factorization (NMF; Smaragdis, 2006). Recent studies utilize deep neural networks (DNNs) following their use in speech separation (Wang and Wang, 2013). In DNN-based speech separation, a DNN is typically trained to learn a mapping function from noisy speech features to some training target. At the inference time, the DNN is fed with mixture features to estimate the training target, from which the enhanced speech signal is constructed. This approach works well for separating two trained speakers in anechoic conditions (Du *et al.*, 2014; Huang *et al.*, 2014, 2015; Zhang and Wang, 2016), a case of talker-dependent speaker separation (Wang and Chen, 2018). DNN-based speaker separation yields significant speech intelligibility improvements for both normal-hearing and hearing-impaired listeners (Healy *et al.*, 2017). More general speaker separation includes target-dependent and gender-dependent methods. In target-dependent separation a trained target speaker can be separated from an untrained speaker (Du *et al.*, 2014; Zhang and Wang, 2016), and gender-dependent separation aims to separate any female speaker from any male speaker (Tan and Wang, 2018).

The most general form is talker-independent speaker separation, in which test speakers cannot be seen during training. It is not straightforward to extend the DNN-based speech separation framework to the talker-independent situation, in which the permutation problem arises during DNN training, as the output layers cannot be uniquely assigned to individual speakers (Kolbæk *et al.*, 2017). Deep clustering (Hershey *et al.*, 2016) and permutation invariant training (PIT; Kolbæk *et al.*, 2017) have been proposed to address this problem. Deep clustering learns a mapping from time-frequency (T-F) units to embedding vectors. Clustering these embeddings yields an estimate of the ideal binary mask (IBM; Wang, 2005), which is utilized to reconstruct separated speech. PIT evaluates all possible losses associated with speaker-output assignments and then optimizes the neural network using the minimum of the losses. One version of this algorithm, known as frame-level permutation invariant training (tPIT; Yu *et al.*, 2017), aims to separate speakers at the frame level without addressing the speaker tracking problem, the issue of tracking the same speaker across consecutive frames. To address the tracking problem, utterance-level permutation invariant training (uPIT; Kolbæk *et al.*, 2017) calculates the loss over the entire mixture signal instead of individual frames. This forces the

[a]Electronic mail: delfarah.1@osu.edu, ORCID: 0000-0002-8354-0832.
[b]ORCID: 0000-0002-7030-9121.
[c]Also at: Center for Cognitive and Brain Sciences, The Ohio State University, Columbus, OH 43210, USA, ORCID: 0000-0001-8195-6319.

frames for each speaker to correspond to the same output layer for the whole utterance, eliminating the need for speaker tracking. However, evaluations show that uPIT significantly underperforms tPIT when an optimal speaker tracker is used for tPIT (Kolbæk *et al.*, 2017).

Following deep clustering and PIT, other talker-independent separation systems are proposed with impressive performance in anechoic conditions. For example, Conv-TasNet (Luo and Mesgarani, 2019) and FurcaNeXt (Shi *et al.*, 2019) report more than 15 dB signal-to-distortion ratio improvement (ΔSDR; Vincent *et al.*, 2006), even surpassing the oracle results of the ideal ratio mask (IRM; Wang *et al.*, 2014). Another state-of-the-art approach is deep computational auditory scene analysis (CASA; Liu and Wang, 2019), in which a two-talker mixture is first separated at the frame level, and then the separated frames are grouped sequentially into individual speakers. This algorithm is inspired by CASA (Wang and Brown, 2006), which performs a simultaneous grouping of *T-F* segments overlapping in time, followed by sequential organization that groups simultaneous streams to sound sources.

To our knowledge, talker-independent speaker separation has been limited to anechoic conditions, except for Wang and Wang (2018) and Wang *et al.* (2018), in which a single-channel scenario is evaluated as a baseline for multi-channel talker-independent speaker separation. In this paper, we address the monaural talker-independent speaker separation in reverberant conditions. To this end, we adopt deep CASA as a baseline and extend it to a two-stage method. In the proposed method, the first stage separates reverberant speech signals, and in the second stage, the separated utterances are further dereverberated. We find it advantageous to address speaker separation and speech dereverberation in two different stages. The idea of two-stage processing has been successfully used in other studies (Grais *et al.*, 2017; Wang *et al.*, 2017; Zhao *et al.*, 2019), including our study on talker-dependent reverberant speaker separation (Delfarah and Wang, 2019).

A preliminary version of this work is presented in Delfarah *et al.* (2020). Compared to the conference version, this paper conducts a much wider range of evaluations. We also discuss how the simultaneous grouping and sequential organization modules each affect the overall system performance. In addition, this paper includes a talker-dependent system as a strong baseline and investigates cross-corpus generalization. The rest of this paper is organized as follows. Section II describes the background of this study including the problem formulation and the deep CASA baseline. In Sec. III, we present the proposed two-stage deep CASA algorithm. Evaluation results and comparisons are presented in Sec. IV. Section V concludes the paper.

II. BACKGROUND

In a reverberant room, a source signal travels in different directions and bounces off of the surrounding surfaces. The reflected signals can be viewed as attenuated and time-delayed copies of the original signal, and these reflected signals combine at the microphone location. As a result, the reverberant signal appears temporally and spectrally smeared compared to the source signal. A room impulse response (RIR) characterizes the reverberation and converts an anechoic source signal to the reverberant received signal. In a reverberant room with two simultaneous talkers, the two-speaker mixture $y(t)$ can be described monaurally as:

$$y(t) = s_{r_1}(t) + s_{r_2}(t) = s_1(t) * h_1(t) + s_2(t) * h_2(t), \tag{1}$$

where s_1 and s_2 are the anechoic speech signals, h_1 and h_2 are their corresponding RIRs, s_{r_1} and s_{r_2} are the reverberant speech signals, and "$*$" denotes the convolution operator. In this study, we aim at extracting s_1 and s_2 from y. In other words, the end goal is to obtain separated and dereverberated speech signals. An alternative goal would be separating reverberant signals s_{r_1} and s_{r_2}. The current study extracts s_1 and s_2 as hearing-impaired listeners show higher speech intelligibility when presented with separated and dereverberated signals compared to separated reverberant signals (Healy *et al.*, 2019).

The speaker separation problem formulated in Eq. (1) has a special case of speaker separation in anechoic conditions, i.e.,

$$h_1(t) = h_2(t) = \delta(t), \tag{2}$$

where δ is the Dirac delta function. The deep CASA approach (Liu and Wang, 2019) addresses anechoic talker-independent speaker separation by first separating the speakers at the frame level and then predicting the correct speaker-frame assignments to sequentially organize the separated frames over the entire mixture to generate separated speaker signals. In deep CASA, these two modules correspond to simultaneous grouping and sequential organization in CASA (Wang and Brown, 2006). The simultaneous grouping module is a Dense-Unet (Huang *et al.*, 2017; Ronneberger *et al.*, 2015), which is trained using a tPIT criterion (Yu *et al.*, 2017). The sequential grouping module utilizes a temporal convolutional network (TCN; Bai *et al.*, 2018; Lea *et al.*, 2016) to predict embeddings followed by a k-means algorithm to cluster the embeddings to produce frame-speaker assignments.

A straightforward way to extend deep CASA to reverberant conditions is as follows. The simultaneous grouping module predicts anechoic speech signals at the frame level, and the sequential grouping module organizes the separated frames to yield s_1 and s_2 estimates. Since this baseline method produces anechoic signals at the output of simultaneous grouping, it performs speaker separation and speech dereverberation in one stage.

We note that speaker separation and speech dereverberation are intrinsically different tasks, and using a single stage to accomplish both tasks may require a mapping from mixture features to a training target that is too complicated

https://doi.org/10.1121/10.0001779

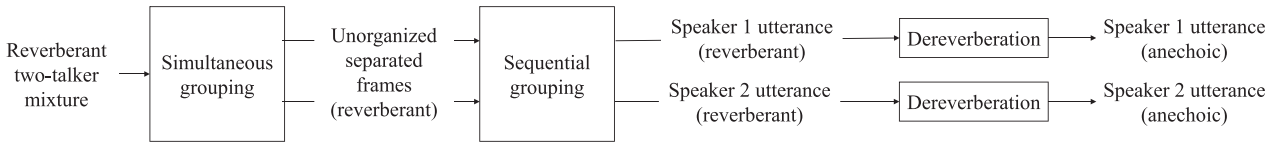


FIG. 1. Overview of the proposed two-stage deep CASA algorithm for speaker separation in reverberant conditions.

for a single DNN to learn effectively. In Sec. III, we propose a two-stage deep CASA algorithm that performs separation and dereverberation separately, reminiscent of the divide and conquer strategy of CASA.

III. TWO-STAGE DEEP CASA FOR SPEAKER SEPARATION IN REVERBERANT CONDITIONS

The algorithm overview is shown in Fig. 1. Given a reverberant two-talker mixture, a simultaneous grouping module is trained to perform frame-level separation of the reverberant speech signals. Then, the sequential grouping module learns to organize those separated reverberant frames over time to form two reverberant streams, each corresponding to one speaker. The simultaneous and sequential grouping modules form the proposed first stage. In the proposed second stage, two separated streams are individually fed into a dereverberation module, which is trained to yield the anechoic source signals. Our system builds upon the deep CASA framework (Liu and Wang, 2019) and is described as follows.

A. Simultaneous grouping

An overview of the simultaneous grouping module is depicted in Fig. 2. Given the mixture signal y, the real and imaginary short-time Fourier transform (STFT) features, Y, are extracted and fed into a DNN, which produces two complex ratio masks cRM_1 and cRM_2 (Williamson *et al.*, 2016). These ratio masks, along with Y, produce two STFT features $\hat{S}_{u_1}$ and $\hat{S}_{u_2}$:

$$\hat{S}_{u_1}(m,f) = \text{cRM}_1(m,f) \otimes Y(m,f), \tag{3}$$

$$\hat{S}_{u_2}(m,f) = \text{cRM}_2(m,f) \otimes Y(m,f), \tag{4}$$

where m is the time frame index, f is the frequency index, and $\otimes$ denotes pointwise matrix multiplication. $\hat{S}_{u_1}$ and $\hat{S}_{u_2}$ are the estimated frame-level reverberant speaker signals, yet, to be organized over time. Using the tPIT loss function (Yu *et al.*, 2017), these frames are reorganized into $\hat{S}_{o_1}$ and $\hat{S}_{o_2}$, which correctly represent the speaker signals. Accordingly, two possible loss functions ℓ_1 and ℓ_2 are calculated per time frame as

$$\ell_1(m) = \Sigma_f |\hat{S}_{u_1}(m,f) - S_{r_1}(m,f)| + \Sigma_f |\hat{S}_{u_2}(m,f) - S_{r_2}(m,f)|, \tag{5}$$

$$\ell_2(m) = \Sigma_f |\hat{S}_{u_1}(m,f) - S_{r_2}(m,f)| + \Sigma_f |\hat{S}_{u_2}(m,f) - S_{r_1}(m,f)|, \tag{6}$$

where S_{r_1} and S_{r_2} are the complex STFT features s_{r_1} and s_{r_1}, accordingly. Then, optimal speaker tracking is performed as follows:

$$\hat{S}_{o_1}(m,f), \hat{S}_{o_2}(m,f) = \begin{cases} \hat{S}_{u_1}(m,f), \hat{S}_{u_2}(m,f), & \text{if } \ell_1(m) \le \ell_2(m), \\ \hat{S}_{u_2}(m,f), \hat{S}_{u_1}(m,f), & \text{otherwise.} \end{cases} \tag{7}$$

Next, inverse STFT converts $\hat{S}_{o_1}$ and $\hat{S}_{o_2}$ into time domain signals $\hat{s}_{o_1}$ and $\hat{s}_{o_2}$, which are the predicted reverberant signals using optimal speaker tracking. Finally, the module is

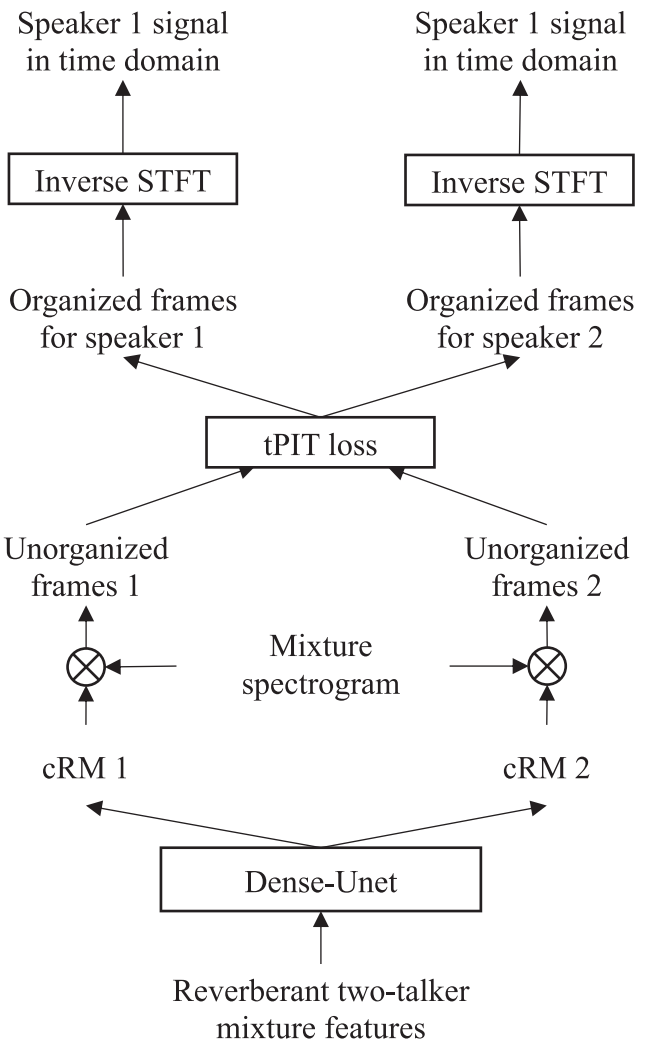


FIG. 2. Overview of the simultaneous grouping module in the proposed two-stage deep CASA algorithm, which separates speakers at the frame level and organizes the time frames with an optimal speaker tracker. cRM refers to a complex ratio mask.

optimized to minimize the signal-to-noise ratio (SNR) loss function $J^{\text{tPIT}-\text{SNR}}$,

$$J^{\text{tPIT}-\text{SNR}} = -10 \sum_{i=1,2} \log \frac{\Sigma_t s_i(t)^2}{\Sigma_t \left[s_i(t) - \hat{s}_{o_i}(t) \right]^2}. \tag{8}$$

A Dense-Unet architecture was used for simultaneous grouping in deep CASA (Liu and Wang, 2019), which consists of a series of upsampling layers, downsampling layers, and dense convolutional blocks. The network can be divided into two halves. In the first half, dense convolutional blocks and downsampling layers are alternated to encode the input *T-F* feature map into a higher level of abstraction. Each dense convolutional block is composed of five densely connected convolutional layers with each layer connected to every other layer in a feedforward fashion. The middle layer in a dense block is a special frequency-mapping layer, which alleviates the inconsistency between different frequencies by reorganizing them in a new space. In the second half of the network, dense blocks and upsampling layers are alternated to project the encoded information back to the original *T-F* resolution. Skip connections are added to link the dense blocks at the same hierarchical level in the two halves so that the fine-grain details in the mixture are preserved.

B. Sequential grouping

As seen in Eqs. (5)–(8), the training of the simultaneous grouping module uses premixed reverberant mixtures s_{r_1} and s_{r_2} to optimize the neural network with the optimal frame-speaker assignments. These signals are not available at test time, and for this reason, a sequential grouping module is trained to predict a temporal organization.

Figure 3 illustrates the sequential organization module. For an M-frame mixture the training target $\boldsymbol{A}$ is an $M \times 2$ matrix,

$$\boldsymbol{A}(m) = \begin{cases} [1,0], & \text{if } \ell_1(m) \le \ell_2(m), \\ [0,1], & \text{otherwise}, \end{cases} \tag{9}$$

where $1 \le m \le M$. One can see that $\boldsymbol{A}$ optimally organizes $\hat{S}_{u_1}$ and $\hat{S}_{u_2}$ as follows:

$$\begin{bmatrix} \hat{S}_{o_1} \\ \hat{S}_{o_2} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A} \\ 1 - \boldsymbol{A} \end{bmatrix} \begin{bmatrix} \hat{S}_{u_1} \\ \hat{S}_{u_2} \end{bmatrix}. \tag{10}$$

As input features, this module uses the magnitude spectrograms $|Y|$, along with the unorganized signals $|\hat{S}_{u_1}|$ and $|\hat{S}_{u_2}|$, produced from Eqs. (3) and (4). The network predicts the embedding matrix $\boldsymbol{V} \in \mathbb{R}^{M \times D}$, where D is the embedding dimension. $\boldsymbol{V}$ encodes the information of the optimal output-speaker pairing. Finally, we optimize the loss function J^{DC} (Hershey *et al.*, 2016),

$$J^{DC} = \|\boldsymbol{W}^{1/2}(\boldsymbol{V}\boldsymbol{V}^T - \boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{W}^{1/2}\|_F^2, \tag{11}$$

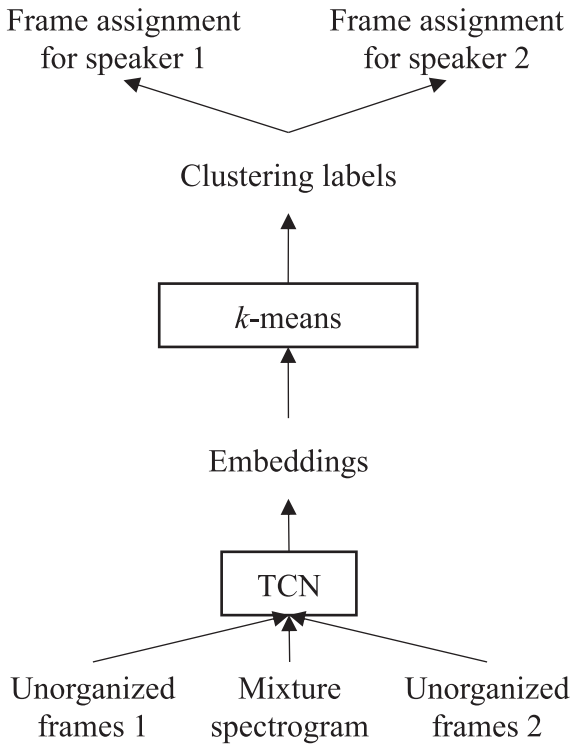


FIG. 3. Overview of the sequential organization module, which assigns the simultaneously separated frames to individual speakers using k-means clustering.

where $\|\cdot\|_F$ is the Frobenius norm and

$$\boldsymbol{W} = \text{diag}\left(\frac{|l_1 - l_2|}{\Sigma_c |l_1 - l_2|} \right), \tag{12}$$

in which l_1 and l_2 are obtained from Eqs. (5) and (6).

At test time, a sequential grouping predicts the embedding $\boldsymbol{V}$ and then a k-means algorithm clusters $\boldsymbol{V}(m)$ vectors, i.e., assigning labels $\hat{a}(m) = \{0, 1\}$. Then, the predicted sequential organization matrix $\hat{\boldsymbol{A}}$ is obtained:

$$\hat{\boldsymbol{A}} = [\hat{\boldsymbol{a}}, 1 - \hat{\boldsymbol{a}}], \tag{13}$$

which is used as in Eq. (9) to generate the estimated complex domain reverberant signals $\hat{S}_{r_1}$ and $\hat{S}_{r_2}$.

A TCN (Bai *et al.*, 2018; Lea *et al.*, 2016) is adopted as the sequence model in this module. In the TCN, input features are first fed to eight consecutive dilated convolutional blocks with an exponentially increasing dilation factor F ($F = 2^0, 2^1, \ldots, 2^7$). Each dilated convolutional block is composed of three convolutional layers. The first layer extends the number of feature maps in the input. The second depthwise dilated layer links the current frame with feature maps, whicch are D frames away, greatly expanding the temporal context. The last layer projects the feature maps back to the input dimension and combines it with the input to form the output of the block. To further expand the receptive field, the eight blocks are repeated three times before embedding the estimation. During the training of the TCN, a dropDilation (Liu and Wang, 2019) technique is utilized to overcome the overfitting problem.

https://doi.org/10.1121/10.0001779

C. Speech dereverberation

In this stage, the estimated reverberant signals are individually dereverberated using a DNN with the same architecture as the Dense-Unet used in the simultaneous grouping module except that the tPIT loss function is not used. $\hat{S}_{r_1}$ (or $\hat{S}_{r_2}$) is fed to the network, and a complex ratio mask is generated to produce the complex domain STFT $\hat{S}_1$ (or $\hat{S}_2$). This is then converted to the time-domain signal $\hat{s}_1$ (or $\hat{s}_2$) and the network is optimized as follows:

$$J^{\mathrm{SNR}} = -10 \log \frac{\Sigma_t s_i(t)^2}{\Sigma_t [s_i(t) - \hat{s}_i(t)]^2}, \quad i = 1, 2. \tag{14}$$

At the inference time, the reverberant two-talker mixture is passed through the simultaneous grouping module to form the frame-level separated reverberant speaker signals. Then, the sequential grouping module uses the k-means algorithm to predict an across-frame organization sequence. This predicted sequence is used to form the two reverberant speech streams. Finally, each of the streams undergoes speech dereverberation to yield the predicted anechoic speaker signals.

IV. EVALUATIONS AND COMPARISONS

A. Experimental setup

The utterances used in this study are sampled at 16 kHz. Spectrograms are generated by dividing the waveform signals into 32-ms frames with a frame shift of 8 ms and applying a 512-point discrete Fourier transform (DFT).

We use the WSJ0 dataset (Garofolo *et al.*, 1993) to generate the two-talker mixtures.[1] To produce the training data, two different speakers from the folder **si_tr_s** are randomly picked each time to produce a pair. First, these anechoic signals are equalized to the same root-mean-square (RMS) level. In the case of speaker pairs with utterances of different lengths, the longer utterance is trimmed to the length of the shorter one. We use the image method to simulate reverberant room conditions. To this end, for each speaker pair, a simulated room with random dimensions of $L \in [6, 7]$ m, $W \in [8, 9]$ m, and $H \in [2.5, 3.5]$ m is picked. Next, a $T_{60} \in [0.3, 1.0]$ s is chosen for the reverberation time. A microphone is placed in this room at a random position of $x \in [2.5, 3.5]$ m, $y \in [3.5, 4.5]$ m, and $z \in [1, 2]$ m, with x, y, and z corresponding to the length, width, and height, respectively. Then, each of the two speakers is placed in the room at different locations with a random distance of $d \in [0.5, 3]$ m from the microphone. The speakers and the microphone have the same elevation within the room. A RIR generator toolbox[2] is used to generate a RIR for each speaker. Then, the reverberant two-talker mixtures are generated as in Eq. (1). The direct-path signals are used as the training targets. These signals can be obtained by convolving the anechoic source signal with a RIR that is generated by retaining only the largest peak in the original RIR. The process of mixture generation is repeated to obtain 201 000 training mixture signals, from which 1000 mixtures are set aside for cross-validation.

One set of experiments involves using speakers from the WSJ0 corpus that were not used during training. For this purpose, speakers are picked from the folders **si_dt_05** and **si_et_05**, and mixture signals are generated using simulated and recorded RIRs. Three simulated rooms are obtained with $T_{60} = 0.3$, 0.6, and 0.9 s. Test room dimensions are $(6.5, 8.5, 3)$ m, and the microphone is fixed at the position of $[3, 4, 1.5]$ m. The same speaker-microphone distances are used as in the training, and all test mixtures have 0 dB target-to-interferer ratio (TIR). To generate mixtures with real room conditions, recorded RIRs from four rooms are chosen with $T_{60} = 0.32$, 0.47, 0.68, and 0.89 s. These RIRs are adopted from Hummersone *et al.* (2010), where detailed room configurations can be found. In each condition, 3000 test mixtures are generated and average results are reported. Our test set comprises 1603 male-female, 530 female-female, and 867 male-male mixtures.

We perform cross-corpus evaluations without retraining using 120 male and 120 female utterances from two Institute of Electrical and Electronics Engineers (IEEE) speakers (IEEE, 1969). For this purpose, 1000 male-female mixtures were generated by randomly selecting and mixing two of those utterances at a time. Furthermore, we evaluate using the LibriSpeech corpus (Panayotov *et al.*, 2015), which is extracted from audiobooks read and recorded by many volunteer speakers. As a result, some level of reverberation would exist in speech signals. We generate 1000 mixtures by randomly selecting utterances from the `test-clean` set of LibriSpeech. In our IEEE and LibriSpeech evaluations, room configurations were the same as those used in the WSJ0 test set, and two speakers were mixed at 0 dB TIR. Note that the sentences and speakers are different in all three corpora.

The networks in all modules were trained using the Adam optimizer (Kingma and Ba, 2015). First, the simultaneous grouping module was trained and then the outputs of this module were used to train the sequential grouping module. The dereverberation network was first trained to dereverberate single-speaker signals and then was connected to the outputs of the simultaneous grouping module and further trained. This step is done to fine-tune the dereverberation module and, therefore, the parameters in the speaker separation stage are not updated. This implies that speaker separation and speech dereverberation are trained successively. We observed slight performance degradation with joint training of the two stages, probably because of the different natures of dereverberation and speaker separation.

Throughout algorithm training, the learning rates were adjusted based on the loss on the validation set, and the DNNs with the smallest errors in the validation set were used to train the consecutive modules and at inference time. Standard performance metrics used for algorithm evaluations are ESTOI (Jensen and Taal, 2016), PESQ (Rix *et al.*, 2001), and ΔSDR (dB) (Vincent *et al.*, 2006). ESTOI measures speech intelligibility and produces a number typically

in the range of [0, 1]. It is an extended version of STOI (Taal *et al.*, 2011) and produces higher correlation with human intelligibility than STOI for modulated noises. Here, we report ESTOI in percent. PESQ is a number in the range of [−0.5 4.5] and measures speech quality. In all three metrics, higher scores indicate better results. The reference signals in the evaluations are the direct-path speech signals, and in each test condition, the average scores of the test mixtures are reported.

B. Comparison systems

We compare the performance of the proposed two-stage deep CASA algorithm with several strong baselines.

1. One-stage deep CASA

This system (Liu and Wang, 2019) represents the direct extension of deep CASA from anechoic to reverberant conditions and is implemented as described in Sec. II.

2. One-stage uPIT Dense-Unet

The Dense-Unet structure of the simultaneous grouping module in deep CASA is trained with the uPIT loss function. This system pulls together the separated frames for each speaker throughout a reverberant mixture and, therefore, does not need sequential organization.

3. Two-stage uPIT Dense-Unet

To further evaluate the proposed two-stage strategy, we also create a two-stage baseline for uPIT Dense-Unet. The first stage uses uPIT Dense-Unet to predict reverberant complex STFT features for each speaker in the mixture. The outputs of this stage are fed to a dereverberation stage to predict direct-path signals. This baseline is similar to two-stage deep CASA except that uPIT Dense-Unet is used as the first stage.

4. One-stage fully convolutional time-domain audio separation network (one-stage Conv-TasNet)

Conv-TasNet (Luo and Mesgarani, 2019) performs uPIT (Kolbæk *et al.*, 2017) in the time domain and demonstrates very good results separating speakers in anechoic conditions. Our implementation of Conv-TasNet uses a TCN structure and predicts frames of length 2 ms with a 1 ms frame shift. It minimizes an utterance-level SNR loss function [Eq. (8)] to predict two direct-path speaker signals from a reverberant mixture in the time domain.

5. Two-stage Conv-TasNet

Similarly, we create a two-stage Conv-TasNet baseline. Conv-TasNet is used in the first stage to predict reverberant speakers, each of which is converted to the complex spectral domain. In the second stage, each signal undergoes dereverberation as done in two-stage deep CASA.

6. IRM

We report the results for signals reconstructed by the IRM defined as (Wang *et al.*, 2014)

$$\mathrm{IRM}_i = \sqrt{\frac{S_i^2(m,f)}{S_{r_1}^2(m,f)+S_{r_2}^2(m,f)}}, \quad i = 1, 2. \tag{15}$$

Note that this system is not trained and the IRM is directly calculated from premixed signals. The overlap-add method uses $\mathrm{IRM}_i \otimes |Y|$ and $\angle Y$ to resynthesize time-domain signals.

TABLE I. ΔSDR, ESTOI (%), and PESQ scores for speaker separation in simulated reverberant conditions using WSJ0 test mixtures. Boldface indicates the best separation score in each condition, and italic indicates the oracle results with the IRM.

Metrics		ΔSDR (dB)				ESTOI (%)				PESQ			
T_{60} (s)	Sequential grouping	0.3	0.6	0.9	Avgerage	0.3	0.6	0.9	Avgerage	0.3	0.6	0.9	Avgerage
Unprocessed	—	0.0	0.0	0.0	0.0	43.2	31.1	23.0	32.5	1.70	1.48	1.37	1.52
IRM	—	*10.0*	*8.8*	*8.6*	*9.1*	*81.3*	*76.5*	*73.1*	*76.9*	*3.10*	*2.85*	*2.68*	*2.88*
One-stage uPIT Dense-Unet	Optimal	7.8	7.6	7.5	7.6	73.6	64.2	55.2	64.3	2.62	2.33	2.06	2.34
	Default	7.6	7.4	7.3	7.4	72.6	63.0	54.0	63.2	2.60	2.31	2.07	2.32
Two-stage uPIT Dense-Unet	Default	9.2	8.3	8.3	8.6	80.7	70.9	63.2	71.6	2.84	2.47	**2.24**	2.52
One-stage Conv-TasNet	Default	7.9	7.4	7.0	7.4	73.2	62.3	52.4	62.6	2.59	2.28	2.02	2.30
Two-stage Conv-TasNet	Default	8.7	8.1	8.1	8.3	78.7	68.5	60.6	69.3	2.69	2.34	2.11	2.38
One-stage deep CASA	Optimal	10.0	9.3	8.9	9.4	80.5	71.6	63.0	71.7	2.90	2.57	2.30	2.59
	Learned	**9.7**	**9.0**	**8.6**	**9.1**	79.6	70.3	61.2	70.4	2.84	2.49	2.21	2.51
Two-stage deep CASA	Optimal	10.1	9.4	9.1	9.6	83.0	75.0	66.9	75.0	3.03	2.65	2.35	2.68
	Learned	9.5	8.8	8.4	8.9	**81.2**	**72.5**	**63.5**	**72.4**	**2.91**	**2.52**	2.18	**2.54**

https://doi.org/10.1121/10.0001779

TABLE II. ΔSDR, ESTOI (%), and PESQ scores for speaker separation in recorded reverberant conditions using WSJ0 test mixtures.

Metrics		ΔSDR (dB)					ESTOI (%)					PESQ				
T_{60} (s)	Sequential grouping	0.32	0.47	0.68	0.89	Avgerage	0.32	0.47	0.68	0.89	Avgerage	0.32	0.47	0.68	0.89	Avgerage
Unprocessed	—	0.0	0.0	0.0	0.0	0.0	41.9	37.9	44.1	33.5	39.4	1.68	1.64	1.76	1.53	1.65
IRM	—	*10.7*	*10.4*	*10.6*	*10.2*	*10.5*	*82.4*	*81.0*	*84.6*	*81.4*	*82.4*	*3.25*	*3.22*	*3.26*	*3.15*	*3.22*
One-stage uPIT Dense-Unet	Optimal	8.7	8.8	9.5	9.7	9.2	72.2	68.3	73.7	65.9	70.0	2.61	2.53	2.66	2.40	2.55
	Default	7.6	7.6	8.5	8.4	8.0	69.9	65.3	71.0	62.9	67.3	2.49	2.39	2.54	2.27	2.42
Two-stage uPIT Dense-Unet	Default	8.8	8.6	9.9	9.4	9.2	76.6	72.3	77.6	69.8	74.1	2.67	2.56	2.71	2.40	2.58
One-stage Conv-TasNet	Default	6.3	6.5	7.4	7.7	7.0	64.8	61.2	66.6	58.6	62.8	2.28	2.23	2.39	2.13	2.26
Two-stage Conv-TasNet	Default	7.7	7.5	8.5	8.5	8.1	70.8	66.5	72.2	65.8	68.8	2.47	2.38	2.53	2.25	2.41
One-stage deep CASA	Optimal	10.4	10.4	11.2	11.1	10.8	76.8	72.9	77.6	70.9	74.6	2.77	2.71	2.80	2.55	2.71
	Learned	9.4	9.4	10.2	10.1	9.8	74.7	70.5	75.3	68.2	72.2	2.62	2.55	2.67	2.40	2.56
Two-stage deep CASA	Optimal	11.1	11.3	12.1	11.8	11.6	79.5	76.9	81.4	75.7	78.4	2.89	2.82	2.95	2.63	2.82
	Learned	**10.0**	**10.2**	**11.1**	**10.6**	**10.5**	**76.8**	**74.0**	**78.7**	**71.9**	**75.3**	**2.71**	**2.64**	**2.80**	**2.43**	**2.64**

7. Talker-dependent speaker separation

A talker-dependent system is trained to separate a known female IEEE speaker and a known male IEEE speaker. To train this system, 100 000 IEEE mixtures are generated using 600 male and 600 female sentences that were not used in testing, and the experimental setup described in Sec. IV A is followed. This system uses a Dense-Unet similar to the simultaneous grouping module of deep CASA without a tPIT criterion. Since the talker-dependent system is specifically designed to separate the two IEEE speakers, it is excluded from the experiments with the WSJ0 corpus.

C. Experimental results using WSJ0 mixtures

In this section, we evaluate and compare the systems using the WSJ0 test speakers. Table I shows ΔSDR, ESTOI (%), and PESQ scores in simulated reverberant room conditions. Considering how frames are sequentially organized, two sets of results are presented for the deep CASA and uPIT Dense-Unet systems. Optimal sequential grouping refers to the ideal organization of Eq. (10), whereas learned sequential organization uses the output of the sequential grouping module [Eq. (13)]. For systems that organize and separate the speakers in a single module at the utterance level, default sequential organization is used in which $\hat{\boldsymbol{A}}(m) = [0, 1], \forall m \in M$.

The results in Table I show that one-stage deep CASA produces slightly higher ΔSDR scores than two-stage deep CASA, whereas the two-stage deep CASA algorithm yields slightly higher ESTOI and PESQ scores. The IRM has significantly better ESTOI and PESQ results than the deep CASA systems, but on ΔSDR, the deep CASA systems have comparable performance to the IRM. From Table I, we also

TABLE III. ΔSDR, ESTOI (%), and PESQ scores for same- and different-gender pairs using WSJ0 test mixtures. Average results in recorded and simulated room conditions are presented. Same genders include both male-male and female-female mixtures and different genders include male-female mixtures.

Metrics		ΔSDR (dB)		ESTOI (%)		PESQ	
mixture genders	Sequential grouping	Same	Different	Same	Different	Same	Different
Unprocessed	—	0.0	0.0	36.1	35.8	1.58	1.55
IRM	—	*9.8*	*9.8*	*79.7*	*79.7*	*3.04*	*3.03*
One-stage uPIT Dense-Unet	Optimal	8.1	9.2	65.7	70.1	2.39	2.55
	Default	6.6	8.9	61.7	69.5	2.21	2.53
Two-stage uPIT Dense-Unet	Default	8.6	10.1	70.5	75.7	2.40	2.68
One-stage Conv-TasNet	Default	6.7	7.7	60.8	64.7	2.20	2.34
Two-stage Conv-TasNet	Default	8.3	9.4	70.7	75.0	2.28	2.55
One-stage deep CASA	Optimal	10.0	10.2	73.0	73.5	2.64	2.65
	Learned	**8.8**	10.0	69.6	72.9	2.44	2.61
Two-stage deep CASA	Optimal	10.5	10.6	76.4	77.1	2.74	2.76
	Learned	**8.8**	**10.4**	**71.1**	**76.3**	**2.45**	**2.71**

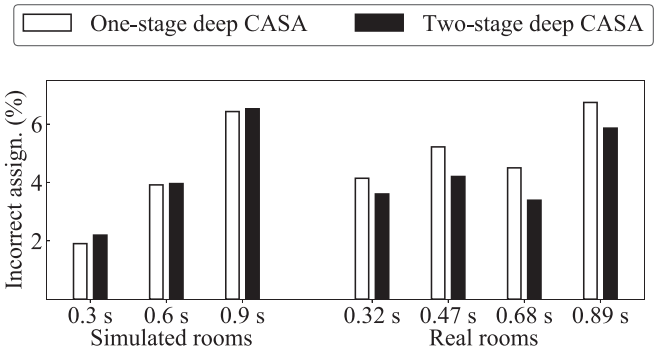


FIG. 4. Percentage of incorrect frame-speaker assignments in deep CASA systems for separating WSJ0 test mixtures. The horizontal axis represents T_{60} values.

observe that for uPIT Dense-Unet and Conv-TasNet, the two-stage versions have uniformly better scores than the one-stage versions. Overall, deep CASA works better than Conv-TasNet and uPIT Dense-Unet. This observation is consistent with speaker separation results in anechoic conditions (Liu and Wang, 2019).

Table II evaluates and compares the systems under the conditions of real RIRs. Note that the training set only contains simulated RIRs. Two-stage systems perform uniformly better than their one-stage counterparts. In particular, two-stage deep CASA has the best performance across all recorded RIR conditions and significantly outperforms the one-stage deep CASA. The average difference in performance is 0.7 dB signal-to-distortion ratio (SDR), 3.1% ESTOI, and 0.08 PESQ points. These results suggest that addressing speech dereverberation and speaker separation in two different stages results in better generalization to unseen, real room conditions. In this case, two-stage deep CASA is able to match the SDR performance of the IRM; however, it underperforms the IRM by 7.1% ESTOI and 0.58 PESQ.

Table III breaks the ΔSDR, ESTOI, and PESQ scores for each system into same and opposite gender conditions. The results indicate that when used with optimal sequential organization, the deep CASA systems can separate the same gender mixtures as well as the different gender mixtures. On the other hand, in one-stage and two-stage deep CASA with learned sequential organization, different gender mixtures yield significantly better results, which is not surprising.

To examine the quality of sequential organization in the deep CASA systems, Fig. 4 plots the rate of the incorrect frame-speaker assignments for these systems in simulated and recorded room conditions. To measure these error rates, we count the total number of different elements between vectors $\boldsymbol{A}$ from Eq. (10) and $\hat{\boldsymbol{A}}$ from Eq. (13), and then divide this number by the total number of frames in the utterance. To eliminate the effects of silent and low-energy frames, this examination considers only frames that have an energy ratio of at least −20 dB relative to the frame with the maximum energy in the mixture. The results indicate that in simulated reverberant conditions, the sequential grouping modules in one-stage and two-stage deep CASA have comparable performances, whereas in recorded room conditions, the two-stage system produces fewer frame organization errors. We think that this is because the output frames from the simultaneous grouping module in two-stage deep CASA are better separated (and, hence, less similar), thus, facilitating sequential organization.

To summarize, the deep CASA algorithms perform much better than Conv-TasNet and uPIT Dense-Unet in simulated and real reverberant conditions. In recorded reverberant conditions, the simultaneous grouping module of two-stage deep CASA appears to outperform this module in one-stage deep CASA, leading to fewer frame-speaker assignment errors. The results show better generalization of the two-stage deep CASA algorithm to real acoustic conditions than were shown for the one-stage CASA algorithm.

TABLE IV. ΔSDR, ESTOI (%), and PESQ scores for speaker separation in simulated reverberant conditions using the IEEE male-female mixtures.

Metrics		ΔSDR (dB)				ESTOI (%)				PESQ			
T_{60} (s)	Sequential grouping	0.3	0.6	0.9	Avgerage	0.3	0.6	0.9	Avgerage	0.3	0.6	0.9	Avgerage
Unprocessed	—	0.0	0.0	0.0	0.0	39.9	29.3	22.2	30.5	1.51	1.30	1.19	1.33
IRM	—	*10.4*	*9.1*	*8.8*	*9.4*	*81.9*	*77.4*	*74.4*	*77.9*	*2.98*	*2.68*	*2.50*	*2.72*
Talker-dependent	Default	11.0	10.1	9.7	10.2	82.1	74.5	66.9	74.5	3.02	2.71	2.45	2.72
One-stage uPIT Dense-Unet	Optimal	9.3	8.7	8.4	8.8	75.9	66.7	57.6	66.7	2.76	2.43	2.14	2.44
	Default	9.2	8.6	8.3	8.7	75.6	66.4	57.3	66.4	2.74	2.42	2.13	2.43
Two-stage uPIT Dense-Unet	Default	9.7	8.7	8.1	8.9	78.4	69.0	58.9	68.8	2.84	2.49	2.18	2.50
One-stage Conv-TasNet	Default	6.0	5.6	5.5	5.7	65.6	53.7	44.3	54.6	2.34	2.02	1.77	2.05
Two-stage Conv-TasNet	Default	8.4	7.7	7.2	7.8	72.8	62.9	52.2	62.6	2.57	2.23	1.92	2.24
One-stage deep CASA	Optimal	10.4	9.6	9.0	9.7	79.4	70.7	62.1	70.8	2.89	2.55	2.27	2.57
	Learned	10.3	9.5	8.9	9.6	79.0	70.2	61.3	70.2	2.86	2.53	2.24	2.54
Two-stage deep CASA	Optimal	10.7	9.8	9.3	9.9	81.9	73.8	65.7	73.8	3.02	2.65	2.34	2.67
	Learned	**10.6**	**9.7**	**9.1**	**9.8**	**81.3**	**73.1**	**64.5**	**72.9**	**2.98**	**2.61**	**2.29**	**2.62**

https://doi.org/10.1121/10.0001779

TABLE V. ΔSDR, ESTOI (%), and PESQ scores for speaker separation in recorded reverberant conditions using the IEEE male-female mixtures.

Metrics		ΔSDR (dB)					ESTOI (%)					PESQ				
T_{60} (s)	Sequential grouping	0.32	0.47	0.68	0.89	Avgerage	0.32	0.47	0.68	0.89	Avgerage	0.32	0.47	0.68	0.89	Avgerage
Unprocessed	—	0.0	0.0	0.0	0.0	0.0	40.1	36.5	42.3	32.3	37.8	1.5	1.46	1.57	1.34	1.47
IRM	—	*10.9*	*10.4*	*10.6*	*10.1*	*10.5*	*83.5*	*81.9*	*85.8*	*82.2*	*83.3*	*3.12*	*3.07*	*3.14*	*2.96*	*3.07*
Talker-dependent	Default	11.2	11.2	12.2	11.9	11.6	81.1	78.0	82.0	75.6	79.2	2.94	2.87	2.98	2.71	2.87
One-stage uPIT Dense-Unet	Optimal	9.9	10.0	10.6	10.7	10.3	75.5	71.6	76.3	68.2	72.9	2.72	2.64	2.75	2.46	2.64
	Default	9.8	9.8	10.5	10.5	10.2	75.4	71.4	76.2	67.9	72.7	2.71	2.62	2.74	2.44	2.63
Two-stage uPIT Dense-Unet	Default	9.1	9.5	9.8	10.0	9.6	76.3	73.1	76.2	69.8	73.8	2.76	2.70	2.75	2.50	2.68
One-stage Conv-TasNet	Default	5.5	5.2	6.4	6.5	5.9	63.2	57.5	64.5	54.4	59.9	2.22	2.06	2.25	1.96	2.12
Two-stage Conv-TasNet	Default	8.2	8.5	9.0	9.4	8.8	70.9	67.7	71.0	64.4	68.5	2.53	2.49	2.57	2.30	2.47
One-stage deep CASA	Optimal	10.9	10.8	11.4	11.5	11.1	77.1	73.4	77.6	71.1	74.8	2.78	2.69	2.80	2.53	2.70
	Learned	10.8	10.7	11.3	11.3	11.0	76.9	73.0	77.1	70.5	74.4	2.75	2.66	2.77	2.50	2.67
Two-stage deep CASA	Optimal	11.6	11.6	12.2	12.2	11.9	79.7	76.7	80.8	75.4	78.2	2.92	2.83	2.96	2.66	2.84
	Learned	**11.5**	**11.4**	**12.1**	**12.0**	**11.8**	**79.4**	**76.1**	**80.3**	**74.5**	**77.6**	**2.89**	**2.79**	**2.93**	**2.60**	**2.80**

D. Cross-corpus generalization

To further investigate the generalization of reverberant speaker separation, we present evaluation results using the IEEE corpus. The objective scores in simulated reverberant conditions are shown in Table IV. Comparing the results in Table IV with the results in Table I indicates that the proposed two-stage deep CASA algorithm generalizes very well to an unseen speech corpus. Our two-stage algorithm performs better than the one-stage deep CASA system as expected (see Delfarah and Wang, 2019). Comparing the results of one-stage and two-stage uPIT Dense-Unet on WSJ0 and IEEE corpora indicates that these systems also generalize well to the unseen IEEE corpus. A large performance drop is observed for one-stage Conv-TasNet when tested with the IEEE speakers; however, the two-stage version performs substantially better. Table IV also presents the oracle results using the IRM, and the two deep CASA and talker-dependent systems achieve higher SDR scores than the IRM. The talker-dependent system matches the PESQ performance of the IRM. This shows the strong representational capacity of the Dense-Unet structure and complex ratio masking used in this study.

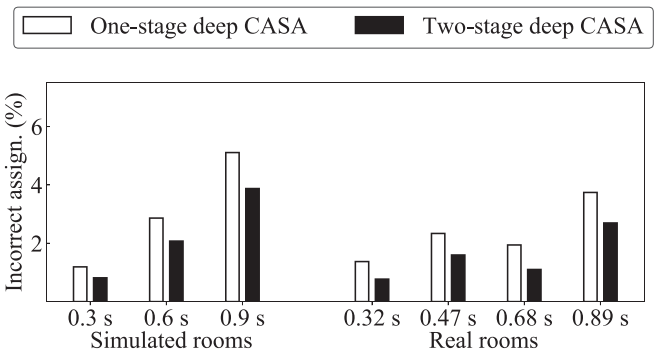


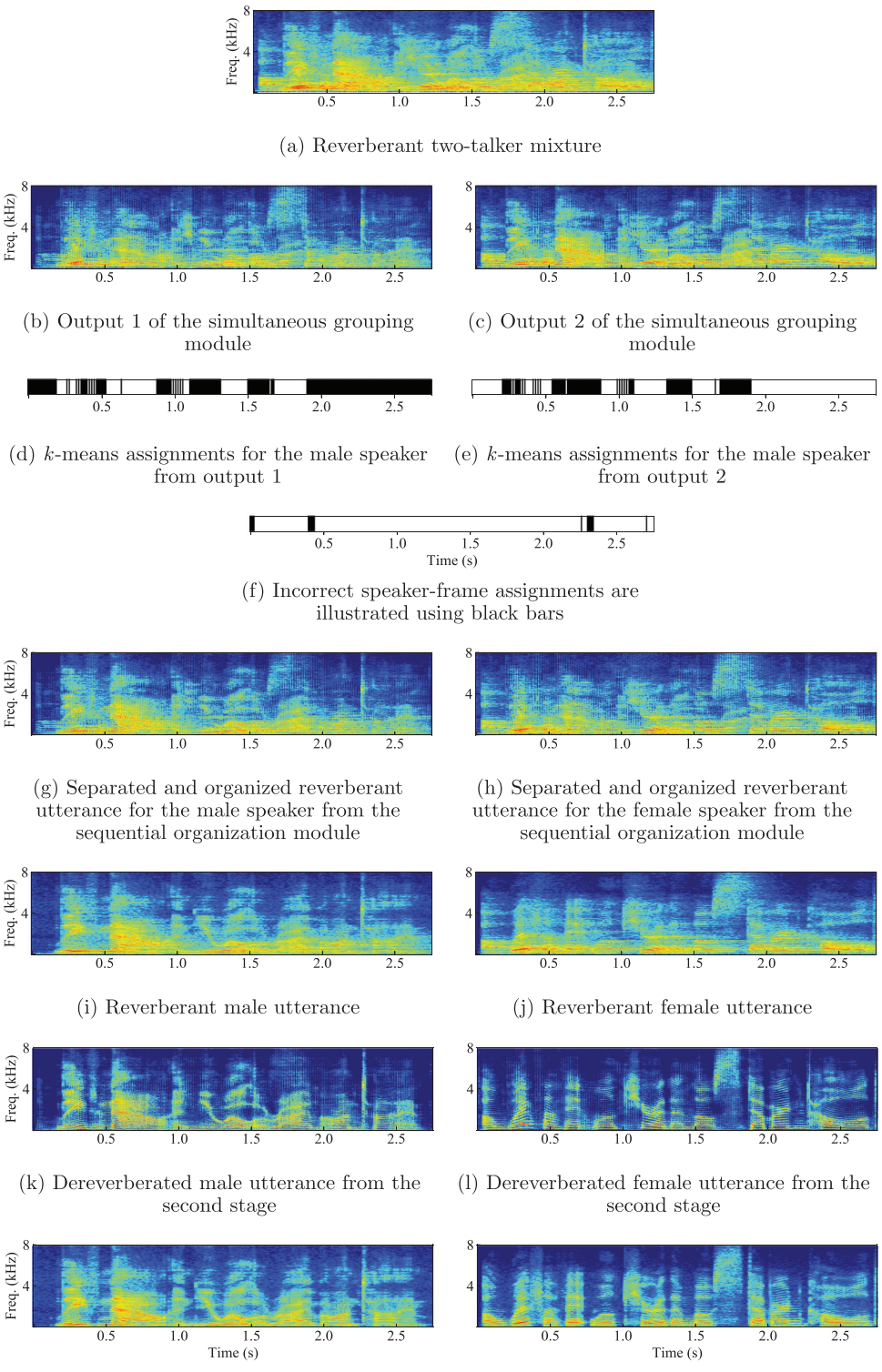
FIG. 5. Percentage of incorrect frame-speaker assignments in deep CASA systems for separating IEEE male-female mixtures.

More challenging are the mixtures of reverberant IEEE speakers using recorded RIRs. Table V shows the performance of different systems in such a condition. The results are consistent with the simulated room results in Table IV. In recorded RIR conditions, two-stage deep CASA performs better than the baseline one-stage system with an average improvement of 0.8 dB SDR, 3.2% ESTOI, and 0.13 PESQ scores. Interestingly, the proposed two-stage deep CASA system approaches the performance of the talker-dependent system—the gold standard for talker-independent speaker separation—with even a slightly 0.2 dB higher SDR score.

In Tables IV and V, we observe that two deep CASA systems exhibit similar performances using optimal or learned sequential organization. These systems make very few errors in organizing separated frames over time, which is expected for different-gender speakers. This effect is illustrated in Fig. 5, where we present the frame-speaker assignment errors in sequential grouping for separating the IEEE mixtures in both simulated and recorded room conditions. Figure 5 shows that two-stage deep CASA makes fewer errors than one-stage deep CASA in both simulated and recorded rooms.

Figure 6 illustrates the separation of a male and a female utterance from the IEEE corpus using the proposed two-stage deep CASA system. Figure 6 shows the separated utterances at the frame level and the sequential organization results. The dereverberated and separated utterances resemble the direct-path premixed speech signals.

Finally, we evaluate on the LibriSpeech corpus. Table VI presents the results using the test speakers from the LibriSpeech corpus. Both same-gender and different-gender cases exist in these mixtures, and recorded RIRs



(a) Reverberant two-talker mixture

(b) Output 1 of the simultaneous grouping module

(c) Output 2 of the simultaneous grouping module

(d) k-means assignments for the male speaker from output 1

(e) k-means assignments for the male speaker from output 2

(f) Incorrect speaker-frame assignments are illustrated using black bars

(g) Separated and organized reverberant utterance for the male speaker from the sequential organization module

(h) Separated and organized reverberant utterance for the female speaker from the sequential organization module

(i) Reverberant male utterance

(j) Reverberant female utterance

(k) Dereverberated male utterance from the second stage

(l) Dereverberated female utterance from the second stage

(m) Anechoic male utterance

(n) Anechoic female utterance

FIG. 6. (Color online) Separating an IEEE male speaker from an IEEE female speaker using the proposed two-stage deep CASA algorithm. The algorithm is trained using mixtures from the WSJ0 speech corpus. Real RIRs from a room with $T_{60} = 0.89$ s were used and mixture TIR was 0 dB. The male sentence was "Leave now and you will arrive on time," and the female sentence was "A whiff of it will cure the most stubborn cold."

are used in generating reverberant mixtures. Similar to earlier results, all three algorithms benefit from two-stage processing with the smallest gain for Conv-TasNet. The two-stage deep CASA system achieves the best performance. Compared to the results in Table V, there are larger gaps between learned and optimal sequential organization, indicating higher frame-assignment errors. This is likely due to the presence of same-gender mixtures in Table VI.

V. CONCLUDING REMARKS

In this paper, we have extended the deep CASA framework to address talker-independent speaker separation in reverberant conditions. We have proposed a two-stage deep CASA algorithm and showed that this algorithm outperforms a baseline deep CASA algorithm in real room reverberant conditions. We have also evaluated an unseen speech

https://doi.org/10.1121/10.0001779

TABLE VI. ΔSDR, ESTOI (%), and PESQ scores for speaker separation in recorded reverberant conditions using the LibriSpeech mixtures.

Metrics		ΔSDR (dB)					ESTOI (%)					PESQ				
T_{60} (s)	Sequential grouping	0.32	0.47	0.68	0.89	Avgerage	0.32	0.47	0.68	0.89	Avgerage	0.32	0.47	0.68	0.89	Avgerage
Unprocessed	—	0.0	0.0	0.0	0.0	0.0	42.0	38.8	46.5	35.0	40.6	1.59	1.56	1.72	1.46	1.58
IRM	—	*10.9*	*10.7*	*11.1*	*10.4*	*10.8*	*83.2*	*82.1*	*86.3*	*82.8*	*83.58*	*3.23*	*3.21*	*3.30*	*3.14*	*3.22*
One-stage uPIT Dense-Unet	Optimal	7.7	8.0	8.8	8.9	8.4	67.6	64.1	70.9	62.2	66.2	2.42	2.33	2.52	2.24	2.38
	Default	6.3	6.3	7.4	7.4	6.9	64.0	59.5	67.3	58.0	62.2	2.23	2.12	2.35	2.06	2.19
Two-stage uPIT Dense-Unet	Default	6.9	6.6	7.8	7.5	7.2	68.9	64.7	70.0	61.1	66.2	2.32	2.21	2.37	2.07	2.24
One-stage Conv-TasNet	Default	5.3	5.8	7.1	7.0	6.3	60.6	58.1	65.9	56.3	60.2	2.13	2.07	2.31	2.02	2.13
Two-stage Conv-TasNet	Default	6.0	6.0	7.0	7.0	6.5	63.9	60.2	66.4	58.3	62.1	2.18	2.12	2.29	1.99	2.14
One-stage deep CASA	Optimal	9.4	9.6	10.5	10.5	10.0	72.9	69.7	75.6	68.1	71.6	2.59	2.53	2.68	2.42	2.55
	Learned	8.1	8.0	9.3	9.1	8.6	69.5	65.6	72.2	63.9	67.8	2.38	2.28	2.48	2.19	2.33
Two-stage deep CASA	Optimal	10.5	10.7	11.6	11.2	11.0	76.4	73.9	79.3	72.3	75.5	2.73	2.65	2.83	2.50	2.68
	Learned	**8.8**	**8.9**	**9.8**	**9.5**	**9.2**	**72.0**	**69.0**	**74.7**	**67.0**	**70.6**	**2.47**	**2.38**	**2.58**	**2.23**	**2.42**

corpus and found that the two-stage system has a superior separation performance.

We have trained Conv-TasNet and uPIT Dense-Unet algorithms in reverberant conditions and compared their performances with the deep CASA algorithms. Deep CASA is shown to have a stronger capacity in generalizing to real-world conditions with unseen room acoustics and speech corpora. A major difference between deep CASA and these comparison systems is their frame organization mechanisms. uPIT ties speaker frames to an output layer throughout a mixture signal, which limits the flexibility of uPIT Dense-Unet and Conv-TasNet. Deep CASA uses tPIT with a strong sequential organization module that predicts the underlying speaker-frame assignments with high accuracy.

The goal of this study is separating and dereverberating two-talker mixtures. Given the same input signal, an alternative goal would be separating reverberant utterances without dereverberation, corresponding to the first stage of our model (see Fig. 1). Based on the first stage evaluation results, deep CASA is equally adept at this task, achieving similar levels of performance. This is to be expected as the separation task is carried out by the first stage.

It is remarkable that our two-stage deep CASA algorithm performs comparably to oracle speaker separation with the IRM as well as talker-dependent speaker separation. In the future, we plan to extend two-stage deep CASA to speaker separation involving more than two speakers and in noisy-reverberant conditions.

ACKNOWLEDGMENTS

This research was supported in part by The National Institute on Deafness and Other Communication Disorders (NIDCD) Grant No. R01DC012048 and the Ohio Supercomputing Center.

[1]The mixture generation script was adopted from http://www.merl.com/demos/deep-clustering/create-speaker-mixtures.zip (Last viewed 8/24/2020).

[2]See https://github.com/ehabets/RIR-Generator (Last viewed 8/24/2020).

Bai, S., Kolter, J. Z., and Koltun, V. (**2018**). "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," arXiv:1803.01271.

Brungart, D. S. (**2001**). "Informational and energetic masking effects in the perception of two simultaneous talkers," J. Acoust. Soc. Am. **109**, 1101–1109.

Culling, J. F., Hodder, K. I., and Toh, C. Y. (**2003**). "Effects of reverberation on perceptual segregation of competing voices," J. Acoust. Soc. Am. **114**, 2871–2876.

Delfarah, M., Liu, Y., and Wang, D. L. (**2020**). "Talker-independent speaker separation in reverberant conditions," in *Proc. ICASSP*, pp. 8723–8727.

Delfarah, M., and Wang, D. L. (**2019**). "Deep learning for talker-dependent reverberant speaker separation: An empirical study," IEEE/ACM Trans. Audio, Speech, Lang. Process. **27**, 1839–1848.

Du, J., Tu, Y., Xu, Y., Dai, L., and Lee, C.-H. (**2014**). "Speech separation of a target speaker based on deep neural networks," in *Proc. ICSP*, pp. 473–477.

Festen, J. M., and Plomp, R. (**1990**). "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing," J. Acoust. Soc. Am. **88**, 1725–1736.

Garofolo, J., Graff, D., Paul, D., and Pallett, D. (**1993**). "CSR-I (WSJ0) complete LDC93S6A," (Linguistic Data Consortium, Philadelphia).

Grais, E. M., Roma, G., Simpson, A. J., and Plumbley, M. D. (**2017**). "Two-stage single-channel audio source separation using deep neural networks," IEEE/ACM Trans. Audio, Speech, Lang. Process. **25**, 1773–1783.

Healy, E. W., Delfarah, M., Johnson, E. M., and Wang, D. L. (**2019**). "A deep learning algorithm to increase intelligibility for hearing-impaired listeners in the presence of a competing talker and reverberation," J. Acoust. Soc. Am. **145**, 1378–1388.

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

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

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

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (**2017**). "Densely connected convolutional networks," in *Proc. CVPR*, pp. 4700–4708.

Huang, P.-S., Kim, M., Hasegawa-Johnson, M., and Smaragdis, P. (**2014**). "Deep learning for monaural speech separation," in *Proc. ICASSP*, pp. 1562–1566.
Huang, P.-S., Kim, M., Hasegawa-Johnson, M., and Smaragdis, P. (**2015**). "Joint optimization of masks and deep recurrent neural networks for monaural source separation," IEEE/ACM Trans. Audio, Speech, Lang. Process. **23**, 2136–2147.
Hummersone, C., Mason, R., and Brookes, T. (**2010**). "Dynamic precedence effect modeling for source separation in reverberant environments," IEEE Trans. Audio, Speech, Lang. Process. **18**, 1867–1871.
IEEE (**1969**). "IEEE recommended practice for speech quality measurements," IEEE Trans. Audio Electroacoust. **17**, 225–246.
Jensen, J., and Taal, C. H. (**2016**). "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," IEEE/ACM Trans. Audio, Speech, Lang. Process. **24**, 2009–2022.
Kingma, D., and Ba, J. (**2015**). "Adam: A method for stochastic optimization," in *Proc. ICML*.
Kolbæk, M., Yu, D., Tan, Z.-H., Jensen, J., Kolbaek, M., Yu, D., Tan, Z.-H., and Jensen, J. (**2017**). "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," IEEE/ACM Trans. Audio, Speech, Lang. Process. **25**, 1901–1913.
Lea, C., Vidal, R., Reiter, A., and Hager, G. D. (**2016**). "Temporal convolutional networks: A unified approach to action segmentation," in *European Conference on Computer Vision*, pp. 47–54.
Liu, Y., and Wang, D. L. (**2019**). "Divide and conquer: A deep CASA approach to talker-independent monaural speaker separation," IEEE/ACM Trans. Audio, Speech, Lang. Process. **27**, 2092–2102.
Luo, Y., and Mesgarani, N. (**2019**). "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation," IEEE/ACM Trans. Audio, Speech, Lang. Process. **27**, 1256–1266.
Miller, G. A. (**1947**). "The masking of speech," Psychol. Bull. **44**, 105–129.
Moore, B. C. (**2007**). *Cochlear Hearing Loss: Physiological, Psychological and Technical Issues* (Wiley, Chichester, UK).
Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (**2015**). "LibriSpeech: An ASR corpus based on public domain audio books," in *Proc. ICASSP*, pp. 5206–5210.
Rix, A. W., Beerends, J. G., Hollier, M. P., and Hekstra, A. P. (**2001**). "Perceptual evaluation of speech quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, pp. 749–752.
Ronneberger, O., Fischer, P., and Brox, T. (**2015**). "U-net: Convolutional networks for biomedical image segmentation," in *Med. Image. Comput. Comput. Assist. Interv.*, pp. 234–241.
Shi, Z., Lin, H., Liu, L., Liu, R., and Han, J. (**2019**). "FurcaNeXt: End-to-end monaural speech separation with dynamic gated dilated temporal convolutional networks," arXiv:1902.04891.
Smaragdis, P. (**2006**). "Convolutive speech bases and their application to supervised speech separation," IEEE Trans. Audio, Speech, Lang. Process. **15**, 1–12.
Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. (**2011**). "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," IEEE Trans. Audio, Speech, Lang. Process. **19**, 2125–2136.
Tan, K., and Wang, D. L. (**2018**). "A two-stage approach to noisy cochannel speech separation with gated residual networks," in *Proc. Interspeech*, pp. 3484–3488.
Vincent, E., Gribonval, R., and Févotte, C. (**2006**). "Performance measurement in blind audio source separation," IEEE Trans. Audio, Speech, Lang. Process. **14**, 1462–1469.
Wang, D. L. (**2005**). "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, edited by P. Divenyi (Springer, Kluwer Academic, Norwell, MA), pp. 181–197.
Wang, D. L., and Brown, G. J., eds. (**2006**). in *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications* (Wiley-IEEE, Hoboken, NJ).
Wang, D. L., and Chen, J. (**2018**). "Supervised speech separation based on deep learning: An overview," IEEE/ACM Trans. Audio, Speech, Lang. Process. **26**, 1702–1726.
Wang, Y., Du, J., Dai, L.-R., and Lee, C.-H. (**2017**). "A gender mixture detection approach to unsupervised single-channel speech separation based on deep neural networks," IEEE/ACM Trans. Audio, Speech, Lang. Process. **25**, 1535–1546.
Wang, Y., Narayanan, A., and Wang, D. L. (**2014**). "On training targets for supervised speech separation," IEEE/ACM Trans. Audio, Speech, Lang. Process. **22**, 1849–1858.
Wang, Y., and Wang, D. L. (**2013**). "Towards scaling up classification-based speech separation," IEEE Trans. Audio, Speech, Lang. Process. **21**, 1381–1390.
Wang, Z.-Q., Le Roux, J., and Hershey, J. R. (**2018**). "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *Proc. ICASSP*, pp. 1–5.
Wang, Z.-Q., and Wang, D. L. (**2018**). "Combining spectral and spatial features for deep learning based blind speaker separation," IEEE/ACM Trans. Audio, Speech, Lang. Process. **27**, 457–468.
Weiss, R. J., and Ellis, D. P. (**2010**). "Speech separation using speaker-adapted eigenvoice speech models," Comput. Speech Lang. **24**, 16–29.
Williamson, D. S., Wang, Y., and Wang, D. L. (**2016**). "Complex ratio masking for monaural speech separation," IEEE/ACM Trans. Audio, Speech, Lang. Process. **24**, 483–492.
Yu, D., Kolbæk, M., Tan, Z.-H., and Jensen, J. (**2017**). "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. ICASSP*, pp. 241–245.
Zhang, X.-L., and Wang, D. L. (**2016**). "A deep ensemble learning method for monaural speech separation," IEEE/ACM Trans. Audio, Speech, Lang. Process. **24**, 967–977.
Zhao, Y., Wang, Z.-Q., and Wang, D. L. (**2019**). "Two-stage deep learning for noisy-reverberant speech enhancement," IEEE/ACM Trans. Audio, Speech, Lang. Process. **27**, 53–62.