Attentive Training: A New Training Framework for Speech Enhancement

Ashutosh Pandey and DeLiang Wang, *Fellow, IEEE*

***Abstract*—Dealing with speech interference in a speech enhancement system requires either speaker separation or target speaker extraction. Speaker separation has multiple output streams with arbitrary assignments while target speaker extraction requires additional cueing for speaker selection. Both of these are not suitable for a standalone speech enhancement system with one output stream. In this study, we propose a novel training framework, called *Attentive Training*, to extend speech enhancement to deal with speech interruptions. Attentive training is based on the observation that, in the real world, multiple talkers very unlikely start speaking at the same time, and therefore, a deep neural network can be trained to create a representation of the first speaker and utilize it to attend to or track that speaker in a multitalker noisy mixture. We present experimental results and comparisons to demonstrate the effectiveness of attentive training for speech enhancement.**

***Index Terms*—Attentive training, speech enhancement, speaker extraction, speaker separation, talker-independent.**

I. Introduction

SPEECH signals in the real world are degraded by acoustic interferences, such as background noise, interfering talkers, and room reverberation. Acoustic interferences degrade the intelligibility and quality of speech for both human and machine listeners. For example, the performance of speech based applications, such as automatic speech recognition (ASR), hearing aids, and telecommunications, deteriorates when dealing with degraded speech. Speech enhancement aims at improving the intelligibility and quality of a degraded signal by removing acoustic interference from it. Monaural speech enhancement utilizes recordings from a single microphone to provide a versatile and cost efficient solution to the problem. This study is focused on monaural speech enhancement that can deal with both speech and nonspeech interference.

Speech enhancement has been widely studied in the signal processing community for decades. Some of the traditional methods include spectral subtraction, Wiener filtering and statistical-model-based methods [1]. The rise of deep learning and its application to speech enhancement has led to dramatic advances over the last decade, and it is firmly established as the mainstream methodology today [2].

Popular approaches to speech enhancement utilize time-frequency representations, such as short-time Fourier transform (STFT), to represent input features and training targets, and aim at enhancing only the spectral magnitude [3], [4], [5], [6], [7], [8], [9], [10], [11]. A recent trend has been to jointly enhance the spectral magnitude and phase by using either complex spectrogram enhancement [12], [13], [14], [15], [16], [17], [18], [19], [20] or time-domain speech enhancement [21], [22], [23], [24], [25], [26], [27], [28], [29].

Speech enhancement is generally formulated as the problem of removing nonspeech interferences from a speech signal. However, in the real world, interfering signals can also be speech from interfering talkers. How to deal with interfering talkers in a speech enhancement system? Dealing with interfering talkers requires two steps: speaker selection and speaker extraction. Human listeners have the amazing ability of auditory perception attending to (hence extracting) a single speaker in a multitalker scenario. This ability is widely referred to as the *cocktail party effect* [30], and has inspired the perceptual theory of selective attention [31]. For humans, speaker selection is dependent on listener attention as well as intention. For machine separation so far, we either separate all speakers from a mixture or provide a cueing signal for speaker selection followed by speaker extraction. The former is called speaker separation and the latter is commonly known as target speaker extraction.

Speaker separation is the task of reconstructing all the speakers from a multitalker mixture. Early works on speaker separation were extended from speech enhancement and talker-dependent, i.e., systems that extract speech signals from only a given speaker and cannot generalize to untrained speakers. When extending to talker-independent speaker separation, these models suffer from a well known permutation ambiguity problem, where a DNN is not able to consistently assign output streams to different speakers during training. Deep clustering [32] and permutation invariant training (PIT) [33] are two representative approaches to resolving the permutation ambiguity problem. Deep clustering and its variants [34] employ a DNN to map each T-F unit of the input mixture to an embedding space, where embeddings are trained to be closer for the units corresponding to a single speaker and far for different speakers. Finally, embedding vectors are clustered into groups corresponding to the different speakers in the mixture to obtain a T-F mask for

Manuscript received 27 July 2022; revised 12 January 2023 and 7 March 2023; accepted 13 March 2023. Date of publication 23 March 2023; date of current version 7 April 2023. This work was supported in part by two NIDCD under Grants R01DC012048 and R02DC015521, in part by the Ohio Supercomputer Center, in part by the Pittsburgh Supercomputing Center, and in part by N S F under Grant ACI-1928147. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Xiao-Lei Zhang. *(Corresponding author: Ashutosh Pandey.)*

Ashutosh Pandey is with the Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210 USA (e-mail: pandey.99@osu.edu).

DeLiang Wang is with the Department of Computer Science and Engineering, Center for Cognitive and Brain Sciences, The Ohio State University, Columbus, OH 43210 USA (e-mail: dwang@cse.ohio-state.edu).

Digital Object Identifier 10.1109/TASLP.2023.3260711

each speaker. In contrast, PIT allows for end-to-end optimization to separate speech signals by dynamically assigning the best matching permutation of the ground-truth signals with the output signals. In particular, the simplicity of PIT has led to many subsequent models for speaker separation [35], [36], [37], [38], [39].

Speaker separation can separate all underlying speakers but it assigns output streams arbitrarily, which is not suitable for speech enhancement systems that need to attend to one output stream. For example, if we design a system that always picks a fixed output stream, it will correspond to either silence or sporadic interruptions when the main speech stream goes to one of the other outputs.

Target speaker extraction is the task of extracting a single speaker from a multitalker mixture, where the target speaker is cued using additional information in the form of audio [40], [41], [42], [43], [44], [45] or image [46], [47], [48]. Recent studies have also explored other kinds of cues, such as spatial location [49], [50], speech activity [51], and onset [52]. Target speaker extraction is similar to auditory selective attention, but requires a priori cueing that may not be available in many applications of speech enhancement.

How to extend a speech enhancement system to deal with speech interruptions without requiring speaker cues? This requires designing an intrinsic speaker selection mechanism. Attention is a major part of perception, and this has inspired us to leverage auditory selective attention to address the problem. If a person is listening (attending) to a talker, he would typically continue listening to that talker irrespective of other speech interruptions, particularly when the interruptions are short. Based on this, we propose a new training framework, which we name *attentive training*, for speech enhancement. In real-world environments, it is very unlikely that multiple talkers start speaking at the same time; such a case would lead to their grouping into the same auditory stream on the basis of common onsets [53]. Therefore, we can assume that a given multitalker mixture has nonoverlapping speech intervals at the beginning. With attentive training, a model presented with a multitalker mixture will start attending to (extracting) nonoverlapping speech segments in the beginning and then continue attending to it while ignoring other speakers. In other words, attentive training treats the speech signals of the first speaker as *target speech*, and the utterances of other speakers plus environmental sounds as *background interference*.

The attentive training framework is consistent with the dominant feature integration theory of attention [54]. According to this perceptual account, attention serves to integrate perceptual features extracted in separate analyses into an object. The attended object forms the target (or foreground), and the remaining objects in a scene become the background. Furthermore learning and attending are integral parts of perception.

Note that attentive training uses the onset of the first speaker as a cue for intrinsic speaker selection. In the context of ASR, a similar idea of using speaker onsets as a cue has been proposed in serialized output training (SOT) [55]. The idea of SOT is to output speaker transcriptions from an ASR system in the order of speaker onsets in the input mixture. The proposed attentive training is fundamentally different from SOT as it is designed for a speech enhancement system that aims at extracting only the first speaker from a mixture.

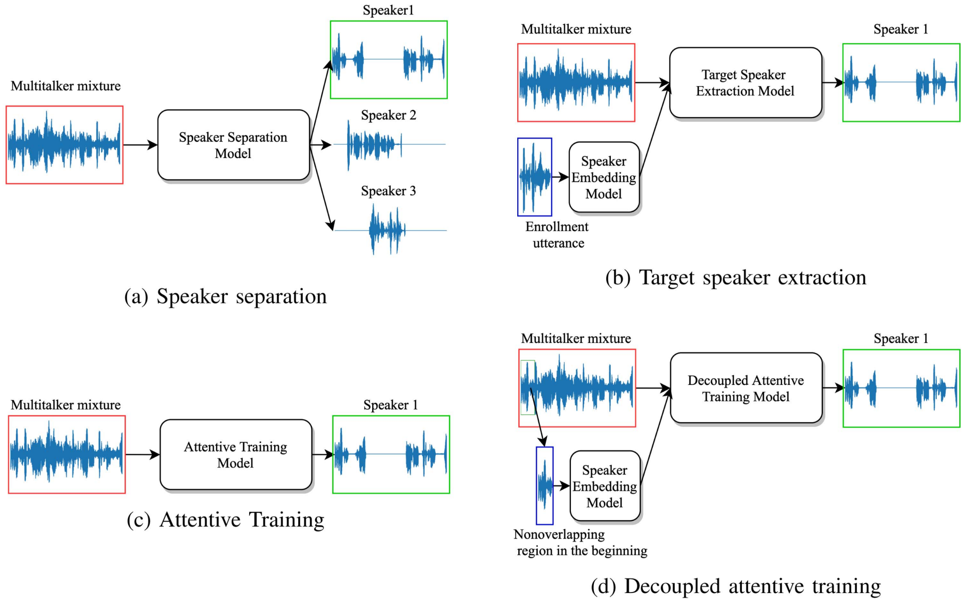
We create a multitalker dataset in a controlled way, where the first speaker is set to start slightly ahead of the rest of the speakers. Next, we train a recently proposed time-domain attentive recurrent network (ARN) [29] with attentive training to estimate the first speaker from a multitalker mixture. We show that ARN is effective in extracting the first speaker and generalizes well to different test conditions, such as an untrained number of speakers, mixtures with larger gaps between the consecutive segments of the target speaker, and smaller speaker overlaps. For instance, a model trained using mixtures with a maximum of 3 speakers obtains strong results for mixtures with 5 speakers.

We compare attentive training with PIT for speaker separation. We find that attentive training obtains substantially better results than PIT when compared on the enhancement metric of the first speaker. We also investigate a decoupled approach to attentive training in which the nonoverlapping speech of the first speaker in the beginning of a mixture is used to create a speaker representation to be used as a cue for target speaker extraction. We observe that end-to-end attentive training obtains better results than decoupled attentive training. We also train a target speaker extraction model using independent enrollment utterances. We find that target speaker extraction with independent enrollment utterances performs slightly better than attentive training. When contrasting target speaker extraction and decoupled attentive training, we conclude that target speaker extraction is slightly better than attentive training only because of additional information in the form of clean enrollment utterances.

We also examine an attentive training model trained with onset differences of more than 1 s between the first and the second speaker, and show that it generalizes well to an onset difference of 0.5 seconds. Also, we train speaker verification systems on top of the hidden layers in ARN to demonstrate that a few of them encode speaker information, which verifies that ARN learns speaker representations implicitly for selection and extraction.

Along the way we introduce a novel data generation technique for mixing an arbitrary number of speakers in a controlled way. Given a set of speakers, their corresponding utterances, and a set of noises, our technique can mix any number of speakers with specified overlaps and speaker orders. Also, mixtures are generated dynamically during training which provides an additional advantage of data augmentation [39]. Our data generation technique should be a useful tool for speaker separation and diarization research, as it can utilize speakers from any corpora and generate mixtures in a flexible way. We provide our data generation script online.

This study focuses on extracting the first speaker from a mixture to illustrate the effectiveness of attentive training. A straightforward and useful extension of attentive training would be to develop a speech enhancement system that aims at removing interfering speech only from the interval of speaker overlaps. The preserved speaker should be the one that enters



(a) Speaker separation

(b) Target speaker extraction

(c) Attentive Training

(d) Decoupled attentive training

Fig. 1. Different methods for extracting the first speaker from a multitalker mixture.

into the overlapping interval from the past. It will reduce to a speech enhancement system handling nonoverlapping speech signals from multiple speakers in the output stream. Designing such a system will require a careful consideration into fixing hyperparameters, such as gaps between consecutive segments of different speakers, to output a perceptually meaningful signal.

We believe that the simple and effective mechanism of attentive training has the potential to be applicable to a variety of selection, tracking, and related tasks, such as multitalker speaker separation and speaker diarization. For speaker separation and diarization, a straightforward extension would be to use an iterative strategy where the first speaker is extracted first, then second, and so on, as in [56].

A preliminary study on attentive training has been published in [57] where a smaller ARN is trained on a smaller dataset and compared only with speaker separation using PIT. The remainder of the paper is organized as follows. A definition and different methods of attentive training are discussed in Section II. Section III describes the data generation algorithm. Section IV details employed DNN models. Experimental settings are given in Section V and results and comparisons are presented in Section VI. Concluding remarks are given in Section VII.

II. Speaker Tracking and Attentive Training

A multitalker mixture $\boldsymbol{y}$ with N samples is modeled as

$$\boldsymbol{y} = \sum_{i=1}^{C} \boldsymbol{s}_i + \boldsymbol{n} \tag{1}$$

where $\{\boldsymbol{y}, \boldsymbol{s}_i\} \in \mathbb{R}^{N\times 1}$, C is the total number of speakers, $\boldsymbol{s}_i$ is the ith speaker, and $\boldsymbol{n}$ is the background noise. Let o_i denote the time sample when ith speaker starts speaking. We assume that speaker indices $i = 1, 2, \ldots, C$ are sorted in the increasing order of onset times. In other words, $i < j$ implies $o_i < o_j$. The goal of attentive training is to separate the first speaker $\boldsymbol{s}_1$ from $\boldsymbol{y}$.

We can extract the first speaker from a mixture using the following methods.

A. Speaker Separation

A speaker separation system has no selection mechanism and reconstructs all the speakers in a mixture. Speaker separation can be utilized to extract the first speaker by first separating all the speakers and then selecting the first speaker using speech onset. Speaker separation is illustrated in Fig. 1(a), and modeled as

$$[\hat{\boldsymbol{s}}_1, \ldots, \hat{\boldsymbol{s}}_C] = f_{SS}(\boldsymbol{y}) \tag{2}$$

where f_{SS} represents a DNN for speaker separation. The speaker separation model is trained using an utterance-level PIT loss defined as

$$\mathcal{L} = \sum_{i=1}^{C} \mathcal{D}\left(\boldsymbol{s}_{\phi^*(i)}, \hat{\boldsymbol{s}}_i\right) \tag{3}$$

where $\mathcal{D}(\boldsymbol{a}, \boldsymbol{b})$ is a distance measure between signals $\boldsymbol{a}$ and $\boldsymbol{b}$ and ϕ^* is a permutation of target signals with the minimum cost, i.e.,

$$\phi^* = \arg\min_{\mathcal{P}} \sum_{i=1}^{C} \mathcal{D}\left(\boldsymbol{s}_{\phi^*(i)}, \hat{\boldsymbol{s}}_i\right) \tag{4}$$

where $\mathcal{P}$ represents the set of all possible permutations. We use an utterance-level negative signal-to-noise ratio (SNR) as the distance measure, defined as

$$\mathcal{D}(\boldsymbol{s}, \hat{\boldsymbol{s}}) = -10 \cdot \log_{10} \frac{||\boldsymbol{s}||^2}{||\boldsymbol{s} - \hat{\boldsymbol{s}}||^2} \tag{5}$$

B. Target Speaker Extraction

Target speaker extraction extracts a single speaker from a mixture with the help of an additional cue for target selection. The speaker selection mechanism is not intrinsic to model training. We assume that we are given additional information in the form of an enrollment utterance $\boldsymbol{e}_1$ corresponding to the first speaker. Target speaker extraction is illustrated in Fig. 1(b). First, a speaker embedding is computed from $\boldsymbol{e}_1$ as

$$\boldsymbol{v}_1 = h(\boldsymbol{e}_1) \tag{6}$$

where $\boldsymbol{v}_1 \in \mathbb{R}^{B\times 1}$, B is the size of the embedding vector, and h is a DNN-based speaker embedding model. Next, $\boldsymbol{v}_1$ and $\boldsymbol{y}$ are used together to estimate $\boldsymbol{s}_1$ as

$$\hat{\boldsymbol{s}}_1 = f_{TSE}\left(\boldsymbol{y}, \boldsymbol{v}_1\right) \tag{7}$$

where f_{TSE} represents a DNN for target speaker extraction. It is trained using a distance between the estimated and the ground-truth signal of the first speaker as defined below.

$$\mathcal{L} = \mathcal{D}\left(\boldsymbol{s}_1, \hat{\boldsymbol{s}}_1\right) \tag{8}$$

C. Attentive Training

Attentive training aims at estimating $\boldsymbol{s}_1$ directly from $\boldsymbol{y}$ as shown in Fig. 1(c). It is defined as

$$\hat{\boldsymbol{s}}_1 = f_{AT}(\boldsymbol{y}) \tag{9}$$

where f_{AT} represents a DNN for attentive training. It is trained using the loss in (8).

D. Decoupled Attentive Training

Decoupled attentive training decouples end-to-end attentive training in two parts. First, it assumes that we are provided with the nonoverlapping speech segment $\boldsymbol{s}_1^{no}$ in the beginning of $\boldsymbol{y}$, defined as

$$\boldsymbol{s}_1^{no} = \boldsymbol{y}[0:M-1] = \boldsymbol{s}_1[0:M-1] + \boldsymbol{n}[0:M-1] \tag{10}$$

where M is the length of $\boldsymbol{s}_1^{no}$. Next, $\boldsymbol{s}_1^{no}$ is used to generate a speaker embedding of the first speaker

$$\boldsymbol{v}_1^{no} = h(\boldsymbol{s}_1^{no}) \tag{11}$$

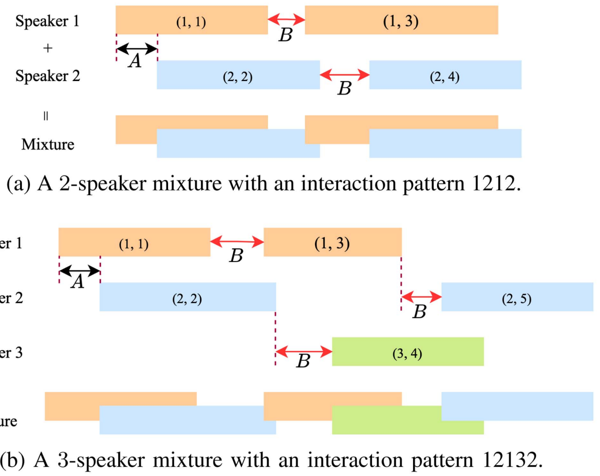
Finally, $\boldsymbol{v}_1^{no}$ and $\boldsymbol{y}$ are used together to estimate $\boldsymbol{s}_1$ as

$$\hat{\boldsymbol{s}}_1 = f_{De\text{-}AT}\left(\boldsymbol{y}, \boldsymbol{v}_1^{no}\right) \tag{12}$$

where $f_{De\text{-}AT}$ represents a DNN for decoupled attentive training. Fig. 1(d) depicts decoupled attentive training. The loss in (8) is used to train a decoupled attentive training model.

III. Data Generation

This section describes our technique for generating multitalker mixtures. Given a set $\boldsymbol{S} = \{S_1, \ldots, S_J\}$ of speakers, their corresponding utterances $\boldsymbol{U}^{S_j} = \{\boldsymbol{s}_1^j, \ldots, \boldsymbol{s}_{Q_j}^j\}$, and a set of noise segments $\boldsymbol{N} = \{\boldsymbol{n}_1, \ldots, \boldsymbol{n}_R\}$, where Q_j denotes the number of utterances of speaker S_j and R is the number of noise segments, we create a multitaker noisy mixture by adding together speech segments of multiple speakers and a noise segment. First, we sort a given set of speech segments in an increasing order of their onset times. Based on this, we define a concept called *interaction pattern* representing the order of speaker segments in a mixture. For example, an interaction pattern of 1212 represents a mixture created by adding 4 segments sorted in the increasing order of their onset times, where the first and the third segments are from the first speaker and the second and the fourth segments are from the second speaker. We also define two parameters A and B, where A is the minimum initial gap between the onset of the first and the second speaker, and B is the gap between two adjacent nonoverlapping segments (regardless of speakers). We illustrate two interaction patterns in Fig. 2. For data generation, we use interaction patterns from a predefined set $\boldsymbol{P} = \{\boldsymbol{p}_1, \ldots, \boldsymbol{p}_P\}$.



(a) A 2-speaker mixture with an interaction pattern 1212.

(b) A 3-speaker mixture with an interaction pattern 12132.

Fig. 2. Examples of interaction patterns with 2 and 3 speakers, and an initial minimum onset gap of A between the first and the second speaker. In pair (a, b) inside a box, a and b respectively index the speaker order and the segment order.

Similar to the LibriCSS dataset [58], we generate mixtures in a way that a given mixture can have an arbitrary number of speakers, but at a given time instant, only a maximum of two speakers can overlap. Algorithm 1 describes the steps used in generating a sample mixture from $\boldsymbol{S}, \boldsymbol{U}, \boldsymbol{N}$, and $\boldsymbol{P}$. In the algorithm, Len$(\boldsymbol{x})$ represents the length of $\boldsymbol{x}$, and Unique$(\boldsymbol{p})$ denotes the set of unique elements in $\boldsymbol{p}$.

In Algorithm 1, the list $\boldsymbol{E}$ is used to keep track of allowed overlap intervals and $E[-k]$ denotes the kth element in $\boldsymbol{E}$ from the end. The allowed interval spans from $E[-2]$ to $E[-1]$, which indicate the ending time samples of the last two segments. The set $\boldsymbol{E}_1$ is used to make sure that two different segments from the same speaker do not overlap (line 27 in Algorithm 1). We remove silences from all utterances and then pad zeroes in the beginning to shift a given segment. We use no padding for the first speaker, the second speaker has a minimum padding of A, and the remaining speakers use zero padding in a way that a maximum of two speakers overlaps at a time.

IV. DNN Models

We employ a recently proposed ARN model for time-domain speech enhancement [29]. The model architecture is shown in Fig. 3. It comprises an input linear layer followed by four

Algorithm 1: A Pseudo Code for Generating a Random Multitalker Noisy Mixture.

```
Input: $\boldsymbol{S}, \boldsymbol{U}, \boldsymbol{N}, \boldsymbol{P}$
Output: $\boldsymbol{y}, \boldsymbol{s}_1, \dots, \boldsymbol{s}_C, \boldsymbol{n}$
Sample a speaker pattern $\boldsymbol{p}$ from $\boldsymbol{P}$
Set $C$ = Len(Unique($\boldsymbol{p}$))
Sample $C$ speakers $S_1, \dots, S_C$ from $\boldsymbol{S}$
Initialize List $\boldsymbol{V}$ = [ ], $\boldsymbol{E}$ = [ ], Set $\boldsymbol{E_1}$ ={ }, Bool
  Overlap = False
for $j$ in $\boldsymbol{p}$ and $i$ in {1, 2,..., Len($\boldsymbol{p}$)} do
  Sample an utterance $\boldsymbol{s}$ from $\boldsymbol{U}^{S_j}$;
  Remove silences in the beginning and the end of $\boldsymbol{s}$
  Sample a value $T$ for the segment length
  Extract a random segment $\boldsymbol{x}$ of length $T$ from $\boldsymbol{s}$
  Set Overlap = True with a probability $p_{overlap}$
  if $i = 1$ then
    PadLeft = 0
  else if $i = 2$ then
    Sample a value for $B$
    if $j$ in $\boldsymbol{E}_1$ then
      Overlap = False
    end if
    if no Overlap then
      Set PadLeft = $E[-1] + B$
    else
      Sample PadLeft from $[A, E[-1]]$
    end if
  else
    Sample a value for $B$
    if $j$ in $\boldsymbol{E}_1$ and $E_1$[j] = $E[-1]$ then
      Overlap = False
    end if
    if no Overlap then
      Set PadLeft = $E[-1] + B$
    else
      Sample PadLeft from $[E[-2] + B, E[-1]]$
    end if
  end if
  Apply a left padding of PadLeft to $\boldsymbol{x}$
  Set $\boldsymbol{V}[i] = \boldsymbol{x}$
  Insert Len($\boldsymbol{x}$) + PadLeft at the end of $\boldsymbol{E}$
  Set $E_1[j]$ = Len($\boldsymbol{x}$) + PadLeft
  Sort $\boldsymbol{E}$ in increasing order
  if Len($\boldsymbol{E}$) > 2 then
    Set $\boldsymbol{E} = [E[-2], E[-1]]$
  end if
end for
Apply right padding to segments in $\boldsymbol{V}$ to match lengths
Sample a separate value of sound level for all segments
Scale all segments to appropriate levels
Create a multitalker mixture by adding all segments
  together
Sample a noise segment from $\boldsymbol{N}$
Sample a value for noise level
Scale the level of the noise segment and add to the
  multitalker mixture
```

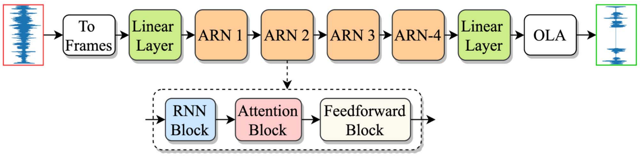


Fig. 3. The model architecture used for attentive training.

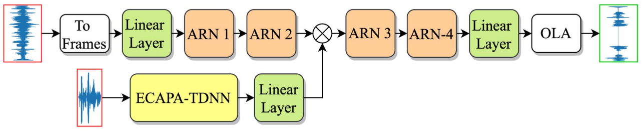


Fig. 4. The model architecture used for target speaker extraction and decoupled attentive training.

ARN layers and an output linear layer. An input mixture $\boldsymbol{y}$ is first converted to frames $\boldsymbol{Y} \in \mathbb{R}^{T \times L}$, where T is the number of frames and L is the frame size. Next, frames in $\boldsymbol{Y}$ are projected to size D, processed by a stack of four ARN layers, and projected back to size L using the output linear layer. Finally, an overlap-and-add (OLA) is used to get the enhanced waveform. An ARN layer comprises an RNN block, a feedforward block, and an attention block. A more detailed description of these blocks can be found in [29]. For speaker separation, we use C linear layers at the output. For decoupled attentive training and target speaker extraction, we utilize a strong speaker embedding model called ECAPA-TDNN [59], which is shown in Fig. 4. The output from ECAPA-TDNN is projected to size D using a linear layer and then multiplied elementwise to the output of the second ARN. We also investigated multiplying to the output of other or all ARNs but observed worse results. We utilize a pretrained ECAPA-TDNN model provided in the SpeechBrain toolkit [60] as it exhibits strong speaker verification performance.

V. Experimental Settings

A. Datasets

We generate training and evaluation data from the LibriSpeech corpus [61]. We use all the speakers from train-clean-100, train-clean-360, and train-other-500 for training. All the speakers from test-clean and dev-clean are used respectively for testing and validation. The training set consists of 960 hours of speech data, which is much larger than the set of 100 hours used in the preliminary study [57].

Noises used are from the WHAM! corpus [62]. First, we split training noises into 10-s chunks, and validation and test noises into 15-s chunks. All chunks shorter than 3 seconds are omitted. We use LKFS based loudness [63] for controlling the SNR. We sample sound levels from $[-25, -30]$ dB for speaker segments and from $[-35, -40]$ dB for noise segments. We provide our dataset generation script along with the test and validation metadata files at https://github.com/ashutosh620/AttentiveTraining.

For target speaker extraction, each multitalker mixture is paired with a randomly sampled enrollment utterance of the first speaker in the mixture. We trim silences from the beginning and the end of an enrollment utterance and truncate longer utterances to a length of 4 seconds.

We also train and evaluate speaker verification systems to examine the speaker information encoded in the hidden layers of the ARN model trained with attentive training. For this, we create a speaker verification dataset using speech from LibriSpeech and noises from WHAM! as in other experiments. We generate training data dynamically by randomly sampling a speech utterance and mixing it with a randomly sampled noise segment. For test and validation, we randomly sample a list of 10000 pairs of noisy speech utterances from different speakers. We sample positive and negative speaker pairs with equal probability.

B. *Training Methodology*

All the utterances are resampled to 16 kHz. A frame size of 16 ms, frame shift of 4 ms, and $D = 1024$ is used for ARN. A smaller ARN model with $D = 512$ was used in the preliminary study [57]. ARN uses BLSTMs with 512 hidden units in both directions. All the models are trained on interaction patterns with 4 segments with a maximum of 3 speakers. In other words, a randomly generated multitalker mixture contains either 1, 2, or 3 speakers. For the PIT model, we use 3 linear layers at the output, and for an input with K ($K <= 3$) speakers, we select the minimum loss assignment from all possible C_K^3 assignments.

All the training samples are randomly and dynamically generated during training, and an episode of 281 k samples (total number of speech utterances) is considered as one epoch. We use $p_{overlap} = 0.75$ and $A = 1$ second. B is sampled from [0.25, 0.50] seconds. Segment length, T, is sampled from [2, 3] seconds for training and from [2, 4] seconds for validation and test. The input and the output are scaled by 25.

All the training utterances longer than 10 seconds are trimmed to 10 seconds. All the models are trained for 100 epochs with a batch size of 32 utterances using the Adam optimizer [64]. The learning rate is initialized with 0.0004 and scaled by 0.98 every two epochs.

Models are evaluated on interaction patterns from {1111, 1212, 1221, 122221, 1231, 123231, 12341, 123451} and three overlap types: {*Max, Half, None*}. Following Algorithm 1, *Max* uses the maximum allowed overlap, *Half* uses half of the allowed regions for overlap, and *None* uses no overlap. We generate 3000 evaluation utterances for each combination of the interaction pattern and overlap type. The pattern 1212 is used to assess performance for an alternating pattern of the target and interfering speaker, 1221 is used to assess performance with a larger gap between two consecutive segments of the target speaker. The pattern 122221 is used to assess performance with an even larger gap not used during training. Similarly, patterns 1231 and 123231 are used to assess performance for 3 speakers with different gaps, where 123231 is not used during training. Patterns 12341 and 123451 are used to assess performance for untrained numbers of 4 and 5 speakers. We use the interaction pattern 1231 with *Max* overlap for validation.

The ECAPA-TDNN model is trained using a set of 7.2 k speakers from the VoxCeleb1 [65] and VoxCeleb2 [66] corpora. Data augmentation techniques, such as additive noise, room reverberation, speed perturbation, and SpecAugment [67] are also utilized. An additive angular margin loss with a margin of 0.2 and scale of 30 is used [68], [69]. A more detailed description can be found in the SpeechBrain toolkit [60].

We also evaluate attentive training for sporadic speech interruptions, which occur often in daily environments. For this, we generate a test dataset with longer interaction patterns from $\{1211111, 1112111, 1111121\}$. Each of these patterns comprises 7 segments, 6 of which correspond to the target speaker and 1 corresponds to the interfering speaker.

For training speaker verification systems on top of the hidden layers of the pretrained ARN model, we use a 1-layered ARN model with $D = 256$ followed by a statistical pooling borrowed from ECAPA-TDNN. The pooling layer uses 128 channels for attention [59], [60]. The embedding size is set to 32. A batch of training data comprises 32 pairs of 3 seconds long utterances, where a pair consists of one noisy and one clean utterance from the same speaker. We trim silences from the beginning and the end. All models are trained with a cyclical learning rate varying between 0.00004 and 0.0004 using the triangular policy as described in [70] in conjunction with the Adam optimizer [64].

We develop all the models in PyTorch [71] and exploit automatic mixed precision training to expedite training [72]. Two NVIDIA Volta V100 32 GB GPUs are utilized to train all attentive training models.

We use scale-invariant SNR (SI-SNR), extended short-time objective intelligibility (eSTOI) [73], and perceptual evaluation of speech quality (PESQ) [74] as evaluation metrics. Objective scores are computed for the first speaker and eSTOI is reported in percentage.

C. *Baseline Models*

We also evaluate the effectiveness of attentive training for two widely used models for speaker separation: convolutional time-domain audio separation network (Conv-TasNet) [36] and dual-path recurrent neural network (DPRNN) [37]. We train these models using the four methods shown in Fig. 1. We modify these models to use one output stream for AT, De-AT and TSE, and 3 output streams for speaker separation. For Conv-TasNet, we utilize the best performing model in [36] which uses $R = 3$ repeats of $X = 8$ convolutional blocks. For De-AT and TSE, speaker embeddings are fused after the first repeat using elementwise multiplication. Similarly, we utilize the best performing DPRNN architecture in [37], which uses a stack of 6 dual-path blocks including intra-chunk and intra-chunk RNN. To train DPRNN for De-AT and TSE, we fuse speaker embeddings after the third dual-path block using elementwise multiplication. We also train a time-domain model called SpEx+ proposed specifically for TSE [75].

VI. Results and Comparisons

We denote speech enhancement as SE, attentive training as AT, speaker separation as PIT, decoupled attentive training as De-AT, and target speaker extraction as TSE in the results. A speech enhancement model is trained only on the interaction pattern 1111, i.e., single-talker utterances with background

TABLE I
COMPARING DIFFERENT METHODS ON THE INTERACTION PATTERN 1111

Metric	Mix.	PIT	AT	De-AT	TSE	SE
SI-SNR	9.5	15.2	**17.5**	17.4	17.3	**19.1**
PESQ	2.33	3.24	**3.40**	3.38	3.38	**3.61**
ESTOI	72.2	87.6	**89.8**	89.2	89.3	**92.6**

TABLE II
COMPARING DIFFERENT METHODS FOR TRAINED NUMBERS OF SPEAKERS

			Type	*Max*			*Half*			*None*		
(a)	(b)	(c)	Metric	SI-SNR	PESQ	ESTOI	SI-SNR	PESQ	ESTOI	SI-SNR	PESQ	ESTOI
2	1212	✓	Mix.	-0.6	1.67	51.4	-0.7	1.86	59.5	-1.0	2.28	71.9
			PIT	11.4	2.64	76.4	12.4	2.79	80.3	14.2	3.04	86.4
			AT	**13.4**	**2.90**	**81.7**	**14.8**	**3.09**	**85.0**	**16.4**	3.38	**89.6**
			De-AT	12.6	2.86	80.3	14.5	3.06	84.3	**16.4**	**3.39**	89.1
			TSE	**13.8**	**2.97**	**82.7**	**15.0**	**3.12**	**85.4**	**16.9**	**3.43**	**89.3**
	1221	✓	Mix.	-0.6	1.77	51.2	-0.8	2.06	60.1	-1.0	2.31	71.6
			PIT	11.3	2.71	76.0	12.5	2.92	80.5	14.3	3.06	86.0
			AT	**13.2**	**2.97**	**81.2**	**14.7**	**3.28**	**85.0**	16.5	3.47	**89.3**
			De-AT	12.5	2.95	80.0	14.5	3.25	84.4	**16.7**	**3.49**	88.9
			TSE	**13.9**	**3.07**	**82.8**	**15.1**	**3.32**	**85.6**	**17.1**	**3.55**	**89.1**
	122221	×	Mix.	-3.7	2.00	51.1	-3.8	2.16	60.2	-3.9	2.32	71.6
			PIT	11.2	2.83	76.1	12.5	2.94	80.5	14.1	3.03	85.7
			AT	**12.9**	**3.23**	**80.9**	**14.6**	**3.44**	**84.9**	16.2	3.52	**89.2**
			De-AT	12.4	3.22	80.0	14.5	3.40	84.5	**16.6**	**3.53**	89.0
			TSE	**13.8**	**3.33**	**82.7**	**15.2**	**3.49**	**85.9**	**17.1**	**3.62**	**89.5**
3	1231	✓	Mix.	-0.6	1.86	55.3	-0.8	2.08	62.7	-1.0	2.32	71.7
			PIT	10.3	2.66	76.0	12.5	2.90	81.5	13.3	3.04	86.4
			AT	**13.2**	**2.97**	**82.4**	**15.2**	**3.26**	**86.0**	**16.3**	3.46	**89.3**
			De-AT	12.2	2.93	80.8	14.7	3.23	85.1	16.2	**3.48**	88.9
			TSE	**14.0**	**3.08**	**83.8**	**15.4**	**3.31**	**86.4**	**16.8**	**3.54**	**89.3**
	123231	×	Mix.	-3.6	1.98	55.6	-3.7	2.14	62.6	-3.9	2.32	71.6
			PIT	9.5	2.68	75.8	11.6	2.87	81.0	12.2	2.97	85.7
			AT	**12.7**	**3.07**	**82.5**	**14.6**	3.30	**85.8**	**15.9**	**3.50**	**89.3**
			De-AT	11.3	3.03	80.5	14.2	**3.31**	84.9	15.8	3.50	89.1
			TSE	**13.6**	**3.18**	**83.8**	**15.3**	**3.40**	**86.5**	**16.6**	**3.60**	**89.5**

(a) number of speakers, (b) interaction pattern, (c) whether trained on interaction pattern.

noise. Background noise is present in all of the following evaluations.

A. Comparing Different Methods

We start by comparing different methods for the interaction pattern 1111. Results are given in Table I. We observe that SE is the best, PIT is the worst, and AT, De-AT, and TSE obtain similar results. We expect SE to obtain best results for this case as it is trained specifically for the matched interaction pattern of 1111. This result suggests that a model capable of dealing with interfering speech performs worse at removing noise than a model trained specifically for removing noise. In other words, the capability of handling interfering speech comes at the expense of noise removal.

Next, we compare different methods for the multitalker case with 2 and 3 speakers and the trained number of speakers. Results are given in Table II. We can observe that a general order of performance among different methods is PIT $<$ De-AT $<$ AT $<$ TSE. In particular, the performance of PIT is far worse than the other methods for all the cases. This highlights a major issue with PIT when dealing with a varying number of speakers and varying degrees of overlaps [39], [58]. We also observe that AT is similar or better than De-AT. This is encouraging because it implies that end-to-end training can better learn the joint task of speaker selection and tracking than a decoupled approach. As expected, TSE obtains the best results since it is provided with additional cueing in the form of an enrollment utterance. It is worth mentioning that De-AT also uses cueing, but the cueing signal comes from the input mixture itself, and hence, it does not provide additional information on top of the input. Also, TSE uses a clean cueing signal in contrast to a noisy one in De-AT.

TABLE III
COMPARING DIFFERENT METHODS FOR THE CASE OF UNTRAINED NUMBER OF SPEAKERS

		Type	*Max*			*Half*			*None*		
(a)	(b)	Metric	SI-SNR	PESQ	ESTOI	SI-SNR	PESQ	ESTOI	SI-SNR	PESQ	ESTOI
4	12341	Mix.	-2.4	1.98	57.5	-2.5	2.12	62.6	-2.8	2.31	71.8
		PIT	7.6	2.65	76.1	7.6	2.77	79.5	9.1	2.92	85.4
		AT	**13.0**	**3.07**	**83.3**	**14.7**	**3.28**	**85.9**	**16.0**	3.49	**89.3**
		De-AT	12.0	3.02	81.7	14.2	3.28	85.0	15.7	**3.50**	89.0
		TSE	**13.9**	**3.17**	**84.4**	**15.2**	**3.36**	**86.3**	**16.5**	**3.57**	**89.3**
5	123451	Mix.	-3.6	1.98	55.7	-3.7	2.14	62.8	-4.0	2.32	71.8
		PIT	4.9	2.55	73.5	4.3	2.67	77.4	6.0	2.85	85.0
		AT	**12.2**	**3.04**	**82.4**	**14.4**	**3.31**	**86.1**	**15.8**	**3.51**	**89.6**
		De-AT	11.4	3.02	81.0	14.0	3.30	85.2	15.5	**3.51**	89.3
		TSE	**13.4**	**3.17**	**83.7**	**15.1**	**3.40**	**86.4**	**16.3**	**3.59**	**89.3**

(a) number of speakers, (b) interaction pattern.

Finally, we present evaluation result for the untrained numbers of 4 and 5 speakers in Table III. We observe similar performance trends to 2 and 3 speakers except for PIT which is much worse because it is not designed to separate the number of speakers not used during training. It is worth noting that AT obtains an SNR improvement of around 15 dB on higher numbers of untrained speakers. This implies that AT does not require training with more than 3 speakers to obtain good generalization.

We plot spectrograms of a sample multitalker mixture enhanced using different methods in Fig. 5. Notice that not only PIT introduces leakage from interfering talkers in the silence intervals but also removes high-frequency speech components. Plots of AT and TSE look very similar with much reduced leakage and well-retained high-frequency components.

B. Comparison With Baselines

Fig. 6 plots the performance of Conv-TasNet, DPRNN, ARN and SpEx+ on interaction pattern 123231. First, we observe a general trend that TSE is the best and AT is better than PIT and De-AT, except for Conv-TasNet with overlap type *None* where AT is worse than PIT and De-AT. This may be due to the fact that Conv-TasNet is a fully convolutional model and it does not have a mechanism to store and propagate speaker identity over time. Additionally, ARN is the best performing model for AT, De-AT and TSE. It is encouraging to observe that ARN outperforms SpEx+, the baseline model proposed specifically for TSE. It is interesting to note that the performance differences between Conv-TasNet and DPRNN are not as significant as observed on WSJ0-2mix and WSJ0-3mix datasets with full overlap. Finally, we notice that even though ARN has the best performance for the cases with a single output stream, it has worst performance for PIT, which uses 3 output streams.

C. Importance of Attentive Training for Speech Enhancement

We have reported in Table I that SE obtains better results than AT when dealing with single-talker input. What happens when a SE model is presented with an input mixture with sporadic speech interruptions? Now, we present results to assess this

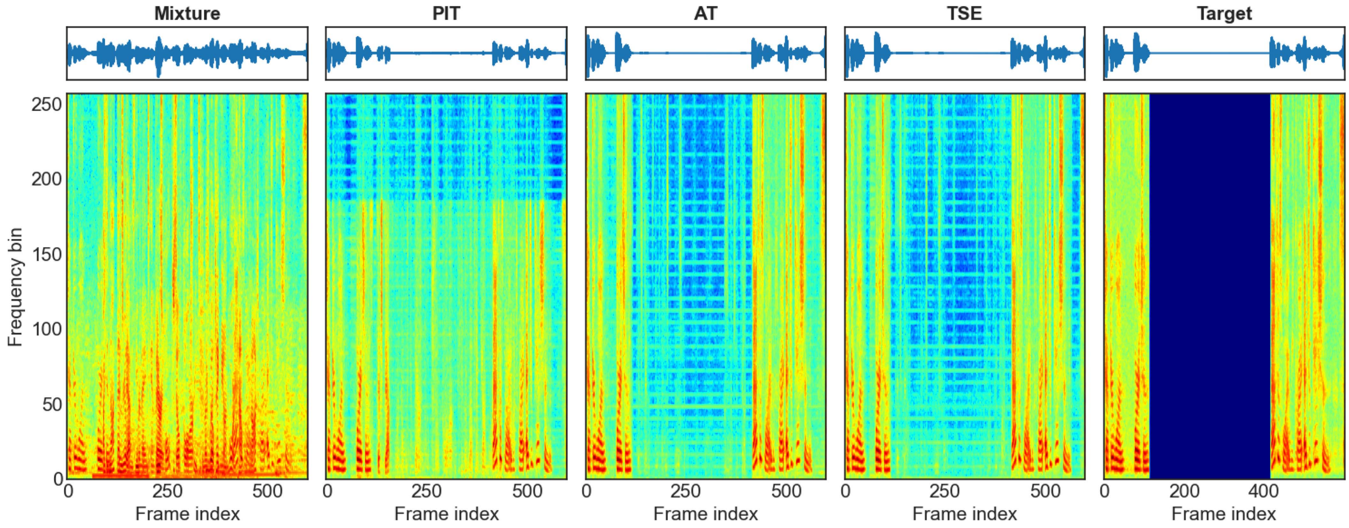


Fig. 5. Spectrograms of a sample multitalker mixture enhanced using different methods.

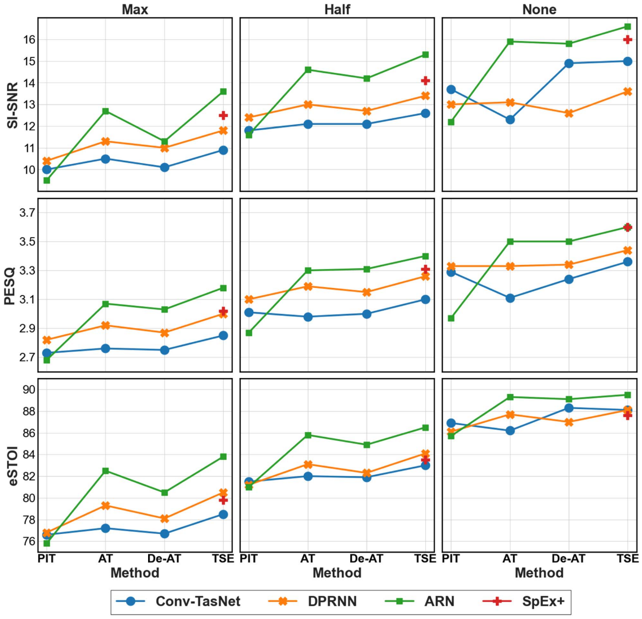


Fig. 6. Comparison of Conv-TasNet, DPRNN, ARN and SpEx+ on interaction pattern 123231 for three types of overlap.

aspect. We compare AT and SE in Table IV on interaction patterns 1211111, 1112111, and 1111121, which are designed to simulate sporadic interruption scenarios.

We observe that AT obtains much better scores in most of the cases, which suggests that speech enhancement fails when presented with speech interruptions. Attentive training enables speech enhancement to deal with speech interruptions, and this is an important advantage of AT. We notice that AT is better for eSTOI for overlap type *Max* and *Half* but worse for *None*. We believe this is because the computation of eSTOI ignores silence intervals in the target signal, hence favoring SE in nonoverlapping intervals.

TABLE IV
COMPARING SE AND AT FOR SPORADIC SPEECH INTERRUPTIONS

	Type	*Max*			*Half*			*None*		
(a)	Metric	SI-SNR	PESQ	eSTOI	SI-SNR	PESQ	eSTOI	SI-SNR	PESQ	eSTOI
1211111	Mix.	5.6	2.19	68.5	5.5	2.23	69.8	5.2	2.33	72.4
	SE	8.7	3.07	85.8	8.2	3.12	87.9	7.7	3.29	**92.5**
	AT	**15.9**	**3.26**	**88.0**	**16.2**	**3.31**	**88.6**	**16.2**	**3.37**	89.4
1112111	Mix.	5.5	2.18	68.5	5.5	2.21	69.4	5.2	2.33	72.5
	SE	8.5	3.05	86.0	8.3	3.09	87.3	7.7	3.28	**92.5**
	AT	**16.0**	**3.29**	**88.4**	**16.5**	**3.33**	**88.9**	**16.9**	**3.42**	90.0
1111121	Mix.	5.5	2.18	68.4	5.5	2.21	69.3	5.2	2.33	72.4
	SE	8.5	3.05	86.1	8.3	3.09	87.3	7.7	3.27	**92.6**
	AT	**16.0**	**3.28**	**88.4**	**16.4**	**3.31**	**88.8**	**17.1**	**3.42**	90.0

(a) interaction pattern.

Next, we analyze behaviors of AT and SE in different segments of interaction patterns with sporadic interruptions. An interaction pattern of 12111111 contains 7 segments including 6 segments from the target and 1 from an interfering speaker. In Fig. 7, we plot objective scores of AT and SE in 6 segments of the target speaker from the beginning to the end. We notice that SE obtains better results than AT in all the segments except for the one before and the one after the interfering talker. Particularly, the performance of SE for the segment before the interfering talker is much worse, implying that it fails in those segments. This establishes that AT is a more robust method than SE and does not fail when presented with speech interruptions.

D. Effects of Speech Onset Differences

The results discussed so far are on test sets in which the onset difference between the first and the second speaker is no smaller than $A = 1$ second. Now, we analyze the behavior of different methods when onset difference is gradually decreased. We plot results for interaction patterns 1221 and 123231 with overlap type *Max* in Fig. 8. The onset difference is gradually decreased from 1 s to 0.25 seconds with a step of 0.25 seconds. We consider two cases of TSE. TSE-1 uses enrollment utterances as specified in the original test set. TSE-2 sets the length of enrollment utterances to the length of onset difference.

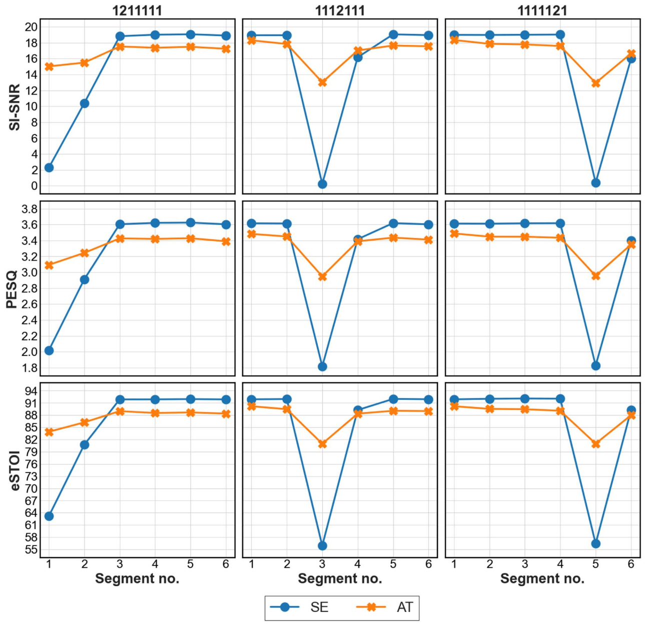


Fig. 7. Comparing AT and SE on 6 segments of the target speaker. Results are plotted for interaction patterns 1211111, 1112111, and 1111121 with overlap type *Max*.

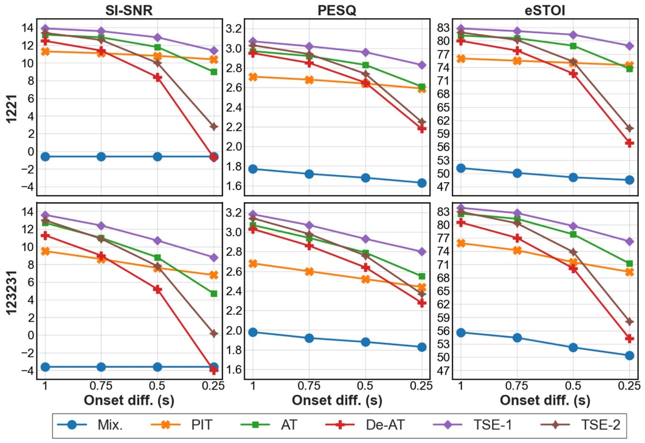


Fig. 8. Performance comparisons with gradually decreasing onset difference.

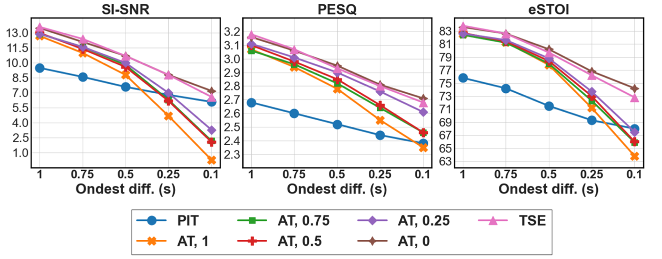


Fig. 9. Comparing ARN trained with AT using a gradually decreasing value of A. AT, a denotes an ARN trained with AT using $A = a$.

TABLE V
PERFORMANCE OF SPEAKER VERIFICATION SYSTEMS TRAINED ON TOP OF THE HIDDEN LAYERS IN THE ARN MODEL

Layer	Raw	Lin-inp	ARN-1	ARN-2	ARN-3	ARN-4
EER (%)	4.5	4.2	3.8	4.7	**3.4**	4.5

We notice that there is a gradual decrease in the performance of all the models as the onset difference is decreased. TSE-2 and De-AT are the most unstable as the performance drops drastically below 0.75 seconds. AT outperforms PIT up to an onset difference of 0.5. The performance of AT drops drastically only for the case of small onset difference of 0.25 s. TSE-1 is the most stable for all the cases. These comparisons indicate that even though AT is sensitive to the onset difference, it generalizes well to smaller onset differences not used during training.

Next, in an attempt to improve the robustness of AT to smaller onset differences, we train ARN with AT using gradually decreasing values of A from $\{1, 0.75, 0.5, 0.25, 0.0\}$ seconds. Note that $A = 0$ does not imply an onset difference of 0, but the minimum allowed onset difference of 0. We plot the performance of these ARN models in Fig. 9 for interaction pattern 123231 and compare it with PIT and TSE (the better-performing TSE-1) plotted in Fig. 8. We see a gradual improvement in the performance with decreasing value of A. Notable, the performance with $A = 0$ matches that of TSE and considerably outperforms PIT. This implies that the robustness of AT to smaller onset differences is easily improved by setting $A = 0$.

E. Speaker Encoding in ARN

The key idea of attentive training is to generate a speaker representation of the first speaker and use it to track this target speaker over the whole mixture. This implies that the hidden layers of the ARN model should have speaker information encoded in them. To investigate this, we present results on training speaker verification models on top of the hidden layers in the pretrained ARN model with frozen parameters. Speaker verification performance in terms of Equal Error Rate (EER) is given in Table V. We observe that training a speaker verification model from raw waveform obtains an EER of 4.5%. The output from the linear layer at the input improves the performance to 4.2%. Layers 2 and 4 do not provide any improvement. However, layers 1 and 3 respectively improve EER to 3.8% and 3.4%, which represents substantial relative EER improvements of 15.6% and 24.4% respectively. This demonstrates that the ARN model is implicitly creating a speaker representation to track the target speaker. We believe that the speaker recognition performance would be even better if we utilized an ARN trained with $A = 0$ instead of $A = 1$.

VII. CONCLUDING REMARKS

We have proposed a novel attentive training framework for speech enhancement. The key idea of attentive training is to attend to a single talker in a given multitalker mixture. Based on the principles of auditory selective attention, attentive training starts attending to (extracting) a speaker based on speech onset and continues attending to it irrespective of other interfering talkers. Attentive training is the first study, to our knowledge, to propose an intrinsic selection mechanism for speaker extraction. We have demonstrated that attentive training has the capability to extend a speech enhancement system to deal with speech interruptions as well as background noises.

We have compared attentive training with different methods of speaker extraction including speaker separation and target speaker extraction. Attentive training is found to be far better than PIT-based speaker separation, which does not have a speaker selection mechanism. Attentive training is competitive with target speaker extraction, which exploits cueing in the form of an enrollment utterance. We have also shown that an approach of decoupling attentive training into speaker selection and tracking obtains similar or worse results than end-to-end training.

Additionally, we have established the importance of attentive training for speech enhancement. We have shown that, when presented with speech interruptions, a speech enhancement system fails during these interruptions. An attentively trained model is found to be far more stable and performs enhancement well during interruptions.

Further, attentive training generalizes to untrained shorter onset differences. For example, a model trained with onset differences of more than 1 s generalizes well to an onset difference of 0.5 seconds. We have also verified that some of the hidden layers of the employed ARN model encode speaker information used for speaker tracking.

We plan to utilize attentive training to train a speech enhancement model to remove interfering speech only from the overlapping intervals instead of tracking the first speaker. Future research also includes investigating attentive training for speaker diarization and separation.

REFERENCES

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 2nd ed. Boca Raton, FL, USA: CRC Press, 2013.

[2] D. 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.

[3] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2013, pp. 436–440.

[4] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 23, no. 1, pp. 7–19, Jan. 2015.

[5] F. Weninger et al., "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *Proc. 12th Int. Conf. Latent Variable Anal. Signal Separation*, 2015, pp. 91–99.

[6] J. Chen, Y. Wang, S. E. Yoho, D. L. Wang, and E. W. Healy, "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises," *J. Acoust. Soc. Amer.*, vol. 139, no. 5, pp. 2604–2612, May 2016.

[7] S.-W. Fu, Y. Tsao, and X. Lu, "SNR-aware convolutional neural network modeling for speech enhancement," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2016, pp. 3768–3772.

[8] S. R. Park and J. Lee, "A fully convolutional neural network for speech enhancement," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2017, pp. 1993–1997.

[9] J. Chen and D. L. Wang, "Long short-term memory for speaker generalization in supervised speech separation," *J. Acoust. Soc. Amer.*, vol. 141, no. 6, pp. 4705–4714, Jun. 2017.

[10] K. Tan, J. Chen, and D. Wang, "Gated residual networks with dilated convolutions for supervised speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 21–25.

[11] A. Pandey and D. L. Wang, "On adversarial training and loss functions for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 5414–5418.

[12] D. S. Williamson, Y. Wang, and D. L. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 24, no. 3, pp. 483–492, Mar. 2016.

[13] H.-S. Choi, J.-H. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee, "Phase-aware speech enhancement with deep complex U-Net," in *Proc. Int. Conf. Learn. Representations*, 2019.

[14] Y. Hu et al., "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 2472–2476.

[15] L. Zhou, Y. Gao, Z. Wang, J. Li, and W. Zhang, "Complex spectral mapping with attention based convolution recurrent neural network for speech enhancement," 2021, *arXiv:2104.05267*.

[16] S.-W. Fu, T.-y. Hu, Y. Tsao, and X. Lu, "Complex spectrogram enhancement by convolutional neural network with multi-metrics learning," in *Proc. IEEE 27th Int. Workshop Mach. Learn. Signal Process.*, 2017, pp. 1–6.

[17] A. Pandey and D. Wang, "Exploring deep complex networks for complex spectrogram enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 6885–6889.

[18] K. Tan and D. Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 380–390, 2020.

[19] A. Pandey and D. L. Wang, "Learning complex spectral mapping for speech enhancement with improved cross-corpus generalization," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 4511–4515.

[20] G. Yu, A. Li, H. Wang, Y. Wang, Y. Ke, and C. Zheng, "DBT-Net: Dual-branch federative magnitude and phase estimation with attention-in-attention transformer for monaural speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 2629–2644, 2022.

[21] S.-W. Fu, Y. Tsao, X. Lu, and H. Kawai, "Raw waveform-based speech enhancement by fully convolutional networks," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2017, pp. 006–012.

[22] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech enhancement generative adversarial network," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2017, pp. 3642–3646.

[23] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 5069–5073.

[24] K. Qian, Y. Zhang, S. Chang, X. Yang, D. Florêncio, and M. Hasegawa-Johnson, "Speech enhancement using Bayesian wavenet," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2017, pp. 2013–2017.

[25] S.-W. Fu, T.-W. Wang, Y. Tsao, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 9, pp. 1570–1584, Sep. 2018.

[26] A. Pandey and D. Wang, "A new framework for CNN-based speech enhancement in the time domain," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 27, no. 7, pp. 1179–1188, Jul. 2019.

[27] A. Pandey and D. Wang, "TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 6875–6879.

[28] K. Wang, B. He, and W.-P. Zhu, "TSTNN: Two-stage transformer based neural network for speech enhancement in the time domain," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 7098–7102.

[29] A. Pandey and D. Wang, "Self-attending RNN for speech enhancement to improve cross-corpus generalization," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 1374–1385, 2022.

[30] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *J. Acoust. Soc. Amer.*, vol. 25, no. 5, pp. 975–979, 1953.

[31] D. Broadbend, *Perception and Communication*. New York, NY, USA: Pergamon Press, 1958.

[32] 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.

[33] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 10, pp. 1901–1913, Oct. 2017.

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

[35] Y. Liu and D. Wang, "Divide and conquer: A deep CASA approach to talker-independent monaural speaker separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 12, pp. 2092–2102, Dec. 2019.

[36] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

[37] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 46–50.

[38] J. Chen, Q. Mao, and D. Liu, "Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation," in *Proc. Interspeech*, 2020, pp. 2642–2646.

[39] N. Zeghidour and D. Grangier, "Wavesplit: End-to-end speech separation by speaker clustering," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 2840–2849, 2021.

[40] M. Delcroix, K. Zmolikova, K. Kinoshita, A. Ogawa, and T. Nakatani, "Single channel target speaker extraction and recognition with speaker beam," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 5554–5558.

[41] Q. Wang et al., "Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking," in *Proc. INTERSPEECH*, 2019, pp. 2728–2732.

[42] C. Xu, W. Rao, E. S. Chng, and H. Li, "SpEx: Multi-scale time domain speaker extraction network," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1370–1384, 2020.

[43] T. Li, Q. Lin, Y. Bao, and M. Li, "Atss-Net: Target speaker separation via attention-based neural network," in *Proc. Interspeech*, 2020, pp. 1411–1415.

[44] Z. Zhang, B. He, and Z. Zhang, "X-tasnet: Robust and accurate time-domain speaker extraction network," in *Proc. Interspeech*, 2020, pp. 1421–1425.

[45] W. Wang, C. Xu, M. Ge, and H. Li, "Neural speaker extraction with speaker-speech cross-attention network," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2021, pp. 3535–3539.

[46] A. Ephrat et al., "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *ACM Trans. Graph.*, vol. 34, 2018, pp. 1–11.

[47] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," in *Proc. Interspeech*, 2018, pp. 3244–3248.

[48] C. Li and Y. Qian, "Listen, watch and understand at the cocktail party: Audio-visual-contextual speech separation," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 1426–1430.

[49] R. Gu et al., "Neural spatial filter: Target speaker speech separation assisted with directional information.," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 4290–4294.

[50] A. Brendel, T. Haubner, and W. Kellermann, "A unified probabilistic view on spatially informed source separation and extraction based on independent vector analysis," *IEEE Trans. Signal Process.*, vol. 68, pp. 3545–3558, 2020.

[51] M. Delcroix, K. Zmolikova, T. Ochiai, K. Kinoshita, and T. Nakatani, "Speaker activity driven neural speech extraction," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 6099–6103.

[52] Y. Hao, J. Xu, P. Zhang, and B. Xu, "WASE: Learning when to attend for speaker extraction in cocktail party environments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 6104–6108.

[53] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA, USA: MIT Press, 1994.

[54] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cogn. Psychol.*, vol. 12, no. 1, pp. 97–136, 1980.

[55] N. Kanda, Y. Gaur, X. Wang, Z. Meng, and T. Yoshioka, "Serialized output training for end-to-end overlapped speech recognition," in *Proc. Interspeech*, 2020, pp. 2797–2801.

[56] T. v. Neumann, K. Kinoshita, M. Delcroix, S. Araki, T. Nakatani, and R. Haeb-Umbach, "All-neural online source separation, counting, and diarization for meeting analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 91–95.

[57] A. Pandey and D. L. Wang, "Attentive training: A new training framework for talker-independent speaker extraction," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2022, pp. 201–205.

[58] Z. Chen et al., "Continuous speech separation: Dataset and analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 7284–7288.

[59] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 3830–3834.

[60] M. Ravanelli et al., "SpeechBrain: A general-purpose speech toolkit," 2021, *arXiv:2106.04624*.

[61] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 5206–5210.

[62] G. Wichern et al., "WHAM!: Extending speech separation to noisy environments," in *Proc. Interspeech*, 2019, pp. 1368–1372.

[63] E. M. Grimm, R. Van Everdingen, and M. J. L. C. Schöpping, "Toward a recommendation for a European standard of peak and LKFS loudness levels," *SMPTE Motion Imag. J.*, vol. 119, no. 3, pp. 28–34, Apr. 2010.

[64] D. Kingma and J. Ba, "ADAM: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015.

[65] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2017, pp. 2616–2620.

[66] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2018.

[67] D. S. Park et al., "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 2613–2617.

[68] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4690–4699.

[69] X. Xiang, S. Wang, H. Huang, Y. Qian, and K. Yu, "Margin matters: Towards more discriminative deep neural network embeddings for speaker recognition," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2019, pp. 1652–1656.

[70] L. N. Smith, "Cyclical learning rates for training neural networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2017, pp. 464–472.

[71] A. Paszke et al., "Automatic differentiation in PyTorch," 2017.

[72] P. Micikevicius et al., "Mixed precision training," in *Proc. Int. Conf. Learn. Representations*, 2018.

[73] 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.

[74] 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.

[75] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, "SpEx+: A complete time domain speaker extraction network," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 1406–1410.