Target Speaker Verification With Selective Auditory Attention for Single and Multi-Talker Speech

Chenglin Xu, *Member, IEEE*, Wei Rao, *Member, IEEE*, Jibin Wu, *Member, IEEE*, and Haizhou Li, *Fellow, IEEE*

***Abstract*—Speaker verification has been studied mostly under the single-talker condition. It is adversely affected in the presence of interference speakers. Inspired by the study on target speaker extraction, e.g., SpEx, we propose a unified speaker verification framework for both single- and multi-talker speech, that is able to pay selective auditory attention to the target speaker. This target speaker verification (tSV) framework jointly optimizes a speaker attention module and a speaker representation module via multi-task learning. We study four different target speaker embedding schemes under the tSV framework. The experimental results show that all four target speaker embedding schemes significantly outperform other competitive solutions for multi-talker speech. Notably, the best tSV speaker embedding scheme achieves 76.0% and 55.3% relative improvements over the baseline system on the WSJ0-2mix-extr and Libri2Mix corpora in terms of equal-error-rate for 2-talker speech, while the performance of tSV for single-talker speech is on par with that of traditional speaker verification system, that is trained and evaluated under the same single-talker condition.**

***Index Terms*—Target speaker verification, speaker extraction, single- and multi-talker speaker verification.**

I. Introduction

TRADITIONAL speaker verification (SV) methods, such as i-vector [1]–[3] with probabilistic linear discriminant analysis (PLDA) [4], x-vector PLDA [5]–[7], assume that input speech is uttered by a single speaker. These methods, however, degrade significantly in the presence of interference speakers. Speaker diarization technique seeks to inform 'who spoke when?' It segments the multi-talker speech temporally into speaker turns, and identifies speaker-overlapping segments [7]–[12]. By doing so, one is able to exclude speaker-overlapping segments from speaker verification [13], [14]. Along the same line of thought, the recent studies on target-speaker voice activity detection (VAD) show that we are able to obtain the target speaker's boundary in a multi-talker speech, e.g. personal VAD [15], target VAD [16]. In general, the speaker diarization technique is helpful only if the speakers overlap sporadically, while it fails when the speakers are heavily overlapped in time.

From the time-frequency analysis point of view, multi-talker speech can be considered as multiple single-talker speech samples overlapping in both temporal and spectral dimensions. To recover a single-talker speech sample, the monaural speech separation techniques could come in handy. Successful implementations include deep clustering [17], deep attractor network [18], permutation invariant training [19]–[21], Conv-TasNet [22], DPRNN [23]. However, speech separation technique seeks to recover the single-talker speech for each individual, that is not only an overkill for speaker verification, but also difficult particularly when we don't know the number of speakers in the multi-talker speech.

For speaker verification, we are only interested in the presence or absence of the target speaker. Speaker extraction technique, a variant of speech separation, that aims to extract one target speaker at a time, is clearly more relevant to the speaker verification task. The question is how to optimize the speaker extraction algorithm so as to better serve the purpose of speaker verification. The speaker extraction techniques typically rely on a reference speech to direct the selective auditory attention to the target speaker in the observed speech [24]. For brevity, the selective auditory attention mechanism is referred to as speaker attention hereafter. Many successful speaker extraction techniques are proposed recently, instances include SpeakerBeam [25], SBF-MTSAL [26], SBF-MTSAL-Concat [26], Voicefilter [27], DENet [28], SpEx [24], and SpEx+ [29]. In speaker verification, a reference speech of the target speaker is always available, that is required by the enrollment process, thus also called enrollment utterance. Such reference speech can be readily used to direct speaker attention.

The idea of speaker extraction (SE) followed by speaker verification, i.e., SE-SV [30] pipeline, was previously studied to address speaker verification for multi-talker speech. The SE-SV system extracts the speech of the target speaker in the first stage, and subsequently processes the extracted speech with a standard speaker verification module, such as i-vector PLDA [1]–[3]. Unfortunately, within the SE-SV framework, the speaker extraction front-end and the speaker verification back-end are optimized separately, leading to a potential mismatch between the two

Manuscript received January 10, 2021; revised June 17, 2021; accepted July 23, 2021. Date of publication July 29, 2021; date of current version August 19, 2021. This work was supported in part by the National Research Foundation, Singapore under its AI Singapore Programme under AISG Award AISG-100E-2018-006, in part by the Science and Engineering Research Council, Agency of Science, Technology and Research, Singapore, through the National Robotics Program under Human-Robot Interaction Phase 1 under Grant 192 25 00054. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. A. Abad. *(Corresponding author: Wei Rao.)*

Chenglin Xu, Jibin Wu, and Haizhou Li are with the Department of Electrical, and Computer Engineering, National University of Singapore, Singapore 119077, Singapore (e-mail: xuchenglin28@gmail.com; jibin.wu@u.nus.edu; haizhou.li@nus.edu.sg).

Wei Rao is with the Tencent Ethereal Audio Lab, Shenzhen 518057, China (e-mail: ellenweiwei@163.com).

Digital Object Identifier 10.1109/TASLP.2021.3100682

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

modules. Furthermore, the SE-SV system is designed for overlapped multi-talker speech input, that presents another mismatch when presented with single-talker speech input. In this paper, we propose an end-to-end neural network architecture for target speaker verification (tSV), which can effectively overcome the aforementioned mismatches.

Specifically, the proposed tSV system consists of three main components: a speaker attention module, a speaker representation module, and a speaker verification decision module. The speaker attention module extracts the target speaker's voice, that is further encoded by the speaker representation module into a discriminative speaker embedding for effective speaker verification. There have been studies on joint optimization between speech enhancement and speaker verification [31]–[33]. Along a similar line of thought, we propose to jointly optimize a speaker attention module and a speaker representation module by simultaneously minimizing a signal reconstruction loss and a speaker identity loss. For speaker verification decision, a PLDA [4] classifier is trained on the speaker embeddings derived from the speaker representation module.

In this work, we seek to develop a unified framework for both single- and multi-talker speech inputs, that is fulfilled by exposing the tSV system to both single- and multi-talker input speech during training. This marks a departure from the single-talker assumption in traditional speaker verification systems, such as x-vector [5], d-vector [34], and other speaker representation techniques [35]. In summary, this paper makes the following contributions:

1) We propose a unified network architecture that performs target speaker verification (tSV) for both single- and multi-talker speech.
2) We formulate a multi-task learning algorithm for joint optimization of a speaker attention module and a speaker representation module in a target speaker verification system.
3) We perform comprehensive studies on each individual network modules as well as at the system level. We successfully show that the proposed tSV framework not only works well under the multi-talker condition, but also performs as competitively as traditional SV under the single-talker condition.

The remainder of the paper is organized as follows. In Section II, we introduce the proposed target speaker verification (tSV) system in details. In Section III, we describe the experiments through which we systematically evaluate the proposed tSV framework. In Section IV, we report the experimental results. Finally, we conclude the study in Section V.

II. Unified Target Speaker Verification System for Single and Multi-Talker Speech

As shown in Fig. 1, the proposed tSV system consists of a speaker attention module, a speaker representation module, and a PLDA classifier. During training, let $x(t)$ be a reference speech from the target single talker, and $y(t)$ be the observed speech that could come from either a single talker or multiple talkers. The speaker attention module is trained to extract the

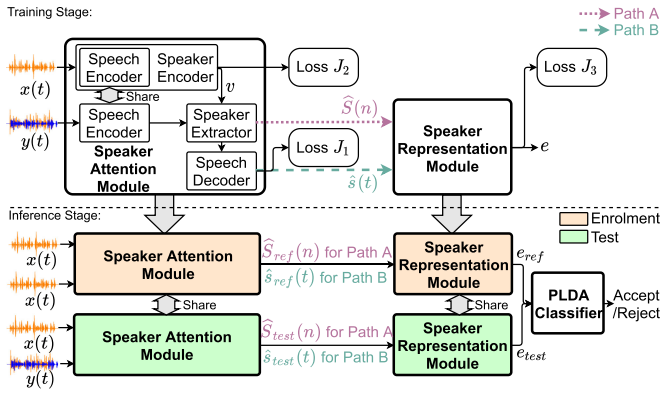


Fig. 1. The proposed tSV system consists of a speaker attention module, a speaker representation module, and a speaker verification decision module (PLDA classifier). $x(t)$ and $y(t)$ denote the single-talker reference speech and the observed speech during the training stage, whereas they are also used to denote the enrollment utterance and test utterance during the inference stage. *Path A* and *Path B* represent two output paths of the speaker attention module. "J_1" is a multi-scale scale-invariant signal-to-distortion loss. "J_2" and "J_3" are cross-entropy losses.

enhanced speech that contains only the target speaker from $y(t)$ with the reference of $x(t)$. The extracted speech is taken by the speaker representation module either in the form of an internal representation $\hat{S}(n)$, (*Path A* in Fig. 1) or as a reconstructed speech $\hat{s}(t)$ (*Path B* in Fig. 1), and further transformed into an appropriate speaker embedding e. We train the speaker attention module and the speaker representation module jointly via a multi-task learning algorithm that will be explained in details in Section II-C. Finally, a PLDA classifier is trained on the derived speaker embeddings e to perform speaker verification.

During inference, let $x(t)$ be an enrollment utterance, a.k.a the reference speech used for speaker attention modeling, from the target single talker, and $y(t)$ be the test utterance for the target speaker verification. We leverage the trained speaker attention module and speaker representation module to derive an enrolled speaker embedding e_{ref} for $x(t)$, and a target speaker embedding e_{test} for $y(t)$. It is commonly assumed that the enrollment utterance is pre-recorded from the target single speaker. Therefore, it is not necessary for the enrollment utterance to pass through the speaker attention module. However, to keep the input to the speaker representation module consistent across both training and inference stages, we pass both the enrollment utterance and the test utterance through the shared speaker attention module as illustrated in Fig. 1. This is referred to as the standard inference configuration.

Under this configuration, $x(t)$ serves as the input reference speech, and at the same time, the input observed speech to obtain an internal representation $\hat{s}_{ref}(t)$ or $\hat{S}_{ref}(n)$ for the enrolled target speaker. Besides, $x(t)$ is also served as the reference speech to extract a target speaker representation $\hat{s}_{test}(t)$ or $\hat{S}_{test}(n)$ for the test utterance $y(t)$. In the case where the target speaker is absent from $y(t)$, no speech content of the target speaker is expected in $\hat{s}_{test}(t)$ or $\hat{S}_{test}(n)$.

The speaker representation module then encodes $\hat{s}_{ref}(t)$ and $\hat{s}_{test}(t)$ (*Path B*), or $\hat{S}_{ref}(n)$ and $\hat{S}_{test}(n)$ (*Path A*), into speaker embeddings e_{ref} and e_{test}, respectively. e_{ref} encodes

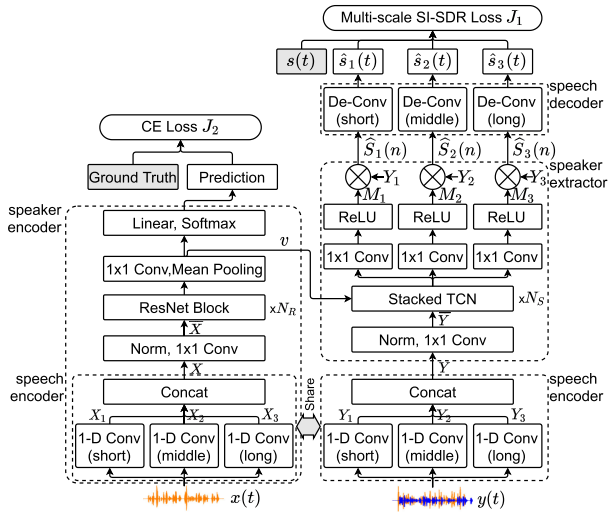


Fig. 2. Illustration of the speaker attention module that extracts the target speaker's voice from $y(t)$ with respect to the reference speech $x(t)$. The outputs are represented either in the form of an internal representation $\hat{S}(n)$ or as a reconstructed speech $\hat{s}(t)$. "De-Conv" is the de-convolution operation to reconstruct the time-domain signal.

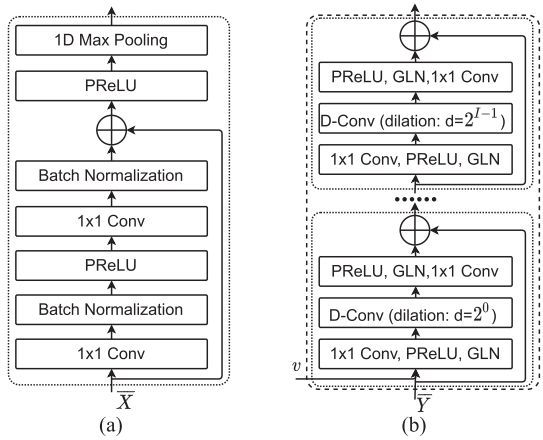


Fig. 3. (a) The diagram of a ResNet block. "1×1 Conv" is a 1-dimensional convolution layer with the kernel size of 1×1. "$\bar{X}$" is the normalized embedding coefficients of the reference speech $x(t)$. (b) The diagram of a stacked TCN module that has I blocks, each of which has a dilated depthwise convolution layer "D-Conv" with dilation ratio of 2^{i-1}, $i = 1, \ldots I$. "GLN" refers to the global layer normalization. The target speaker representation v is only presented at the first block. "$\bar{Y}$" is the normalized embedding coefficients of the observed speech $y(t)$.

the enrolled target speaker, and e_{test} encodes the extracted target speaker as long as he/she is present in $y(t)$. Finally, the PLDA classifier compares e_{ref} and e_{test} to accept or reject the speaker identity claim.

A. Speaker Attention Module

The speaker attention module in Fig. 1 consists of a shared *speech encoder*, a *speaker encoder*, a *speaker extractor*, and a *speech decoder* [24]. As a more detailed illustration shown in Fig. 2, the *speech encoder* encodes the observed speech $y(t)$, either single-talker or multi-talker, into spectrum-like embedding coefficients. Built on top of the shared *speech encoder*, the *speaker encoder* encodes the reference speech $x(t)$ into an utterance-level latent representation v that represents the target speaker. It guides the *speaker extractor* to estimates an auditory mask for the target speaker, which only lets pass the target speaker's voice. Finally, the *speech decoder* reconstructs the time-domain speech signal from the modulated embedding coefficients of the observed speech.

1) Speech Encoder: Inspired by the multi-scale time domain speech encoder introduced in [24], we use three parallel 1-D convolutional layers, each has a different temporal resolution, to encode the speech inputs $y(t) \in \mathbb{R}^{1\times T_1}$ and $x(t) \in \mathbb{R}^{1\times T_2}$. Each of these convolutional layers has N filters but with different kernel sizes of $L_1(short), L_2(middle), L_3(long)$, respectively. The rectified linear unit (ReLU) activation function is used in these layers to produce non-negative embedding coefficients as the following

$$Y_i = \text{ReLU}(y * U_i), \quad i = 1, 2, 3 \tag{1}$$

$$X_i = \text{ReLU}(x * U_i), \quad i = 1, 2, 3 \tag{2}$$

where $*$ denotes the 1-D convolutional operator. $U_i \in \mathbb{R}^{N\times L_i}$ refers to the N convolutional filters that has a kernel size of L_i each. These filters are applied to the speech inputs at a fixed stride of $L_1/2$ samples to allow easy concatenation of the filter outputs at different temporal resolutions. $Y_i \in \mathbb{R}^{N\times K_1}$ and $X_i \in \mathbb{R}^{N\times K_2}$ are the encoded embedding coefficients, wherein $K_1 = 2(T_1 - L_1)/L_1 + 1$ and $K_2 = 2(T_2 - L_1)/L_1 + 1$ are the number of output frames. The outputs from these three parallel convolution layers are aligned and concatenated into multi-scale embedding coefficients $Y \in \mathbb{R}^{3N\times K_1}$ or $X \in \mathbb{R}^{3N\times K_2}$.

2) Speaker Encoder: Following SpEx+ [29], we introduce a speaker encoder to encodes the reference speech $x(t)$ into an utterance-level representation v, which can characterize the voiceprint of the target speaker. Built on top of the shared speech encoder, a 1-D convolutional layer with a kernel size of 1×1, that is called a 1×1 Conv, is applied to the normalized (channel-wise) embedding coefficients, followed by N_R identical residual network (ResNet) blocks to progressively modify the representation. The details of each residual block are provided in Fig. 3(a). Finally, a 1×1 Conv layer together with a mean pooling is used to transform the frame-level feature representations into a fixed dimensional utterance-level representation $v \in \mathbb{R}^{D\times 1}$. To summarise, the speaker encoder can be expressed as a function $g(\cdot)$ of the input speech $x(t)$,

$$v = g(x(t)) \tag{3}$$

To train this speaker encoder, we attach an output layer to classify the speaker identity of each utterance, which is trained jointly with other modules within a multi-task learning framework as will be explained in Section II-C. It worth mentioning that the output layer and the loss function are not required during inference.

3) Speaker Extractor: The speaker extractor estimates a selective filter (i.e., mask), that only lets pass speech content related to the target speaker while masks off that of other interference speakers. To this end, the multi-scale embedding coefficients Y

of the observed speech is first normalized channel-wise before being filtered by a 1×1 Conv layer with O filters. Then, stacked temporal convolutional network (TCN) modules are repeated N_S times to transform the input embedding into an effective representation for mask estimation. The details of one such stacked TCN module are provided in Fig. 3(b), wherein I dilated depthwise separable convolution layers are stacked with dilation factors of $2^{(i-1)}, i = 1, 2, \ldots, I$. The depthwise convolution with P filters and a kernel size of $1 \times Q$ is coupled with a 1×1 Conv layer with O fiters and global layer normalization ("GLN") to keep the total number of parameters manageable.

To achieve target speaker extraction, the mask estimation is conditioned on the target speaker representation v derived from the speaker encoder during both training and inference. Specifically, the representation v is repeated and concatenated to the normalized embedding coefficients $\bar{Y}$ as the inputs $\hat{Y} \in \mathbb{R}^{(O+D)\times K_1}$ to each stacked TCN module. Next, three parallel 1×1 convolutions layers with N filters each are used to estimate mask $M_i \in \mathbb{R}^{N\times K_1}, i = 1, 2, 3$, one for each encoded embedding coefficient $Y_i \in \mathbb{R}^{N\times K_1}$.

Finally, the modulated embedding coefficients $\hat{S}_i \in \mathbb{R}^{N\times K_1}$ of the target speaker are obtained, for different scale $i = 1, 2, 3$, by applying the estimated mask M_i onto the corresponding embedding coefficients Y_i,

$$\begin{aligned} \hat{S}_i &= M_i \otimes Y_i \\ &= f(Y, v) \otimes Y_i \end{aligned} \tag{4}$$

where $\otimes$ denotes the element-wise multiplication, and $f(\cdot)$ represents the operations of the speaker extractor. $\hat{S}_i$ represents the temporal sequence of $\hat{S}_i(n), n \in [1, K_1]$.

4) Speech Decoder: The speech decoder reconstructs the time-domain signal $\hat{s}_i(t)$ from the modulated embedding coefficients $\hat{S}_i$ through a de-convolutional layer that performs 1-D transposed convolution operations as follows,

$$\hat{s}_i(t) = \hat{S}_i * V_i, \quad i = 1, 2, 3 \tag{5}$$

where $V_i \in \mathbb{R}^{N\times L_i}$ is the decoder basis.

B. Speaker Representation Module

Speaker representation module seeks to encode the speaker characteristics of an utterance of variable duration into a fixed-length vector, that is called speaker embedding, for example, x-vector [5]. The speaker representation with a fixed-length vector greatly facilitates the speaker comparison. In this paper, the speaker embedding, denoted as e, is derived from a speaker attention-representation pipeline, therefore, it is called target-speaker-vector, or ts-vector for short. It should be noted that the speaker embedding e, as the output of the speaker representation module, is derived for speaker comparison. Here, e is not to be confused with the reference utterance latent representation v in 3, that is derived for the speaker attention purpose.

As introduced in the earlier section, the speaker attention module is trained to produce time-domain speech signal for the target speaker. As reported in [24], for speaker attention module with multi-scale speech encoding, the quality of the reconstructed speech signal $\hat{s}_1(t)$ with a high temporal resolution (a small kernel size) outperforms $\hat{s}_2(t)$ and $\hat{s}_3(t)$ with middle and low temporal resolutions, as shown in Fig. 2. Therefore, $\hat{s}_1(t)$ from the speaker attention module is chosen in this paper as the time-domain input to the speaker representation module (*Path B* in Fig. 1). Meanwhile, we also study the use of modulated embedding coefficients $\hat{S}(n)$ (*Path A* in Fig. 1) as the input to the speaker representation module. We next discuss four different ts-vector speaker embedding schemes derived from them.

1) Speaker Embedding With Embedding Coefficients $\hat{S}(n)$: As shown in Fig. 4(a), following a similar network architecture as the speaker encoder introduced in Section II-A2, we design a network to transform the temporal sequence $\hat{S}(n)$ into a speaker embedding. To ensure a fixed dimensional output from the network, we use a mean pooling layer to read out the target speaker embedding e, or ts-vector. We refer to the resulting speaker embedding e as ts-vector-R, and the tSV system as tSV-R hereafter. Since the temporal sequences $\hat{S}_1(n)$, $\hat{S}_2(n)$ and $\hat{S}_3(n)$ are directly taken by the speaker representation module, the speech decoder of the speaker attention module is not involved during run-time inference.

2) Speaker Embedding With Time-Domain Encoder: We now take the best reconstructed signal $\hat{s}_1(t)$ from *Path B* of the speaker attention module as the input to the speaker representation module. As the reconstructed speech $\hat{s}_1(t)$ is a time-domain signal, the time-domain processing is required to encode the speech signal. For fair comparison with $\hat{S}(n)$ from *Path A* that spans across multiple time scales, $\hat{s}_1(t)$ is first encoded by a time-domain encoder with the same temporal resolutions to $\tilde{S}(n)$.

For simplicity, we adopt the same speaker encoder as introduced in Section II-A2, which consists of a trainable speech encoder front-end. As illustrated in Fig. 4(b), the encoded multi-scale representation is further transformed into a speaker embedding using the same network structure as adopted for tSV-R. We denote the resulting speaker embedding e as ts-vector-T, and the tSV system as tSV-T hereafter. With such a trainable time-domain speech encoder, we would like to compare tSV-T with tSV-R, and other frequency-domain speech encoders in the experiments. This speaker embedding scheme is also introduced for the study of an alternative inference workflow in Fig. 6, as will be discussed in Section III-E and IV-E.

3) Speaker Embedding With Frequency-Domain Encoder: We further use $\hat{s}_1(t)$ from *Path B* of the speaker attention module as the input but in a different way. As illustrated in Fig. 4(c), instead of using a trainable time-domain speech encoder, we study the use of frequency domain processing via the short-time Fourier transform (STFT) analysis.

Comparing to the tSV-T shown in Fig. 4(b), we note that, besides the difference in speech encoder front-end, the frequency-domain speaker representation module shares the same architecture with that of tSV-T. The resulting speaker embedding e is referred to as ts-vector-F, and the tSV system as tSV-F hereafter.

4) Speaker Embedding With Frequency-Domain Attention: For frequency-domain SV, i.e., i-vector PLDA [1]–[3], the static and dynamic features are shown to improve the effectiveness of speaker embeddings, for example, mel-frequency cepstral coefficients (MFCC) with its delta and acceleration [36]. In

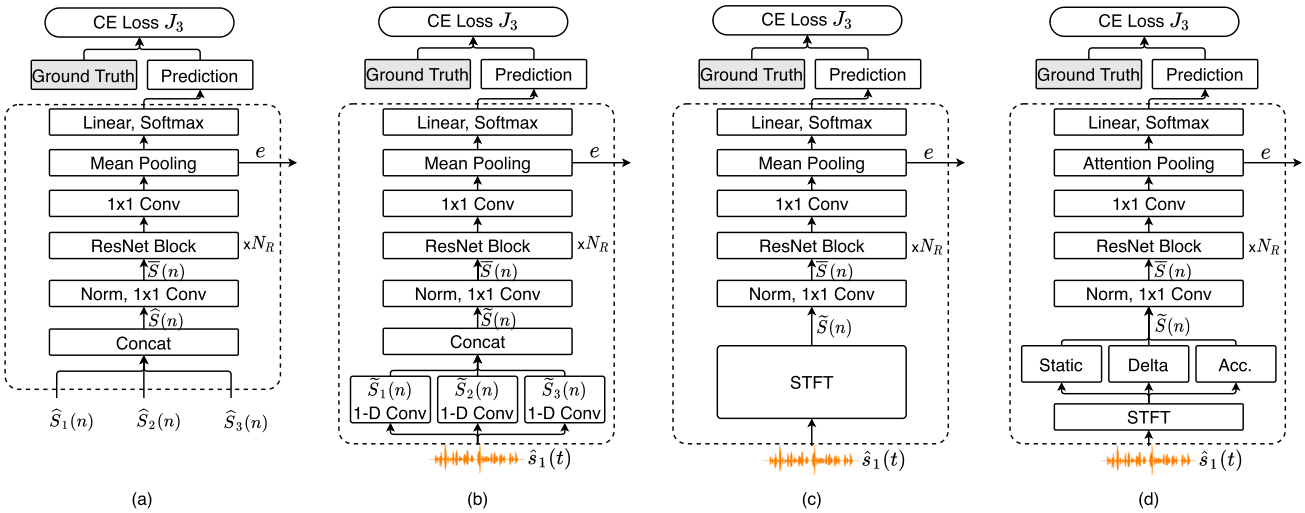


Fig. 4. The implementation of the speaker representation module with four speaker embedding schemes, where tSV-R follows *Path A* of tSV system, while tSV-T, tSV-F, and tSV-FA follow *Path B* of tSV system. (a) tSV-R with ts-vector-R; (b) tSV-T with ts-vector-T; (c) tSV-F with ts-vector-F; (d) tSV-FA with ts-vector-FA.

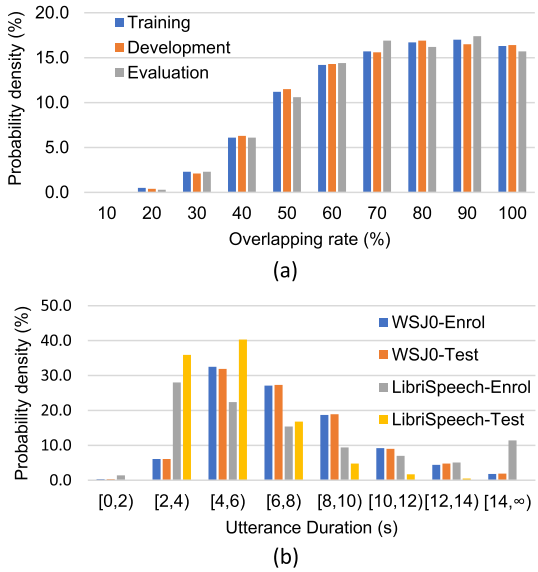


Fig. 5. (a) The overlapping rate distribution of 2-talker mixture on training, development, and evaluation sets of the WSJ0-2mix-extr corpus. (b) The duration distribution of the evaluation set for both enrol and test utterances.

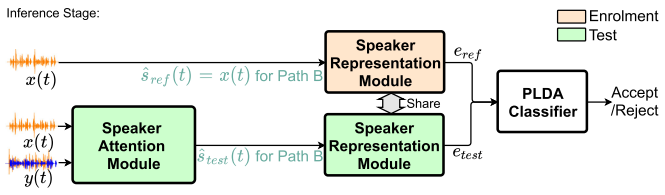


Fig. 6. An alternative inference flow-chart for the tSV system with *Path B* in Fig. 1, where the enrollment utterance is directly taken by speaker representation module, bypassing the speaker attention module.

addition, when deriving an utterance-level speaker embedding from a frame-level representation, studies have shown the superiority of using the attention mechanism to replace the mean pooling [37], [38]. Therefore, we extend the frequency-domain speaker representation module tSV-F by 1) including delta and acceleration of the magnitude spectrum as additional features, and 2) replacing the mean pooling layer with an attentive statistic pooling layer. The details of the proposed model are illustrated in Fig. 4(d). We refer to the resulting speaker embedding e as ts-vector-FA, and the tSV system as tSV-FA hereafter.

To train these speaker representation modules, we leverage a speaker classification task that forms part of the multi-task learning framework as introduced in the following section.

C. Multi-Task Learning

Speaker attention module primarily seeks to enhance the perceptual quality of the target speaker, which does not directly contribute to the improvement of speaker discrimination at the speaker representation module. Therefore, to unify their contributions, we propose to jointly optimize the speaker attention module and the speaker representation module via a multi-task learning framework. In the following, we will explain how multiple loss functions are contributed individually as well as at the system level to improve the speaker discrimination.

As discussed in Section II-A, the main objective of the speaker attention module is to extract target speaker's voice with a high perceptual quality. This can be achieved by optimizing a multi-scale scale-invariant signal-to-distortion ratio (SI-SDR) loss [39], denoted as J_1.

$$J_1 = -[(1-\alpha-\beta)\rho(\hat{s}_1, s) + \alpha\rho(\hat{s}_2, s) + \beta\rho(\hat{s}_3, s)] \quad (6)$$

where α and β are tunable hyperparameters to adjust the contributions at different temporal scales. $\hat{s}_1$, $\hat{s}_2$ and $\hat{s}_3$ are the reconstructed signals from modulated embedding coefficients $\hat{S}_1$, $\hat{S}_2$ and $\hat{S}_3$, respectively. s is the target clean signal. The SI-SDR loss [39], denoted as $\rho(\cdot,\cdot)$,[1] measures the error between

[1]We may add a small value to the numerator and denominator of the loss function to handle situations where the target speaker is absent from the input speech mixture during training [40].

the reconstructed and the target clean signals.

$$\rho(\hat{s}, s) = 10\log_{10}\left(\frac{||\frac{\langle \hat{s}, s\rangle}{\langle s, s\rangle}s||^2}{||\frac{\langle \hat{s}, s\rangle}{\langle s, s\rangle}s - \hat{s}||^2}\right) \tag{7}$$

where $\langle \cdot, \cdot \rangle$ indicates the inner product. In order to ensure scale invariance, the signals $\hat{s}$ and s are normalized to zero-mean in the above formulation.

Since speaker representation v is crucial to the success of speaker extraction, we further introduce a speaker classification task [24] to the speaker encoder within the speaker attention module, so as to improve the quality of speaker representation v. Specifically, the cross-entropy loss J_2 is applied to the speaker classification task.

$$J_2 = -\sum_{c=1}^{C} p_c \log(P(c|x)) \tag{8}$$

where C is the total number of speakers in the speaker classification task. p_c is 1 if the segment x belongs to speaker c, otherwise, p_c is 0. $P(c|x)$ is the probability for the segment x to be speaker c. Besides optimized for the speaker classification task through J_2, the speaker encoder is also jointly optimized for speaker extraction via J_1.

The speaker representation module seeks to encode the enrollment utterance and the test utterance into speaker embeddings e_{ref} and e_{test} that are suitable for speaker comparison. To effectively characterize the speakers, the speaker representation module is trained together with the speaker attention module under another speaker classification task with a cross-entropy loss J_3, as shown in Fig. 4,

$$J_3 = -\sum_{c=1}^{C} p_c \log(P(c|\hat{S})) \tag{9}$$

or

$$J_3 = -\sum_{c=1}^{C} p_c \log(P(c|\hat{s}_1)) \tag{10}$$

where $\hat{S} = [\hat{S}_1; \hat{S}_2; \hat{S}_3]$ and $\hat{s}_1$ are the modulated embedding coefficients and the reconstructed signal from the speaker attention module for *Path A* and *Path B*, respectively.

Finally, we jointly optimize the proposed framework with a total loss J, that is the weighted sum of J_1, J_2 and J_3,

$$J = J_1 + \gamma J_2 + \eta J_3 \tag{11}$$

where γ and η are tunable hyperparameters to align different objectives.

D. Target Speaker Verification

The PLDA classifier [4] has been applied widely to many speaker embeddings, such as, i-vector [1]–[3], x-vector [5]–[7], and d-vector [34]. With the four target speaker embedding schemes implemented in the speaker representation module, we are now ready to train a PLDA classifier for target speaker verification.

Specifically, prior to the PLDA classification, the speaker embeddings (ts-vectors) are centered following a linear discriminate analysis (LDA) for dimensionality reduction and a length normalization. To improve the system generalizability, we train the PLDA classifier using the extracted speaker embeddings from the training set that consists of both single talker and multi-talker speech samples. During the run-time inference, the PLDA classifier compares the enrolled target speaker and the extracted target speaker using their respective speaker embeddings, i.e., e_{ref} and e_{test}. Finally, PLDA scores are normalized with an adaptive s-norm [41], and used to accept or reject the speaker identity claim.

III. Experimental Setup

A. Speech Corpora

To evaluate the performance of the proposed target speaker verification system, we conducted experiments on two standard corpora:[2] 1) WSJ0 [42] and its 2-talker mixture version (WSJ0-2mix-extr) [26] with a total number of 119 speakers, and 2) LibriSpeech [43] and its 2-talker mixture version (Libri2Mix) [44] with a total number of 1212 speakers.

1) WSJ0 and WSJ0-2mix-Extr Corpora: WSJ0 corpus [42] consists of read speech from the Wall Street Journal. It has three original collections: "si_tr_s," "si_dt_05," and "si_et_05". In this work, we first generated a single-talker dataset (WSJ0-1talker) by randomly selecting speech from the WSJ0 corpus for single-talker SV. WSJ0-1talker consists of three subsets: training, development, and evaluation. Specifically, 11 560 utterances from a total of 101 speakers (50 male and 51 female speakers) were selected from WSJ0 "si_tr_s" set, and split into the training set (8769 utterances) and the development set (the rest 2791 utterances). The evaluation set includes 1857 enrollment utterances and 1478 test utterances of 18 speakers (10 male and 8 female speakers) from the WSJ0 "si_dt_05" and "si_et_05" sets. From this evaluation set, we further constructed 3000 target trials and 48 000 non-target trials. The speech samples were down-sampled to 8 kHz.

WSJ0-2mix-extr [26] corpus was generated from the WSJ0 Corpus by mixing two randomly selected utterances as a 2-talker dataset (WSJ0-2talker).[3] WSJ0-2talker also consists of training, development, and evaluation sets. Specifically, the training set included 20 000 mixtures that were generated by mixing two randomly selected utterances from the aforementioned 101 speakers in the WSJ0 corpus. Similarly, the development set with 5000 mixtures was also generated from the same 101 speakers. The evaluation set has 3000 mixtures generated from the aforementioned 18 different speakers as the test utterances, which are unseen during training. The signal-to-noise ratio (SNR) between the target speaker and the interference speaker of each mixture was randomly chosen between 0 dB and 5 dB. For the training and evaluation of the speaker attention module, each mixture

[2]All the configurations (including the utterances list for training, development and evaluation) and the code could be found here: https://github.com/xuchenglin28/target_speaker_verification

[3]The WSJ0-2mix-extr corpus simulation code is available at: https://github.com/xuchenglin28/speaker_extraction

has a corresponding reference speech and a single talker target speech. For the multi-talker SV evaluation, 3000 target trials and 48 000 non-target trials were generated from the same 1875 enrollment utterances from WSJ0-1talker dataset and the above 3000 2-talker test utterances.

The mixtures in the WSJ0-2talker were generated according to the maximum duration protocol, where the shorter utterance was appended with zeros to match the duration of the longer utterance. The maximum duration protocol leads to various overlapping rates for the mixtures, as summarized in Fig. 5(a). Most 2-talker speech samples are heavily overlapped. The average duration of enrollment and test utterances is 7.2 s and 7.3 s, respectively. The distribution of utterance duration for the enrollment and evaluation sets is summarized in Fig. 5(b). Most utterances used for both enrollment and test are less than 10.0 seconds.

In both the WSJ0-1talker and WSJ0-2talker, the speakers in the evaluation set are excluded from the training and development sets.

2) LibriSpeech and Libri2Mix Corpora: LibriSpeech [43] is derived from audiobooks that are part of the LibriVox[4] project. Libri2Mix [44] is a clean 2-talker mixture corpus generated from the LibriSpeech corpus by mixing two randomly selected utterances. Libri2Mix[5] consists of several subsets: "train-360" with 50 800 utterances (212 hours, 921 speakers), "train-100" with 13 900 utterances (58 hours, 251 speakers), "dev" with 3000 utterances (11 hours, 40 speakers), and "test" with 3000 utterances (11 hours, 40 speakers).

To study the speaker attention mechanism, each 2-talker mixture speech could be reused in the following way. In the first instance, we take one speaker, e.g., *Speaker A* as the target speaker, and another speaker e.g., *Speaker B* as the interference speaker. By swapping the role of *Speaker A* and *Speaker B*, we can use the same speech sample in the second instance.

We generated a 2-talker dataset (Libri-2talker) with a training set, a development set, and an evaluation set. The training set (127 056 examples, 1172 speakers) and development set (2344 examples, 1172 speakers) are randomly chosen from the "train-360" and "train-100" sets of the Libri2Mix corpus. The evaluation set includes 2260 enrollment utterances from the "test-clean" set of the LibriSpeech corpus and 6000 test utterances from "test" set of the Libri2Mix corpus. With these enrollment and test utterances, 6000 target trials and 114 000 non-target trials were generated to evaluate the performance of proposed system for multi-talker SV.

We also generated a single-talker dataset (Libri-1talker), which consisted of training, development, and evaluation sets. The training set (125 925 utterances) and development set (6628 utterances) were randomly selected from the original "train-clean-100" and "train-clean-360" of the LibriSpeech corpus with a total of 1172 speakers and around 460 hours of clean speech. The evaluation set includes the same 2260 enrollment utterances and 6000 test utterances of target single-talker speech from the "test" set in Libri2Mix corpus with 40 speakers. Similarly, 6000 target trials and 114 000 non-target trials were generated to evaluate the performance of the SV system under the single speaker's condition.

The Libri2Mix was simulated according to the minimum duration protocol, where the longer utterance was cut short to match the duration of the shorter one. The minimum duration protocol leads to roughly 100% overlapping rate. The average duration of enrollment and test utterances is 7.4 s and 5.0 s, respectively. The duration distribution of enrollment and test set is reported in Fig. 5(b). Most utterances for both enrollment and test are less than 8.0 seconds.

B. Configuration of Speaker Attention Module

We adopt the complete time-domain implementation [29] as the speaker attention mechanism for target speaker extraction. The shared speech encoder encodes both the mixture speech and the reference speech into multi-scale embedding coefficients by three parallel 1-D convolutions with N (=256) filters each. The three 1-D convolutions apply different filter lengths of L_1(=2.5 ms), L_2(=10 ms), L_3(=20 ms) for complementary temporal resolutions. The multi-scale embedding coefficients are firstly normalized by their mean and variance with trainable gain and bias on the channel dimension in the speaker extractor. Then the 1×1 convolution linearly transforms the normalized embedding coefficients to have O(=256) channels. Each TCN block has a dilated depthwise convolution with P (=512) filters and a kernel size of $1 \times Q$(=3). $I(=8)$ TCN blocks are stacked into a module and further repeated for $N_S(=4)$ times. The ReLU activation function is adopted to estimate the mask instead of sigmoid by following the recommendation in [24], [29]. The speech decoder reconstructs the modulated embedding coefficients into time-domain speech signals. The configuration of the de-convolution is kept the same as that used in the speech encoder.

Since the extraction pipeline requires the information of the target speaker, the speaker encoder employs same normalization and 1×1 convolution as in speaker extractor. Then the normalized embedding coefficients are used as inputs to the following N_R(=3) ResNet block. The ResNet block has two 1×1 convolutions layers with 256 filters each, and a 1-D max-pooling layer with the kernel size of 1×3. The whole network is trained with random initialization.

C. Configuration of Speaker Representation and Speaker Verification Modules

As shown in Fig. 4, the four speaker representation modules share a similar network architecture except for some variations in the speech encoding front-end. For tSV-R, it directly takes the modulated embedding coefficients $\hat{S}(n)$ from *Path A* as inputs. The tSV-T uses three parallel 1-D convolution blocks that has the same network configurations as the speech encoder in the speaker attention module. Both tSV-F and tSV-FA systems employ the STFT with a hamming window of size 32 ms (=256 samples), and a stride of 16 ms to perform signal analysis. In addition, the delta and acceleration features are calculated with an order of 2 inside the tSV-FA system.

[4]https://librivox.org

[5]The simulation code of Libri2Mix is:https://github.com/JorisCos/LibriMix

These speech encoding front-ends are followed by channel-wise normalization, 1×1 Conv with 256 filters, and $N_R(=3)$ stacked ResNet blocks. Each ResNet block has two 1×1 Conv layers with 256 filters and a 1-D max-pooling layer with the kernel size of 1×3. The resulting output speaker embedding from the mean pooling layer has a dimension of 256. For tSV-FA, the attentive statistic pooling layer is formed by two fully-connected layers that have 500 and 1 neurons, respectively. The ReLU activation function is applied after the first layer. A softmax activation function is applied on the temporal dimension to obtain a normalized weight coefficient for each frame-level feature. The output speaker embedding, with a dimension of 512, is generated by concatenating the weighted mean and standard deviation vectors.

For the speaker verification module, the LDA and Gaussian PLDA models with 100 latent variables are trained on the derived speaker embeddings from the training data.

D. Training Procedure

We organized the training of the proposed tSV system into three stages. At the first stage, following the same configuration in [29], we trained the speaker attention module with 2-talker mixture speech segments that had a fixed duration of 4 s. The module was trained with the weighted loss $J = J_1 + \gamma J_2$, where J_1 was defined in 6 and J_2 was defined in 8. The weights α, β in J_1 and γ were fine-tuned to take the values of 0.1, 0.1, and 10, respectively. The learning rate was initialized at 10^{-3}, and halved whenever the loss was stagnant for 3 consecutive epochs. To ensure the front-end speaker attention module also works well under the single talk condition, we further fine-tuned the module with both the same 2-talker mixtures and the additional single speaker's speech at a reduced learning rate of 10^{-4}.

At the second stage, we froze the speaker attention module and further trained the speaker representation module using the extracted speech from both the 2-talker mixture and the single speaker. A learning rate of 10^{-4} was used at this stage. Lastly, to improve the synergy between the front-end speaker attention module and the back-end speaker embeddding extractor, they were jointly fine-tuned using the weighted loss $J = J_1 + \gamma J_2 + \eta J_3$ at a reduced learning rate of 10^{-5}, where a same value of 10 were used for both γ and η. The Adam algorithm [45] was adopted to optimize the network across all three stages.

E. Alternative Configuration for tSV Inference With Path B

In Fig. 1, to keep the same processing procedures for both e_{ref} and e_{test} during inference, we pass the enrollment utterance and the test utterance through the shared speaker attention module in the same way. We note an alternative implementation to *Path B* is bypassing the speaker attention module and directly feeding the enrollment utterance into the speaker representation module.

In this way, the speaker representation model can generate e_{ref} directly from $x(t)$ instead of using its reconstructed version $\hat{s}_{ref}(t)$, as illustrated in Fig. 6. The resulting speaker embedding e_{ref} and e_{test} can be compared by the PLDA classifier to make the speaker verification decision. However, e_{ref} is derived from the enrollment utterance $x(t)$, while e_{test} is derived from the extracted speech $\hat{s}_{test}(t)$. This marks a mismatch between their representations, which may lead to a sub-optimal decision. To evaluate the effect of this mismatch, we conducted an experiment on this alternative configuration, and reported the results in Section IV-E.

TABLE I
PERFORMANCE OF SPEX+ ON WSJ0-2MIX-EXTR DATABASE UNDER MISMATCH CONDITIONS BETWEEN TRAINING AND EVALUATION

Systems	Training data	Test data	SI-SDR (dB)
SpEx+ I	Mixture	Mixture	18.1
	Mixture	Single	10.9
SpEx+ II	Mixture+Single	Mixture	17.9
	Mixture+Single	Single	56.7

F. Evaluation Metrics

The speaker attention module was evaluated using the SI-SDR [39] loss defined in 7 for both 2-talker mixture and clean single speaker conditions. The speaker verification decision module (i.e., PLDA classifier) was evaluated using equal error-rate (EER), minimum of the normalized detection cost function (DCF) with P_Target = 0.01 (denoted as DCF08 [46]), and DCF with P_Target = 0.001 (denoted as DCF10 [47]), respectively.

IV. EXPERIMENTAL RESULTS

We start by validating the proposed speaker attention module and speaker representation module through experiments. We then evaluate the proposed tSV system with four target speaker embedding schemes on the WSJ0 and WSJ0-2mix-extr corpora in Section IV-A to Section IV-E. We also compare the proposed tSV with other state-of-the-art solutions to multi-talker speaker verification on the WSJ0-2mix-extr corpus in Section IV-F. We further evaluate the proposed tSV system on LibriSpeech and Libri2Mix corpora, that are larger datasets, in Section IV-G. Finally, we evaluate our target speaker embedding scheme across WSJ0-2mix-extr and Libri2Mix corpora in Section IV-H.

A. Speaker Attention to Single- Vs. 2-Talker Speech on WSJ0-1talker and WSJ0-2talker Datasets

While SpEx+ system [29] demonstrates superior target speaker extraction performance when trained and evaluated on multi-talker speech, it remains unknown how such system performs in face of the single talker speech. To understand the impact of mismatch between training and testing, we use the single-talker speech to test the SpEx+ system that trained solely on the 2-talker mixture data. We refer to the resulted system as SpEx+ I. From Table I, we notice that the SpEx+ I system achieves 18.1 dB and 10.9 dB in terms of SI-SDR for 2-talker and single talker speech, respectively. This huge performance gap can be explained by the mismatch of training and testing conditions.

To study how much the multi-condition training, using both single and 2-talker speech data, helps in improving the system generalizability, we fine-tuned the SpEx+ I system using data from both conditions, which we referred to as SpEx+ II. From

TABLE II
A COMPARISON BETWEEN TRADITIONAL SV SPEAKER EMBEDDINGS AND THE SPEAKER EMBEDDINGS BY THE PROPOSED SPEAKER REPRESENTATION MODULE IN FIG. 1 UNDER SINGLE-TALKER CONDITION ON WSJ0-1TALKER DATASET

Systems	EER (%)	DCF08	DCF10
i-vector PLDA [1]	3.00	**0.360**	0.522
x-vector PLDA [5]	5.87	0.692	0.888
SV-T	4.40	0.450	0.630
SV-F	4.37	0.416	0.587
SV-F w/ Δ & $\Delta\Delta$	3.27	0.458	0.662
SV-FA	**2.90**	0.363	**0.517**

Table I, we note the performance on single talker speech improved significantly from 10.9 dB to 56.7 dB, while for 2-talker speech is still remains comparable. Motivated by this result, we apply the multi-condition training protocol to construct a unified tSV system that works for both single- and multi-talker speech in this work.

B. Speaker Representation for SV on WSJ0-1talker Dataset

The speaker representation module in Fig. 1 is designed to work with the speaker attention module for target speaker recognition. Therefore, the speaker embedding process in tSV is different from that in other traditional speaker verification systems where single talker speech is assumed. To justify the effectiveness of proposed speaker representation module, we compare various speaker embedding SV systems, namely, time-domain SV (SV-T), the frequency SV (SV-F) and the frequency-domain attention-based SV (SV-FA), as shown in Fig. 4(b), 4(c) and 4(d) respectively, with the traditional single-talker systems, i.e., x-vector PLDA [5], under the single talker test condition. The traditional SV system with the single-talker assumption is also referred to as the single-talker SV system. In this study, the speaker attention module is not required for our system.

We observe from Table II that 1) SV-T, SV-F, and SV-FA systems consistently outperform the x-vector PLDA system in terms of EER; 2) both temporal dynamic features (delta and acceleration) and attention pooling contribute to the performance improvement of SV-FA by comparing SV-FA with SV-F and SV-F w/ $\Delta\&\Delta\Delta$; 3) SV-FA system achieves the best EER among all systems. The results suggest that the proposed speaker representation module is as competitive as, if not better than, i-vector or x-vector systems for single talker speech. We proceed with our proposed speaker representation module in the subsequent tSV studies.

C. Evaluating Target Speaker Embeddings on WSJ0-2talker Dataset

To investigate the effect of multi-talker speech on single-talker SV, we compare the traditional SV and the proposed tSV with four different target speaker embedding schemes, namely tSV-R, tSV-T, tSV-F, and tSV-FA, on 2-talker mixture evaluation dataset, i.e. WSJ0-2talker.

The system setups and experiment results are summarised in Table III(a). First, we report three zero-effort baselines (System 1, 10, and 17) that follow traditional speaker embedding schemes, SV-T, SV-F, and SV-FA. They are trained on single talker dataset (WSJ0-1talker) and evaluated on 2-talker mixture evaluation set. We further report 3 reference systems (System 9, 16, and 24) that are trained and evaluated both on single-talker speech. As they don't include interference speakers, the speaker attention module is not involved in the study. For multi-talker experiments, System 9 represents the upper-bound performance of tSV-R and tSV-T; Systems 16 and 24 represent the upper-bound performances of tSV-F and tSV-FA respectively.

For the upper-bound reference systems, we observe that the SV-FA system outperforms the SV-T and SV-F systems, benefiting from the dynamic features and attention pooling. Comparing System 1 and 9, 10 and 16, and 17 and 24, we observe that the performance of the single-talker SV systems seriously degrade in the presence of interference speakers.

We also report three competitive baselines (System 2, 11, and 18) that follow the target speaker extraction-verification (TSE-SV) pipeline [30], where speaker extraction and speaker verification modules are trained separately. Between the zero-effort baselines and competitive baselines, in particular, between System 17 and 18, we observe the followings: 1) The target speaker extraction front-end greatly improves the SV performance under multi-talker test condition; 2) Among System 2, 11 and 18, the frequency-domain SV systems (SV-F and SV-FA) appear to be more robust than the time-domain counterpart (SV-T). This suggests that the time-domain SV system is sensitive to the mismatch between speaker verification module trained with single talker speech, and the extracted speech from speaker extraction module.

The difference between the target speaker extraction-verification pipeline [30] and the proposed tSV system mainly lies in their training schemes. The former trains the extraction module and verification module separately, while the latter enables a multi-task joint training between speaker attention and speaker representation modules. To reduce the mismatch of the pipeline system, we train the back-end speaker representation module with extracted speech as System 19. We observe that System 19 outperforms System 18.

We investigate a sharing and unsharing schemes between the speaker encoder in the speaker attention module and the speaker representation module. We find that the unsharing scheme achieves better performance. Therefore, we adopt this unsharing scheme in the subsequent experiments. Next we summarize the results for the four target speaker embedding schemes under multi-talker test condition.

Now we look into 4 groups of experiments for 4 target speaker embedding schemes in Table III(a). Let's examine how the systems perform on 2-talker speech including System 3 (tSV-R), Systems 5 (tSV-T), Systems 12 (tSV-F), and Systems 20 (tSV-FA). We observe the followings: 1) tSV-FA achieves significantly better performance than other three schemes, which deliver comparable results. 2) tSV-FA improves the performance on 2-talker speech by 76.0%, 45.6% and 32.4% relative improvement over the single talker trained SV baseline (System 17) in terms of EER, DCF08 and DCF10, respectively. 3) tSV-FA achieves 24.1%, 23.0% and 20.4% relative improvement over the competitive baseline (System 18) in terms of EER, DCF08 and DCF10, respectively, that is attributed to the joint optimization between the speaker attention and speaker representation

TABLE III

A SUMMARY OF TARGET SPEAKER VERIFICATION EXPERIMENTS ON (A) WSJ0-1TAKER AND WSJ0-2TALKER, (B) LIBRI-1TALKER AND LIBRI-2TALKER DATASETS. "PATH" DENOTES THE OUTPUT PATH OF THE SPEAKER ATTENTION MODULE IN FIG. 1. "SPEECH DATA TYPE" COLUMN INDICATES THE TYPE OF TRAINING AND EVALUATION DATA INVOLVED ('M' FOR MULTI-TALKER, 'S' FOR SINGLE-TALKER). IN "SPEAKER EMBEDDING INPUT" COLUMN, "TRAINING," "TEST," AND "ENROL" INDICATE THE TYPE OF SPEECH REPRESENTATION FOR SPEAKER EMBEDDING, WHERE "s" AND "y" REPRESENT SINGLE- AND 2-TALKER SPEECH; "x" IS THE ORIGINAL SINGLE-TALKER ENROLLMENT SPEECH; "$\hat{S}$," "$\hat{S}_{test}$" AND "$\hat{S}_{ref}$" ARE THE MODULATED EMBEDDING COEFFICIENTS FROM THE SPEAKER ATTENTION MODULE; "$\hat{s}$," "$\hat{s}_{test}$" AND "$\hat{s}_{ref}$" ARE THE RECONSTRUCTED SIGNALS FROM THE SPEAKER ATTENTION MODULE. "BASELINE1" REPRESENTS THE ZERO-EFFORT TEST CASE WHERE SV SYSTEM IS TRAINED WITH SINGLE-TALKER SPEECH AND EVALUATED ON 2-TALKER SPEECH. "BASELINE2" REPRESENTS THE SPEAKER EXTRACTION-VERIFICATION PIPELINE SYSTEMS. "UPPER BOUND" DENOTES THE CASE WHERE SINGLE-TALKER SPEECH IS USED IN BOTH TRAINING AND EVALUATION. "†" REPRESENTS THE ALTERNATIVE ENROLLMENT PROCESS IN FIG. 6. "PARA." INDICATES THE NUMBER OF PARAMETERS DURING INFERENCE

System ID	Architecture	Path	Joint Opt.	Speech Data Type		Speaker Embedding Input			Para. (M)	EER (%)	DCF08	DCF10
				Training	Test	Training	Test	Enrol				
(a) Single- and multi-talker speaker verification experiments on WSJ0-1talker and WSJ0-2talker												
1 (Baseline1)	SV-T	-	-	S	M	s	y	x	1.77	19.37	0.864	0.928
2 (Baseline2)	TSE-SV-T	-	No	M+S	M	s	$\hat{s}_{test}$	x	12.64	16.97	0.815	0.897
3	tSV-R	A	Yes	M+S	M	$\hat{S}$	$\hat{S}_{test}$	$\hat{S}_{ref}$	12.57	6.67	0.580	0.774
4	tSV-R	A	Yes	M+S	S	$\hat{S}$	$\hat{S}_{test}$	$\hat{S}_{ref}$	12.57	4.00	0.371	0.561
5	tSV-T	B	Yes	M+S	M	$\hat{s}$	$\hat{s}_{test}$	$\hat{s}_{ref}$	12.64	6.60	0.582	0.759
6	tSV-T	B	Yes	M+S	S	$\hat{s}$	$\hat{s}_{test}$	$\hat{s}_{ref}$	12.64	4.63	0.424	0.671
7	tSV-T	B	Yes	M+S	M	$\hat{s}$	$\hat{s}_{test}$	$x^{\dagger}$	12.64	62.13	1.000	1.000
8	tSV-T	B	Yes	M+S	S	$\hat{s}$	$\hat{s}_{test}$	$x^{\dagger}$	12.64	77.93	1.000	1.000
9 (Upper Bound)	SV-T	-	-	S	S	s	s	x	1.77	4.40	0.450	0.630
10 (Baseline1)	SV-F	-	-	S	M	s	y	x	1.08	20.07	0.849	0.894
11 (Baseline2)	TSE-SV-F	-	No	M+S	M	s	$\hat{s}_{test}$	x	12.08	10.53	0.686	0.789
12	tSV-F	B	Yes	M+S	M	$\hat{s}$	$\hat{s}_{test}$	$\hat{s}_{ref}$	12.08	6.67	0.573	0.721
13	tSV-F	B	Yes	M+S	S	$\hat{s}$	$\hat{s}_{test}$	$\hat{s}_{ref}$	12.08	3.87	0.352	0.531
14	tSV-F	B	Yes	M+S	M	$\hat{s}$	$\hat{s}_{test}$	$x^{\dagger}$	12.08	6.53	0.570	0.708
15	tSV-F	B	Yes	M+S	S	$\hat{s}$	$\hat{s}_{test}$	$x^{\dagger}$	12.08	4.03	0.380	0.553
16 (Upper Bound)	SV-F	-	-	S	S	s	s	x	1.08	4.37	0.416	0.587
17 (Baseline1)	SV-FA	-	-	S	M	s	y	x	1.33	20.97	0.825	0.893
18 (Baseline2)	TSE-SV-FA	-	No	M+S	M	s	$\hat{s}_{test}$	x	12.19	6.63	0.583	0.759
19	tSV-FA	B	No	M+S	M	$\hat{s}$	$\hat{s}_{test}$	$\hat{s}_{ref}$	12.19	5.27	0.500	0.680
20	tSV-FA	B	Yes	M+S	M	$\hat{s}$	$\hat{s}_{test}$	$\hat{s}_{ref}$	12.19	5.03	0.449	0.604
21	tSV-FA	B	Yes	M+S	S	$\hat{s}$	$\hat{s}_{test}$	$\hat{s}_{ref}$	12.19	2.63	0.325	0.505
22	tSV-FA	B	Yes	M+S	M	$\hat{s}$	$\hat{s}_{test}$	$x^{\dagger}$	12.19	5.13	0.443	0.602
23	tSV-FA	B	Yes	M+S	S	$\hat{s}$	$\hat{s}_{test}$	$x^{\dagger}$	12.19	2.73	0.325	0.492
24 (Upper Bound)	SV-FA	-	-	S	S	s	s	x	1.33	2.90	0.363	0.517
(b) Single- and multi-talker speaker verification experiments on Libri-1talker and Libri-2talker												
25 (Baseline1)	SV-FA	-	-	S	M	s	y	x	1.33	17.62	0.855	0.938
26 (Baseline2)	TSE-SV-FA	-	No	M+S	M	s	$\hat{s}_{test}$	x	12.19	10.80	0.684	0.812
27	tSV-FA	B	No	M+S	M	$\hat{s}$	$\hat{s}_{test}$	$\hat{s}_{ref}$	12.19	8.22	0.539	0.743
28	tSV-FA	B	Yes	M+S	M	$\hat{s}$	$\hat{s}_{test}$	$\hat{s}_{ref}$	12.19	7.88	0.508	0.681
29	tSV-FA	B	Yes	M+S	S	$\hat{s}$	$\hat{s}_{test}$	$\hat{s}_{ref}$	12.19	6.22	0.348	0.675
30	tSV-FA	B	Yes	M+S	M	$\hat{s}$	$\hat{s}_{test}$	$x^{\dagger}$	12.19	7.80	0.514	0.685
31	tSV-FA	B	Yes	M+S	S	$\hat{s}$	$\hat{s}_{test}$	$x^{\dagger}$	12.19	6.33	0.354	0.677
32 (Upper Bound)	SV-FA	-	-	S	S	s	s	x	1.33	5.70	0.293	0.403

modules. 4) tSV with all four target speaker embedding schemes consistently outperform zero-effort baselines and competitive baselines for multi-talker speech.

It is encouraging to see that the proposed tSV systems evaluated on multi-talker speech achieve comparable performance with the upper-bound reference systems evaluated on single speaker speech.

D. Evaluating Target Speaker Embeddings on WSJ0-1talker Dataset

In real world applications, one expects that the same speaker verification system is able to handle single talker and multi-talker speech seamlessly because we don't know in advance whether the speakers would overlap. We have studied the performance of tSV systems for multi-talker speech at different overlapping rate in Section IV-C. It is interesting to know the performance of the proposed tSV systems on single talker's evaluation set, i.e. with zero overlapping rate. The results are summarized as System 4, 6, 13 and 21 in Table III(a).

To evaluate under single talker condition, we adopt the traditional single talker SV system as a baseline. Comparing among Systems 4, 6 and 9, we observe that the tSV-R system (System 4) achieves significant better performance than SV-T system (System 9) in terms of EER, DCF08 and DCF10. While this is a pleasantly surprising result, we consider that the target speaker extraction system in tSV-R does the first-round verification by only extracting the target speaker's voice. This extra step helps to decline the non-target trials. Although the tSV-T system only shows superior performance in terms of DCF08, we observe the same findings as tSV-R system in both tSV-F and tSV-FA systems (System 13 vs. 16 and System 21 vs. 24) in terms of EER, DCF08 and DCF10. With the same single talker evaluation condition, the tSV-FA method achieves the best performance of 2.63% (EER), 0.325 (DCF08) and 0.505 (DCF10) among the proposed four approaches. Comparing with the SV-FA system (System 24) trained and evaluated on same clean single speaker

condition, the tSV-FA method achieves 9.3%, 10.5% and 2.3% relative improvements in terms of EER, DCF08 and DCF10.

From Section IV-C, we conclude that the proposed tSV systems significantly improve multi-talker speaker verification. When the same model is evaluated in this section on single talker speech, they even outperform the traditional SV systems trained and evaluated for single talker condition. With these findings, we confirm that the proposed tSV system with various target speaker embedding schemes represents an unified framework for single and multi-talker speech, with tSV-FA as the top achiever.

E. Alternative Configuration for Target Speaker Enrollment

As enrollment utterance always contains single talker's voice, Fig. 6 shows an alternative to the standard configuration in Fig. 1. In the alternative configuration, we directly extract e_{ref} from the enrollment utterance $x(t)$ bypassing the speaker attention module. This alternative configuration works for all *Path B* schemes, i.e., tSV-T, tSV-F and tSV-FA.

In Table III(a), the systems that follow the alternative configuration in Fig. 6 are marked with "†". We observe that alternative tSV-F and tSV-FA systems (system 14, 15, 22, and 23) achieve comparable performance to those with the standard configurations (System 12, 13, 20 and 21). This suggests that the extracted speech from the speaker attention module is of similar quality as the original signal as far as speaker embedding is concerned. The SI-SDR performance of the extracted speech (56.7 dB) leads to the same conclusion, as reported in Table I.

However, we also observe that the alternative tSV-T system fails short of the expectation when comparing with the standard configuration, as reported in System 7 and 8 of Table III(a). We consider that the speaker representation module adopts a time-domain trainable speech encoder that is accustomed to extracted speech from speaker attention module. Therefore, there is a mismatch between the extracted speech and the original clean speech, that adversely affect the SV performance. This corroborates that the time-domain speaker representation module is sensitive to the mismatch between the extracted speech for training and clean speech for inference.

F. Benchmarking Against the State-of-The-Art on WSJ0-2talker Dataset

Recently studies, such as personal VAD [15], target VAD [16], seek to address speaker diarization problem for multi-talker speech. They could be used as the front-end for multi-talker SV when the speakers don't overlap. We are interested in the performance of multi-talker SV with such an oracle speaker diarization front-end. We obtain the ground-truth VAD by firstly applying the energy-based VAD on the single talker speech that makes up the mixture speech, and use the VAD labels as the diarization label for mixture speech.

Since the speech in WSJ0 corpus is quite clean and the energy-based VAD works well on the clean speech, we consider the obtained diarization labels as oracle diarization labels. With the oracle speaker diarization, the average percentage of removed non-target speech frames in the mixture speeches is around 24%. "OSD-SV" in Table IV shows the performance of multi-talker SV system with the oracle speaker diarization front-end, which also means the best performance achieved by speaker diarization for multi-talker SV. Compared with the multi-talker speaker verification results by tSV-R, tSV-T, tSV-F and tSV-FA, as reported from Table III(a), it is shown that the proposed tSV system significantly outperforms "OSD-SV".

TABLE IV
COMPARISONS BETWEEN TSV SYSTEMS AND OTHER STATE-OF-THE-ART SYSTEMS ON WSJ0-2TALKER DATASET. "OSD-SV" REPRESENTS THE CASE WHERE WE REPLACE THE SPEAKER ATTENTION MODULE IN FIG. 1 WITH AN ORACLE SPEAKER DIARIZATION (OSD). "SE-SV1" AND "SE-SV2" ARE TWO SPEAKER EXTRACTION-VERIFICATION PIPELINE SYSTEMS WITH SBF-MTSAL AND SBF-MTSAL-CONCAT [26] AS THE SPEAKER EXTRACTION FRONT-END, RESPECTIVELY

Systems	EER (%)	DCF08	DCF10
OSD-SV	14.60	0.851	0.908
SE-SV1 [31]	8.30	0.643	0.777
SE-SV2 [31]	7.77	0.631	0.747
tSV-R	6.67	0.580	0.774
tSV-T	6.60	0.582	0.759
tSV-F	6.53	0.570	0.708
tSV-FA	**5.03**	**0.449**	**0.604**
Upper-Bound	2.90	0.363	0.517

We further compare the proposed tSV system with the speaker extraction-verification pipeline, as summarized in Table IV. "SE-SV1" and "SE-SV2" are speaker extraction-verification pipeline systems [30], where SBF-MTSAL and SBF-MTSAL-Concat methods [26] serve as the speaker extraction front-end, and i-vector PLDA serves as the speaker verification back-end, without joint optimization. We observe the followings: 1) The proposed tSV system with various target speaker embedding schemes consistently outperform SE-SV1 [30] and SE-SV2 [30]; 2) The tSV-FA system achieves 35.3%, 28.8% and 19.1% relative improvement over SE-SV2, the most competitive system, in terms of EER, DCF08 and DCF10, respectively.

G. Evaluating tSV-FA on Libri-1talker/Libri-2talker Datasets

We further conduct experiments on a large Libri2Mix corpus with 1172 speakers in the training set. As the tSV-FA system achieves the best performance on WSJ0-1talker and WSJ0-2talker datasets, we only train and evaluate the tSV-FA system on the Libri-1talker and Libri-2talker datasets.

As summarized in Table III(b), System 25 and 32 are the baseline and the upper-bound systems for multi-talker speaker verification. Both are traditional SV systems trained under single talker condition, while the baseline is evaluated on 2-talker speech, and the upper bound system is evaluated on single talker speech. Similar to System 18 and 19, the pipeline systems are also trained on the Libri-1talker and Libri-2talker datasets as System 26 and 27.

We also evaluate the trained tSV-FA model on both 2-talker and single talker speech, denoted as Systems 28 and 29, just like in Section IV-C and Section IV-D. We observe that the tSV-FA system (System 28) achieves 55.3%, 40.6% and 27.4% relative improvements over the baseline (System 25) in terms of EER, DCF08 and DCF10, respectively. The tSV-FA system (System 28) also achieves better performance than the pipeline

TABLE V
EVALUATION OF THE TSV-FA SYSTEM ACROSS CORPORA

Training Set	Test Set	EER (%)	DCF08	DCF10
WSJ0-1talker & WSJ0-2talker	WSJ0-2talker	5.03	0.449	0.604
	WSJ0-1talker	2.63	0.325	0.505
	Libri-2talker	16.85	0.936	0.988
	Libri-1talker	11.33	0.744	0.907
Libri-1talker & Libri-2talker	WSJ0-2talker	4.9	0.444	0.558
	WSJ0-1talker	2.43	0.322	0.467
	Libri-2talker	7.88	0.508	0.681
	Libri-1talker	6.22	0.348	0.675

systems (System 26 and 27). The performance of tSV-FA system, evaluated either on 2-talker speech (System 28) or on single talker speech (System 29) is approaching that of the upper-bound system (System 32). The results further validate that tSV-FA system represents a unified solution for both single and multi-talker speech.

We further evaluate the alternative configurations in Fig. 6. Bypassing the speaker attention module, we directly extract e_{ref} from the enrollment utterance $x(t)$, and evaluate such an alternative tSV-FA system on the same 2-talker (System 30) and single talker (System 31) datasets. We observe that the alternative systems perform similarly to the standard configurations (Systems 28 and 29).

H. Evaluating tSV-FA Across Corpora

We further evaluate tSV-FA, that is trained on either WSJ0 or LibriSpeech dataset, but evaluated across WSJ0 and LibriSpeech datasets. The results are summarized in Table V.

Among the training-evaluation data pairs, we observe a low EER for WSJ0-WSJ0, and LibriSpeech-WSJ0 pairs. The EER for LibriSpeech-WSJ0 is even lower than that for WSJ0-WSJ0, which we consider is due to the fact that LibriSpeech is a larger database with many more speakers, thus, leading to a more robust tSV-FA system. This is an encouraging result, that confirms the ability of cross-corpus generalization from LibriSpeech-trained model to WSJ0 evaluation condition.

However, we haven't observed the same generalization ability for WSJ0-trained tSV-FA model. The EER for WSJ0-LibriSpeech is higher than that of LibriSpeech-LibriSpeech. Nonetheless, these results match our expectation, as we know that system performance under mismatched training-evaluation conditions is always poorer than that under matched conditions. We understand that WSJ0 is a smaller database with a small number of speakers. Furthermore, WSJ0 is recorded in a quiet acoustic environment with only one type of microphone channel, that may leads to the poor generalization under the LibriSpeech test condition.

V. CONCLUSIONS AND DISCUSSIONS

In this paper, we present a unified target speaker verification framework tSV for both single- and multi-talker speech. This framework jointly optimizes a speaker attention module and a speaker representation module via a multi-task learning framework. We systematically evaluated the performance of each individual module as well as the whole framework using two standard corpora. Experimental results show that the proposed framework significantly improves speaker verification performance for multi-talker speech, approaching that of traditional SV systems under the single-talker condition. It also outperforms other solutions, for instance, combining the state-of-the-art speaker extraction and speaker verification systems in tandem. Through multi-condition training, using both single- and multi-talker speech, we show that the proposed framework can easily generalize across different testing conditions.

There have been many studies on speaker extraction and speaker verification in literature. While these studies are usually carried out in isolation. For the first time, we show that with an advanced speaker attention-verification mechanism, we can achieve accurate speaker verification in multi-talker environment using the same enrollment utterance as what we have in the conventional single-talker speaker verification system.

We would like to acknowledge that it remains a challenge for a system to work in an unknown real-world acoustic environment. A universal solution is beyond the scope of this work. Instead, we have devoted this work to the exploration of a unified speaker verification solution for both single and multi-talker speech. Although we only conducted experiments on single and 2-talker speech, the proposed framework can be easily extended for scenarios with three or more speakers, together with noise and reverberation. As a future work, we are particularly interested in a training strategy, that can be enabled by self-supervised learning [48] and incremental learning [49], to allow the tSV system to adapt itself in the real-world changing environments.

It is in our plan to further the study of tSV by exploring the use of state-of-the-art loss functions and network structures, i.e., triplet loss [50]. We will also look into optimization of system architecture for computational efficiency, such as, network pruning and parameter quantization. By so doing, we will study the proposed system on an even larger database and benchmark against the pipeline baseline systems.

REFERENCES

[1] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 788–798, May 2011.
[2] P. Kenny, "Bayesian speaker verification with heavy-tailed priors," in *Proc. Odyssey: Speaker Lang. Recognit. Workshop*, Vol. 14, 2010.
[3] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Proc. 12th Annu. Conf. Int. Speech Commun. Assoc.*, 2011, pp. 249–252.
[4] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2007, pp. 1–8.
[5] D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, and S. Khudanpur, "Deep neural network-based speaker embeddings for end-to-end speaker verification," in *IEEE Workshop Spoken Lang. Technol.*, 2016, pp. 165–170.
[6] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5329–5333.
[7] D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey, and S. Khudanpur, "Speaker recognition for multi-speaker conversations using x-vectors," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 5796–5800.
[8] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 2, pp. 356–370, Feb. 2012.
[9] O. Kudashev, S. Novoselov, K. Simonchik, and A. Kozlov, "A speaker recognition system for the SITW challenge," in *Proc. INTERSPEECH*, 2016, pp. 833–837.

[10] Y. Liu, Y. Tian, L. He, and J. Liu, "Investigating various diarization algorithms for speaker in the wild (SITW) speaker recognition challenge," in *Proc. INTERSPEECH*, 2016, pp. 853–857.

[11] O. Novotnỳ, P. Matejka, O. Plchot, O. Glembek, L. Burget, and J. Cernockỳ, "Analysis of speaker recognition systems in realistic scenarios of the SITW 2016 challenge," in *Proc. INTERSPEECH*, 2016, pp. 828–832.

[12] H. Ghaemmaghami *et al.*, "Speakers in the wild (SITW): The QUT speaker recognition system," in *Proc. 17th Annu. Conf. Int. Speech Commun. Assoc.*, 2016, pp. 838–842.

[13] D. Charlet, C. Barras, and J.-S. Liénard, "Impact of overlapping speech detection on speaker diarization for broadcast news and debates," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2013, pp. 7707–7711.

[14] S. H. Yella and H. Bourlard, "Overlapping speech detection using long-term conversational features for speaker diarization in meeting room conversations," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1688–1700, Dec. 2014.

[15] S. Ding, Q. Wang, S.-y. Chang, L. Wan, and I. L. Moreno, "Personal vad: Speaker-conditioned voice activity detection," in *Proc. Odyssey: Speaker Lang. Recognit. Workshop*, pp. 433–439, 2019.

[16] I. Medennikov *et al.*, "Target-speaker voice activity detection: A novel approach for multi-speaker diarization in a dinner party scenario," in *Proc. INTERSPEECH, 2020*, pp. 274–278.

[17] 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.

[18] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 246–250.

[19] 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.

[20] C. Xu, W. Rao, X. Xiao, E. S. Chng, and H. Li, "Single channel speech separation with constrained utterance level permutation invariant training using grid lstm," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 6–10.

[21] C. Xu, W. Rao, E. S. Chng, and H. Li, "A shifted delta coefficient objective for monaural speech separation using multi-task learning," in *Proc. INTERSPEECH*, 2018, pp. 3479–3483.

[22] 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.

[23] 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.

[24] C. Xu, W. Rao, C. E. Siong, and H. Li, "SpEx: Multi-scale time domain speaker extraction network," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1370–1384, 2020, doi: 10.1109/TASLP.2020.2987429.

[25] 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.

[26] C. Xu, W. Rao, E. S. Chng, and H. Li, "Optimization of speaker extraction neural network with magnitude and temporal spectrum approximation loss," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6990–6994.

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

[28] J. Wang *et al.*, "Deep extractor network for target speaker recovery from single channel speech mixtures," in *Proc. INTERSPEECH*, 2018, pp. 307–311.

[29] M. Ge, C. Xu, L. Wang, E. S. Chng, D. Jianwu, and H. Li, "SpEx+ : A complete time domain speaker extraction network," in *Proc. INTERSPEECH*, pp. 1406–1410, 2020.

[30] W. Rao, C. Xu, E. S. Chng, and H. Li, "Target speaker extraction for multi-talker speaker verification," in *Proc. INTERSPEECH*, 2019, pp. 1273–1277.

[31] H. Taherian, Z.-Q. Wang, J. Chang, and D. Wang, "Robust speaker recognition based on single-channel and multi-channel speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1293–1302, 2020, doi: 10.1109/TASLP.2020.2986896.

[32] S. Kataria, S.-X. Zhang, and D. Yu, "Exploring multi-channel features for speaker verification with joint vad and speech enhancement," 2020, *arXiv:2010.12692*.

[33] C.-J. Peng, Y.-J. Chan, C. Yu, S.-S. Wang, Y. Tsao, and T.-S. Chi, "Attention-based multi-task learning for speech-enhancement and speaker-identification in multi-speaker dialogue scenario," in *Proc. IEEE Int. Symp. Circuits Syst.*, pp. 1–5, 2021.

[34] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 4879–4883.

[35] Z. Huang, S. Wang, and K. Yu, "Angular softmax for short-duration text-independent speaker verification," in *Proc. INTERSPEECH*, 2018, pp. 3623–3627.

[36] W. Rao and M. W. Mak, "Boosting the performance of i-vector based speaker verification via utterance partitioning," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 21, no. 5, pp. 1012–1022, May 2013.

[37] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," in *Proc. INTERSPEECH*, 2018, pp. 2252–2256.

[38] Q. Wang, K. Okabe, K. A. Lee, H. Yamamoto, and T. Koshinaka, "Attention mechanism in speaker recognition: What does it learn in deep speaker embedding," in *IEEE Spoken Lang. Technol. Workshop*, 2018, pp. 1052–1059.

[39] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr-half-baked or well done," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 626–630.

[40] M. Borsdorf, C. Xu, H. Li, and T. Schultz, "Universal speaker extraction in the presence and absence of target speakers for speech of one and two talkers," in *Proc. INTERSPEECH*, 2021.

[41] D. E. Sturim and D. A. Reynolds, "Speaker adaptive cohort selection for tnorm in text-independent speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2005, vol. 1, pp. I- 741.

[42] J. Garofolo, D. Graff, D. Paul, and D. Pallett, "Csr-i (wsj0) complete ldc93s6a," *Philadelphia: Linguistic Data Consortium*, 1993.

[43] 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.

[44] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "Librimix: An open-source dataset for generalizable speech separation," 2020, *arXiv:2005.11262*.

[45] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015, *arXiv:1412.6980*.

[46] A. F. Martin and C. S. Greenberg, "Nist 2008 speaker recognition evaluation: Performance across telephone and room microphone channels," in *Proc. INTERSPEECH*, 2009, pp. 2579–2582.

[47] A. F. Martin and C. S. Greenberg, "The nist 2010 speaker recognition evaluation," in *Proc. INTERSPEECH*, 2010, pp. 2726–2729.

[48] C. Doersch and A. Zisserman, "Multi-task self-supervised visual learning," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2070–2079.

[49] Y. Wu *et al.*, "Large scale incremental learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 374–382.

[50] C. Zhang, K. Koishida, and J. H. Hansen, "Text-independent speaker verification based on triplet convolutional neural network embeddings," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 9, pp. 1633–1644, Sep. 2018.



Chenglin Xu (Member, IEEE) received the B.Eng. and M.Sc. degrees from Northwestern Polytechnical University, Xi'an, China, in 2012 and 2015, respectively, and the Ph.D. degree from Nanyang Technological University, Singapore, in 2020. He is currently a Research Fellow with HLT Lab, Department of Electrical and Computer Engineering, National University of Singapore. His research interests include speech extraction, speech enhancement, source separation, multitalker speaker verification, and robust speech recognition.



Wei Rao (Member, IEEE) received the B.Eng. degree in electronic information engineering and the M.Eng. degree in information and telecommunication engineering from the China University of Geosciences, Wuhan, China, in 2007 and 2010, respectively, and the Ph.D. degree from The Hong Kong Polytechnic University, Hong Kong, in 2015. She is currently a Senior Researcher with Tencent Ethereal Audio Lab, Shenzhen, China.

Prior to joining Tencent, she was a Research Scientist with Temasek Laboratories, Nanyang Technological University, Singapore, from 2015 to 2018, and with HLT Lab, Department of Electrical and Computer Engineering, National University of Singapore, Singapore, from 2018 to 2020. Her research interests include robust speaker recognition, speech signal processing, and machine learning. She was the recipient of Best Technology Award from Temasek Laboratories, NTU. She was the Area Chair of ICTAI 2016, a Member of the technical program committee of ICASSP and INTERSPEECH, and the Local Arrangement Co-Chair of ASRU 2019.



Jibin Wu (Member, IEEE) received the B.E. and Ph.D. degrees in electrical engineering from the National University of Singapore, Singapore, in 2016 and 2020, respectively. He is currently a Research Fellow with HLT Lab, Department of Electrical and Computer Engineering, National University of Singapore. His research interests include spiking neural network, neuromorphic computing, auditory modelling, and automatic speech recognition.



Haizhou Li (Fellow, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in electrical and electronic engineering from the South China University of Technology, Guangzhou, China, in 1984, 1987, and 1990, respectively. He is currently a Professor with the Department of Electrical and Computer Engineering, National University of Singapore (NUS), Singapore. Prior to joining NUS, he taught in the University of Hong Kong, Hong Kong, from 1988 to 1990 and South China University of Technology, Guangzhou, China, from 1990 to 1994. He was a Visiting Professor with CRIN, France, from 1994 to 1995, the Research Manager with Apple-ISS Research Centre from 1996 to 1998, the Research Director of Lernout & Hauspie Asia Pacific from 1999 to 2001, the Vice President of InfoTalk Corp. Ltd., Hong Kong, from 2001 to 2003, and the Principal Scientist and Department Head of Human Language Technology in the Institute for Infocomm Research, Singapore, from 2003 to 2016. His research interests include automatic speech recognition, speaker and language recognition, and natural language processing. He was the Editor-in-Chief of the IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING from 2015 to 2018, has been a Member of the Editorial Board of Computer Speech and Language since 2012, an elected Member of the IEEE Speech and Language Processing Technical Committee from 2013 to 2015, the President of the International Speech Communication Association from 2015 to 2017, the President of Asia Pacific Signal and Information Processing Association from 2015 to 2016, and the President of Asian Federation of Natural Language Processing from 2017 to 2018. He was the General Chair of ACL 2012, INTERSPEECH 2014, and ASRU 2019. He is a Fellow of the ISCA. He was the recipient of the National Infocomm Award 2002, and the President's Technology Award 2013 in Singapore. He was named one of the two Nokia Visiting Professors in 2009 by the Nokia Foundation, and U Bremen Excellence Chair Professor in 2019.