Interspeech 2024
1-5 September 2024, Kos, Greece

Joint Speaker Features Learning for Audio-visual Multichannel Speech Separation and Recognition

Guinan Li[1], *Jiajun Deng*[1], *Youjun Chen*[1], *Mengzhe Geng*[2], *Shujie Hu*[1], *Zhe Li*[3], *Zengrui Jin*[1], *Tianzi Wang*[1], *Xurong Xie*[4], *Helen Meng*[1], *Xunying Liu*[1]

[1]The Chinese University of Hong Kong; [2]National Research Council Canada; [3]The Hong Kong Polytechnic University; [4]Institute of Software, Chinese Academy of Sciences

{gnli,jjdeng, yjchen, sjhu, zrjin, twang, hmmeng, xyliu}@se.cuhk.edu.hk,
Mengzhe.Geng@nrc-cnrc.gc.ca, lizhe.li@connect.polyu.hk, xurong@iscas.ac.cn

Abstract

This paper proposes joint speaker feature learning methods for zero-shot adaptation of audio-visual multichannel speech separation and recognition systems. xVector and ECAPA-TDNN speaker encoders are connected using purpose-built fusion blocks and tightly integrated with the complete system training. Experiments conducted on LRS3-TED data simulated multichannel overlapped speech suggest that joint speaker feature learning consistently improves speech separation and recognition performance over the baselines without joint speaker feature estimation. Further analyses reveal performance improvements are strongly correlated with increased inter-speaker discrimination measured using cosine similarity. The best-performing joint speaker feature learning adapted system outperformed the baseline fine-tuned WavLM model by statistically significant WER reductions of 21.6% and 25.3% absolute (67.5% and 83.5% relative) on Dev and Test sets after incorporating WavLM features and video modality.

Index Terms: Speaker features, Zero-shot adaptation, Speech separation, Speech recognition

1. Introduction

Despite the rapid progress of automatic speech recognition (ASR) in recent decades, accurate recognition of cocktail party speech [1] remains a highly challenging task to date. To this end, microphone arrays play a key role in state-of-the-art speech separation and recognition systems designed for such data [2]. The required array beamforming techniques used to perform multichannel signal integration are implemented as time or frequency domain filters. These are represented by time-domain delay and sum [3, 4], frequency-domain minimum variance distortionless response (MVDR) [5] and generalized eigenvalue (GEV) [6] based multichannel integration approaches.

In recent years, end-to-end DNN-based microphone array beamforming techniques represented by a) neural time-frequency (TF) masking approaches [7]; b) neural Filter and Sum methods [8, 9]; and c) mask-based MVDR [10] and generalized eigenvalues (GEV) [11] approaches have been widely adopted. In addition, incorporating visual information into either multi speech separation front-ends alone [12], or further into speech recognition back-ends [13], can further improve the overall system performance.

Natural speech represented by cocktail party speech is highly heterogeneous. To this end, speaker adaptation techniques have been widely studied as powerful solutions to customise ASR systems for individual users. These include, but not limited to: 1) auxiliary speaker embedding based approaches [14–18], e.g. iVector [14] and xVector [15]; 2) feature transformation based methods, e.g., feature-space MLLR [19]; and 3) model-based methods [20–23] that estimate speaker dependent (SD) adapter parameters implemented as, e.g. learning hidden unit contributions (LHUC) [21], during speaker adaptive training (SAT) and test-time unsupervised adaptation [22, 23].

Recent researches have also demonstrated that overlapping speech separation [24–32] and recognition [24, 26, 31, 32] systems also benefit from modelling speaker-level features . These include, but not limited to, SpeakerBeam [24, 25], VoiceFilter [26], TEA-PSE [27, 28], SpEx [29, 30] and PSE [31, 32]. However, these prior studies suffer from several limitations: **1)** Lack of tight integration between speaker feature learning and backbone speech enhancement-recognition model training. Speaker features are learned either separately and de-coupled from the backbone speech enhancement-recognition model without any form of joint training [26–28, 31, 32], or only partially jointly learned with the speech enhancement front-end alone, while separated from the speech recognition back-end [24, 25, 29, 30]. **2)** Lack of zero-shot, instantaneous adaptation methods for speech enhancement models. Instead the commonly used enrollment-based speaker adaptation approaches [24–32] require clean speech samples to be explicitly recorded at the onset of user personalization. This not only incurs processing latency but also leads to privacy concerns. **3)** The efficacy of speaker adaptation was predominantly evaluated on speech separation systems alone [24–32], while there is a lack of holistic incorporation of the speaker features into both the speech separation front-end and recognition back-end components.

To this end, this paper proposes joint speaker features learning approaches for zero-shot, instantaneous adaptation of audio-visual multichannel speech separation and recognition systems. xVector [15] or ECAPA-TDNN [33] speaker feature extraction modules are connected with the backbone system using purpose-built fusion blocks, and tightly integrated with the complete system training. The resulting systems support zero-shot, enrollment-free and on-the-fly adaptation to unseen speakers without requiring pre-recorded user data. Experiments conducted on simulated multichannel overlapped and noisy speech data constructed using the benchmark LRS3-TED [34] dataset suggest that joint speaker features learning consistently produces speech enhancement and recognition performance improvements over the baselines without joint speaker feature estimation. Further analyses reveal that these performance improvements are strongly correlated with the increase of inter-speaker discrimination measured using cosine similarity. The best-performing joint speaker feature learning adapted system outperformed the baseline fine-tuned WavLM model by statistically significant WER reductions of **21.6% absolute** (**67.5% relative**) and **25.3% absolute** (**83.5% relative**) on Dev and Test sets after incorporating WavLM features and video modality. The main contributions of this paper are summarized below:

10.21437/Interspeech.2024-1063

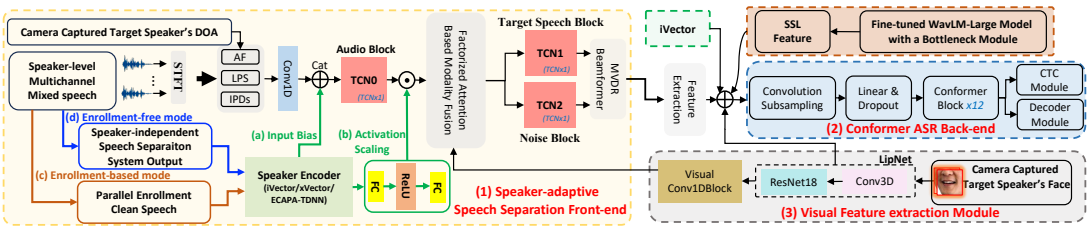

Figure 1: *Example of joint speaker features learning for an audio-visual multichannel speech separation and recognition system including the following components:* ***(1) Speaker-adaptive speech separation front-end*** *implemented using TCNs and mask-based MVDR beamforming;* ***(2) Conformer ASR Back-end****; and* ***(3) Visual Feature extraction Module****. The Speaker Encoder (iVector, xVector or ECAPA-TDNN) module is connected with the backbone system using purpose-built fusion blocks based on either* ***(a) Input Bias*** *or* ***(b) Activation Scaling****, and tightly integrated with complete system training. Speaker adaptation is performed in either* ***(c) enrollment-based mode*** *[24–32] requiring pre-recorded speaker-level parallel clean-noisy speech; or* ***(d) zero-short, enrollment-free mode*** *that does not require pre-recorded speaker-level clean speech data. "Cat" and "⊙" denote the concatenation and element-wise product operation, respectively.*

1) To the best of our knowledge, this paper presents the first use of joint speaker features learning approaches for zero-shot user adaptation of audio-visual multichannel speech separation and recognition systems. In contrast, in prior researches speaker features are learned either separately and de-coupled from the backbone system without any form of joint training [26–28, 31, 32], or only partially jointly learned with the speech enhancement front-end alone, while separated from the speech recognition backend [24, 25, 29, 30].

2) To the best of our knowledge, this paper pioneers the use of zero-shot, enrollment-free speaker adaptation for speech separation and recognition tasks. In contrast, the related prior studies only focus on less practical enrollment-based adaptation techniques [24–32] that require clean speech samples to be explicitly recorded at the onset of user personalization.

3) This paper presents the first investigation of the complete incorporation of speaker features into all the components of a complete end-to-end audio-visual multichannel speech separation and recognition system. In contrast, prior researches consider speaker adaptation of either the speech separation front-end [12, 24–32] alone, or the speech recognition back-end [14, 16–19, 21–23] only.

2. Audio-visual Speech Separation

2.1. Mask-based MVDR

In MVDR beamforming [5, 6], a linear filter $\mathbf{w}(f) \in \mathbb{C}^R$ is applied to the multichannel mixture speech short-time Fourier transform (STFT) spectrum $\mathbf{y}(t,f) \in \mathbb{C}^R$ to produce the filtered output $\hat{S}(t,f)$ as:

$$\hat{S}(t,f) = \mathbf{w}(f)^H \mathbf{y}(t,f) = \underbrace{\mathbf{w}(f)^H \mathbf{x}(t,f)}_{\text{target speech component}} + \underbrace{\mathbf{w}(f)^H \mathbf{n}(t,f)}_{\text{residual noise}}, \tag{1}$$

where R denotes the channel number of a microphone array. t and f denote the indices of time and frequency bins, respectively. $(\cdot)^H$ denotes the conjugate transpose operator. $\mathbf{x}(t,f) \in \mathbb{C}^R$ is a complex vector containing the clean speech signals. $\mathbf{n}(t,f) \in \mathbb{C}^R$ represents either the interfering speaker's speech or additive background noise alone, or a combination of both.

By minimizing the residual noise output while imposing a distortionless constraint on the target speech [5], the MVDR beamforming filter is estimated as

$$\mathbf{w}(f) = \frac{\mathbf{\Phi}_n(f)^{-1}\mathbf{\Phi}_x(f)}{\mathrm{tr}\left(\mathbf{\Phi}_n(f)^{-1}\mathbf{\Phi}_x(f)\right)}\mathbf{u}_r, \tag{2}$$

where $\mathbf{u}_r = [0, \ldots, 1, \ldots, 0]^T \in \mathbb{R}^R$ is a one-hot reference vector where its r-th component equals to one. $\mathrm{tr}(\cdot)$ denotes the trace operator. Without loss of generality, we select the first channel as reference in this paper. The target speaker power spectral density (PSD) matrix $\mathbf{\Phi}_x(f)$ and noise-specific PSD matrix $\mathbf{\Phi}_n(f)^{-1}$ are computed using DNN predicted complex TF masks. More details can be found in the paper [13].

2.2. Audio and Visual Modality Fusion

Audio modality: Three types of audio features including the complex STFT spectrum of all the microphone array channels, the inter-microphone phase differences (IPDs) [35] and angle feature (AF) [36] are adopted as the audio inputs (Fig. 1, upper left, in light grey). Following [13], temporal convolutional networks (TCNs) [37] are used. Each TCN block is stacked by 8 Dilated1DConvBlock with exponentially increased dilation factors $2^0, 2^1, \ldots, 2^7$. The log-power spectrum (LPS) features of the reference microphone channel are concatenated with the IPDs and AF features before being fed into a Conv1D module followed by a single TCN-based Audio Block to compute audio embeddings.

Visual modality: The target speaker's lip region obtained via face tracking is fed into a LipNet [38] containing a 3D convolution layer (Fig. 1, bottom right, in pink) and 18-layer ResNet [39] (Fig. 1, bottom right, in light turquoise) to extract visual features before being fed into a Visual Conv1DBlock. Then, the output of the Visual Conv1DBlock is up-sampled and time synchronised with the audio frames via linear interpolation to compute visual embeddings.

Modality fusion: A factorized attention-based modality fusion module (Fig. 1, middle, in grep) following the prior works in [13] is utilized to integrate the audio and visual embeddings.

Separation Network Training Cost Function: The mask-MVDR based multichannel speech separation network is trained separately to maximize the SISNR metric before joint fine-tuning using the back-end ASR error loss in Eqn. (3).

3. Speaker Adaptation of Audio-visual Speech Separation Front-end

Speaker features based on either iVector [14], xVector [15] or ECAPA-TDNN [33] are connected with the speech separation backbone system using either input bias or activation scaling. Among these, xVector and ECAPA-TDNN speaker encoders are tightly integrated with the entire speech enhancement-recognition model training. Speaker adaptation is performed in either enrollment-based or zero-shot enrollment-free modes.

3.1. Speaker Feature Fusion

The utterance-level speaker features extracted from each speaker encoder are incorporated into the backbone system using either **input bias** or **activation scaling** with purpose-built fusion blocks for adaptation. **a) Input bias fusion** refers to the utterance-level speaker features being concatenated directly

with the output audio embeddings from the Conv1D module along the feature dimension (Fig. 1, middle, green line marked with "Input Bias"). **b) Activation scaling fusion** feeds the speaker features through a fusion module before being further applied to the TCN outputs using element-wise product. The fusion module consists of a fully connected (FC) layer with 256 x 200 dimensions and a ReLU activation followed by an output FC layer with 256 output units (Fig. 1, middle, green line marked with "Activation Scaling").

3.2. Speaker Feature Learning

Speaker features are learned either **separately from** or **tightly integrated with** the backbone speech enhancement-recognition model training. **a)** In **non-joint speaker feature learning**, speaker encoders are trained in an offline manner using the speaker recognition error loss [26–28, 31, 32]. The features extracted from these speaker encoders are fed into the backbone speech enhancement-recognition system using the above fusion methods. **b)** In **joint speaker feature learning**, xVector or ECAPA-TDNN speaker encoders are tightly integrated with the backbone enhancement-recognition model. They are updated in turn using the SI-SNR cost for speech separation front-end alone, and followed by ASR cost for the entire enhancement-recognition system. Further analyses in Sec. 5.2 reveal that these performance improvements brought by jointly speaker feature learning are strongly correlated with the increase of inter-speaker discrimination measured using cosine similarity.

3.3. Adaptation Supervision

Speaker adaptation is performed in either **enrollment-based** mode [25–33] or **zero-short, enrollment-free** mode. **a)** In **enrollment-based supervised** adaptation [24–32] (Fig. 1, bottom left, light brown box), pre-recorded speaker-level parallel clean-noisy speech is required at the onset of user personalization. For example, SpeakeBeam [24] uses 100% of the test set data' ground truth clean speech to construct speaker-level enrollment data[1]. **b)** In **enrollment-free, zero-shot** adaptation (Fig. 1, middle left, blue box), speaker-level parallel clean-noisy speech is not required. The target clean speech is replaced by the enhanced speech outputs produced by a speaker-independent speech separation system before being fed into the speaker feature encoder. This form of zero-shot speaker feature learning directly from untranscribed overlapped and noisy speech alone alleviates the processing latency in speaker clean-noisy parallel speech pre-recording and privacy issues brought by enrollment-based adaptation.

4. Audio-visual Conformer ASR Back-end

Speech separation outputs are used to extract Mel-filterbank features. They are concatenated with visual features and fed into the ASR back-end. The hybrid CTC-attention Conformer ASR model [40] consists of an encoder module and a decoder module. Fig. 1 (middle right, in light blue) shows an example of a Conformer ASR system. More details can be found in [40]. The following multi-task criterion interpolation between the CTC and attention error costs [41] is used in Conformer model training,

$$\mathcal{L}_{\text{ASR}} = (1-\beta)\mathcal{L}_{att} + \beta\mathcal{L}_{ctc}, \tag{3}$$

where $\beta \in [0, 1]$ is a tunable hyper-parameter and empirically set as 0.3 for training and 0.4 for recognition in this paper.

End-to-end joint fine-tuning of speech separation and the recognition components [13] using the ASR cost function of Eqn. 3 is applied in this paper.

[1]https://github.com/BUTSpeechFIT/speakerbeam

iVector-based Speaker Adaptation: iVector features extracted with a 100ms window are up-sampled to be time synchronised with the Mel-filterbank audio frames [42] (Fig. 1, top middle, in light green). Then, the iVector, audio and visual features are concatenated together and fed into the ASR back-end.

SSL pre-trained WavLM features: The enhanced speech fine-tuned WavLM-Large models [43] contain a Bottleneck Module [44]. It is used to extract SSL pre-trained features that are fed into Conformer ASR systems via input feature concatenation with Mel-filterbanks (Fig. 1, top left, in light orange).

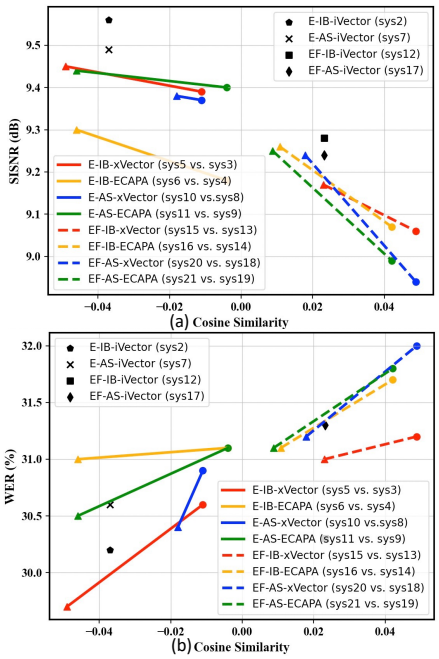

Figure 2: *Correlation between cosine similarity and SISNR or overall WER on "Dev" set for systems in Table 1. Comparable systems with/without joint and non-joint speaker features learning marked as "△" and "◯" respectively at two ends of each colored line. "EF", "E", "IB", and "AS" denote "enrollment free", "enrollment", "input bias" and "activation scaling".*

5. Experiments

5.1. Experimental Setup

Simulated Multichannel Mixture Speech: Experiments were conducted on the overlapped and noisy speech simulated using the LRS3-TED dataset [34]. A 15-channel symmetric linear array with non-even inter-channel spacing [7,6,5,4,3,2,1,1,2,3,4,5,6,7]cm is used in the simulation process. 843 point-source noises [46] and 20000 room impulse responses (RIRs) generated by the image method [47] in 400 different simulated rooms are used in our experiment. The distance between a sound source and the microphone array center is uniformly sampled from a range of 1m to 5m and the room size ranges from 4m×4m×3m to 10m×10m×6m (length×width×height). The reverberation time T_{60} is uniformly sampled from a range of 0.14s to 0.92s. The average overlapping ratio is around 80%. The signal-to-noise ratio (SNR) is uniformly sampled from {0, 5, 10, 15, 20}dB, and the signal-to-interference ratio (SIR) is uniformly sampled from {-6, 0, 6}dB. In addition, the angle difference relative to the microphone array between the target and interfering speakers is uniformly sampled from four ranges of the angle difference $\{[0°, 15°), [15°, 45°), [45°, 90°), [90°, 180°)\}$. The final simulated multichannel datasets contain four subsets with 110354,

Table 1: *Performance of joint speaker features learning for audio-only speech separation and recognition systems in either enrollment-based or enrollment-free mode. The ASR back-end used to evaluate the WER metric is fine-tuned in a pipelined manner by the speech separation output. "Enroll.", "Spk. Feat.", and "O.V." denote "enrollment", "speaker feature" and "overall", respectively. "†" represents a statistically significant (MAPSSWE, α=0.05 [45]) WER difference over the SI baseline (sys. 1).*

Sys.	Speaker Adaptation for Speech Separation				SISNR(↑)/PESQ(↑)/STOI(↑)/WER (↓)					
	Joint Spk. Feat. Learning	Adaptation Supervision	Spk. Feat. Fusion	Spk. Feat. Encoding	Dev					Test
					[0°, 15°)	[15°, 45°)	[45°, 90°)	[90°, 180°)	O.V.	O.V.
The raw first channel of overlapped speech					0.05/1.92/65.16/55.4	-0.05/1.91/64.99/56.3	-0.02/1.87/65.25/53.7	-0.04/1.95/65.34/55.0	-0.02/1.91/65.18/55.1	0.06/1.83/64.52/48.6
1	Speaker-independent (SI) baseline				4.26/2.44/76.37/48.8	9.16/2.86/85.45/29.6	10.58/2.95/87.79/26.0	10.84/3.05/88.80/24.7	8.69/2.82/84.56/32.5	7.62/2.75/84.94/25.1
2	✗	Enroll.	Input Bias	iVector	6.56/**2.60/80.80/41.2**	**9.71/2.89**/85.89/**29.0**	**10.91/2.97**/88.05/**25.4**	**11.13/3.06**/88.94/24.6	**9.56/2.88/85.90**/30.2†	**8.72/2.83/87.08/21.6**†
3	✗			xVector	**6.63**/2.59/80.66/42.0	9.43/2.88/85.79/29.4	10.61/2.95/87.79/25.9	10.93/3.04/88.80/24.5	9.39/2.86/85.74/30.6†	8.50/2.81/86.79/22.6†
4	✗			ECAPA-TDNN	5.73/2.53/79.19/44.1	9.35/2.87/85.63/**29.0**	10.70/2.96/87.93/26.3	10.98/3.05/88.86/24.3	9.18/2.85/85.38/31.1†	8.30/2.79/86.48/23.0†
5	✓			xVector	6.37/2.58/80.17/42.9	9.56/**2.89/85.97**/29.1	10.85/**2.97/88.10/25.4**	11.07/**3.06/88.97/24.2**	9.45/2.87/85.78/**29.7**†	8.56/2.82/86.84/21.9†
6	✓			ECAPA-TDNN	6.04/2.55/79.54/43.5	9.38/2.87/85.73/29.7	10.78/2.96/88.03/26.0	11.07/**3.06/88.97**/24.3	9.30/2.86/85.54/31.0†	8.46/2.80/86.78/22.7†
7	✗		Act. Scaling	iVector	**7.13/2.63/81.51/40.4**	9.34/2.87/85.54/30.3	10.67/2.95/87.67/26.3	10.86/3.04/88.70/25.0	**9.49/2.87/85.84**/30.6†	**8.74/2.82/87.17/22.4**†
8	✗			xVector	6.59/2.58/80.37/42.4	9.31/2.87/85.62/29.6	10.63/**2.96**/87.95/26.6	10.99/**3.05**/88.87/24.6	9.37/2.86/85.68/30.9†	8.47/2.80/86.79/23.0†
9	✗			ECAPA-TDNN	6.58/2.58/80.73/43.1	9.35/2.87/85.63/29.7	10.68/2.95/87.91/26.5	**11.02**/3.04/88.83/24.4	9.40/2.86/85.75/31.1†	8.47/2.80/86.71/22.7†
10	✓			xVector	6.36/2.57/80.20/41.7	9.47/**2.88**/85.78/**29.0**	**10.74/2.96/87.97**/26.0	10.98/**3.05/88.93**/24.3	9.38/**2.87**/85.70/**30.4**†	8.55/2.81/86.87/22.7†
11	✓			ECAPA-TDNN	6.62/2.58/80.67/42.2	**9.54/2.88/85.89**/29.3	10.67/**2.96**/87.89/**25.8**	10.99/**3.05**/88.89/**24.1**	9.44/**2.87**/85.82/30.5†	8.57/2.81/86.82/22.6†
12	✗	Enroll. Free	Input Bias	iVector	**6.01/2.57/80.09/42.0**	9.40/2.89/85.72/29.0	10.80/2.96/88.05/**25.2**	10.95/3.06/88.91/24.6	**9.28/2.87/85.67/30.3**†	7.97/2.78/85.84/23.6†
13	✗			xVector	4.82/2.50/77.21/45.4	9.65/**2.90**/86.00/28.8	10.82/2.97/88.04/25.6	11.05/3.06/89.00/**24.2**	9.06/2.86/85.03/31.2†	8.16/2.80/85.87/23.7†
14	✗			ECAPA-TDNN	4.94/2.51/77.41/45.6	9.55/2.89/85.82/29.4	10.81/2.96/88.01/26.7	11.04/3.05/88.90/24.4	9.07/2.85/85.00/31.7†	8.20/2.80/85.92/23.2†
15	✓			xVector	4.98/2.51/77.47/44.1	9.68/2.89/85.93/28.8	10.95/2.97/88.06/25.6	11.16/3.06/89.02/24.7	9.17/2.86/85.09/31.0†	8.25/2.80/85.84/23.2†
16	✓			ECAPA-TDNN	5.12/2.52/77.56/44.9	**9.80/2.90/86.18/28.4**	**10.97/2.98/88.19**/25.8	**11.24/3.07/89.09**/24.3	9.26/2.86/85.22/31.1†	**8.31/2.81/86.13/22.7**†
17	✗		Act. Scaling	iVector	**6.45/2.59/80.28/42.5**	9.15/2.86/85.21/31.0	10.60/2.95/87.63/26.2	10.81/3.04/88.64/24.9	9.24/2.86/**85.42**/31.3†	8.15/2.78/85.89/23.6†
18	✗			xVector	4.79/2.49/77.09/46.3	9.42/2.88/85.68/30.2	10.61/2.96/87.89/25.9	11.00/3.05/88.89/24.8	8.94/2.85/84.85/32.0	8.03/2.78/85.47/24.2 †
19	✗			ECAPA-TDNN	4.79/2.50/77.09/45.7	9.38/2.88/85.57/30.6	10.80/2.97/88.00/**25.8**	11.07/**3.07**/89.00/**24.4**	8.99/2.85/84.88/31.8†	8.08/2.79/85.65/23.6†
20	✓			xVector	5.08/2.52/77.51/44.5	9.80/2.90/**86.14**/28.9	**10.96/2.98**/88.21/**25.8**	11.22/3.06/89.06/24.9	9.24/2.86/85.19/31.2†	8.29/2.81/**86.01**/23.2†
21	✓			ECAPA-TDNN	5.07/2.52/77.57/44.7	**9.82/2.91**/86.06/**28.8**	10.95/**2.98/88.24**/25.9	**11.26/3.07/89.14**/24.5	**9.25/2.87**/85.22/**31.1**†	**8.30/2.82**/85.96/**22.5**†

3122, 3136 and 1321 utterances each for training (111.47 hours, 4886 speakers), val (3.15 hours, 157 speakers), dev (3.14 hours, 158 speakers) and test (0.84 hours, 412 speakers).

Baseline System Description: The configuration settings of the baseline speech separation front-end, visual feature extraction and Conformer ASR back-end are referred to [13].

Implementation Details: Speaker features are empirically configured for iVector, xVector or ECAPA-TDNN as follows: **a)** speaker features are fed into the system at the input and output of TCN0 Audio Block (Fig. 1, upper middle, in pink) respectively for input bias and activation scaling based fusion; **b)** features dimensionality set as 100, 256 and 256 respectively for iVector, xVector and ECAPA-TDNN in all the experiments.

5.2. Experimental Results

Performance of joint speaker features learning is shown in Table 1. Several trends can be found. **1)** The proposed joint speaker features learning brings improvements consistently for both enrollment-based and enrollment-free zero-shot adaptation compared to non-joint speaker features learning (sys. 5,6,10,11,15,16,20,21 vs. sys. 3,4,8,9,13,14,18,19), with a statistically significant WER reduction up to **0.7%** and **1.1%** absolute (**2.2%** and **4.7%** relative) (sys. 21 vs. sys. 19) on the dev and test sets. Consistent improvements are also found on speech enhancement metrics (SISNR [7], PESQ [48] and STOI [49]) scores. **2)** These improvements are consistently correlated with inter-speaker discrimination measured in cosine similarity. In terms of the correlation between cosine similarity and SISNR in Fig 2(a) or overall WER in Fig 2(b), using joint speaker features learning can boost the system performance characterized by lower cosine similarity scores[2]. **3)** The performance of enrollment-free adaptation is comparable to more expensive enrollment-based adaptation when varying the inter-speaker angle difference between 15° to 180°, except for the most challenging subset when the inter-speaker angle difference is under 15°. This is expected as the quality of SI system speech separation outputs, which are used to extract target speaker features, is heavily degraded and thus negatively impacts speaker adaptation performance. How to mitigate the sensitivity to inter-speaker angle difference for enrollment-free adaptation will be studied in future research.

Performance of speaker adaptation for end-to-end speech separation and recognition is shown in Table 2. The baseline system (sys. 1, in Table 1) and the respective best-performed systems from both enrollment-based and zero-shot enrollment-free adaptation (sys. 2,21, in Table 1) are used correspondingly for end-to-end audio-only systems training (sys. 1,2,3, in Table 2). Such selected systems are jointly fine-tuned except for using a stronger Conformer ASR back-end integrated with fine-tuned WavLM SSL features. When applying speaker adaptation to both stages, the proposed speaker-adaptive audio-visual systems consistently outperformed the SI baseline by statistically significant WER reductions of up to **2.3**% and **2.7**% absolute (**18.1**% and **35.1**% relative) on dev and test sets (sys. 5 vs. sys. 4). It can be observed that the proposed best-performed speaker-adaptive audio-visual system (sys. 5) can significantly outperform the baseline fine-tuned WavLM Model using the raw first channel of overlapped speech by statistically significant WER reductions of up to **21.6%** and **25.3%** absolute (**67.5%** and **83.5%** relative) across dev and test sets.

Table 2: *Performance of speaker-adaptive audio-visual end-to-end speech separation and recognition systems integrated with fine-tuned WavLM SSL features. "Sep." and "Recog." denote "separation" and "recognition". "$*$", † and "‡" represent a statistically significant WER difference (MAPSSWE, α=0.05 [45]) over the audio-only SI baseline (sys. 1), audio-visual SI baselines (sys. 4) and baseline fine-tuned WavLM model.*

Sys.	Adaptation Supervision	+Visual		+Speaker Adaptation		WER (↓)	
		Sep.	Recog.	Sep.	Recog.	Dev	Test
Fine-tuned WavLM Model by raw first channel of overlapped speech						32.0	30.3
1	SI Baseline			✗	✗	15.1	11.2
2	Enroll.	✗	✗	✓	✓	12.1*‡	7.9*‡
3	Enroll. Free			✓	✓	14.1*‡	10.1*‡
4	SI Baseline			✗	✗	12.7	7.7
5	Enroll.	✓	✓	✓	✓	10.4†‡	5.0†‡
6	Enroll. Free			✓	✓	12.5 ‡	7.0 ‡

6. Conclusions

This paper proposes joint speaker feature learning methods for zero-shot adaptation of audio-visual multichannel speech separation and recognition systems. xVector and ECAPA-TDNN speaker encoders are connected using purpose-built fusion blocks and tightly integrated with the complete system training. Experiments conducted on LRS3-TED data simulated multichannel overlapped speech suggest that joint speaker feature learning consistently improves speech separation and recognition performance over the baselines without using such. Further analyses reveal performance improvements are strongly correlated with increased inter-speaker discrimination measured using cosine similarity. Future research will focus on mitigating the sensitivity to inter-speaker angle differences for enrollment-free adaptation.

[2]Cosine similarity score is computed using the speaker features extracted from the target and interfering speaker's speech respectively.

7. Acknowledgements

This research is supported by Hong Kong RGC GRF grant No. 14200021, 14200220, TRS under Grant T45-407/19N, Innovation & Technology Fund grant No. ITS/218/21, National Natural Science Foundation of China (NSFC) Grant 62106255, and Youth Innovation Promotion Association CAS Grant 2023119.

8. References

[1] Y. Qian *et al.*, "Past review, current progress, and challenges ahead on the cocktail party problem," *FRONT INFORM TECH EL*, 2018.

[2] R. Haeb-Umbach *et al.*, "Far-field automatic speech recognition," *Proceedings of the IEEE*, 2020.

[3] X. Anguera *et al.*, "Acoustic beamforming for speaker diarization of meetings," *TASLP*, 2007.

[4] Y. Xu *et al.*, "Neural spatio-temporal beamformer for target speech separation," in *INTERSPEECH*, 2020.

[5] M. Souden *et al.*, "On optimal frequency-domain multichannel linear filtering for noise reduction," *TASLP*, 2009.

[6] E. Warsitz *et al.*, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *TASLP*, 2007.

[7] F. Bahmaninezhad *et al.*, "A comprehensive study of speech separation: spectrogram vs waveform separation," in *INTERSPEECH*, 2019.

[8] T. N. Sainath *et al.*, "Multichannel signal processing with deep neural networks for automatic speech recognition," *TASLP*, 2017.

[9] X. Xiao *et al.*, "Deep beamforming networks for multi-channel speech recognition," in *ICASSP*, 2016.

[10] T. Yoshioka *et al.*, "Multi-microphone neural speech separation for far-field multi-talker speech recognition," in *ICASSP*, 2018.

[11] J. Heymann *et al.*, "Beamnet: End-to-end training of a beamformer-supported multi-channel ASR system," in *ICASSP*, 2017.

[12] R. Gu *et al.*, "Multi-modal multi-channel target speech separation," *IEEE J-STSP*, 2020.

[13] G. Li *et al.*, "Audio-visual end-to-end multi-channel speech separation, dereverberation and recognition," *TASLP*, 2023.

[14] G. Saon *et al.*, "Speaker adaptation of neural network acoustic models using i-vectors," in *ASRU*, 2013.

[15] D. Snyder *et al.*, "X-vectors: Robust dnn embeddings for speaker recognition," in *ICASSP*, 2018.

[16] K. Veselỳ *et al.*, "Sequence summarizing neural network for speaker adaptation," in *ICASSP*, 2016.

[17] N. Kanda *et al.*, "End-to-end speaker-attributed asr with transformer," in *INTERSPEECH*, 2021.

[18] N. Kanda *et al.*, "Investigation of end-to-end speaker-attributed asr for continuous multi-talker recordings," in *SLT*, 2021.

[19] M. J. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," *Comput. Speech Lang.*, 1998.

[20] T. Ochiai *et al.*, "Speaker adaptation for multichannel end-to-end speech recognition," in *ICASSP*, 2018.

[21] P. Swietojanski *et al.*, "Learning hidden unit contributions for unsupervised acoustic model adaptation," *TASLP*, 2016.

[22] T. Ochiai *et al.*, "Speaker adaptive training using deep neural networks," in *ICASSP*, 2014.

[23] J. Deng *et al.*, "Confidence score based speaker adaptation of conformer speech recognition systems," *TASLP*, 2023.

[24] K. Žmolíková *et al.*, "Speakerbeam: Speaker aware neural network for target speaker extraction in speech mixtures," *IEEE J-STSP*, 2019.

[25] T. Ochiai *et al.*, "Multimodal speakerbeam: Single channel target speech extraction with audio-visual speaker clues." in *INTERSPEECH*, 2019.

[26] Q. Wang *et al.*, "VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking," in *INTERSPEECH*, 2019.

[27] Y. Ju *et al.*, "Tea-pse: Tencent-ethereal-audio-lab personalized speech enhancement system for icassp 2022 dns challenge," in *ICASSP*, 2022.

[28] Y. Ju *et al.*, "Tea-pse 2.0: Sub-band network for real-time personalized speech enhancement," in *SLT*, 2023.

[29] C. Xu *et al.*, "Spex: Multi-scale time domain speaker extraction network," *TASLP*, 2020.

[30] M. Ge *et al.*, "Spex+: A complete time domain speaker extraction network," in *INTERSPEECH*, 2020.

[31] S. E. Eskimez *et al.*, "Personalized speech enhancement: New models and comprehensive evaluation," in *ICASSP*, 2022.

[32] H. Taherian *et al.*, "One model to enhance them all: array geometry agnostic multi-channel personalized speech enhancement," in *ICASSP*, 2022.

[33] B. Desplanques *et al.*, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *INTERSPEECH*, 2020.

[34] T. Afouras *et al.*, "Lrs3-ted: a large-scale dataset for visual speech recognition," *arXiv preprint arXiv:1809.00496*, 2018.

[35] T. Yoshioka *et al.*, "Recognizing overlapped speech in meetings: A multichannel separation approach using neural networks," in *INTERSPEECH*, 2018.

[36] Z. Chen *et al.*, "Multi-channel overlapped speech recognition with location guided speech extraction network," in *SLT*, 2018.

[37] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *TASLP*, 2019.

[38] T. Afouras *et al.*, "Deep lip reading: a comparison of models and an online application," in *INTERSPEECH*, 2018.

[39] K. He *et al.*, "Deep residual learning for image recognition," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016.

[40] A. Gulati *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," in *INTERSPEECH*, 2020.

[41] S. Watanabe *et al.*, "Hybrid ctc/attention architecture for end-to-end speech recognition," *IEEE J-STSP*, 2017.

[42] M. Geng *et al.*, "Speaker adaptation using spectro-temporal deep features for dysarthric and elderly speech recognition," *TASLP*, 2022.

[43] S. Chen *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE J-STSP*, 2022.

[44] S. Hu *et al.*, "Exploring self-supervised pre-trained asr models for dysarthric and elderly speech recognition," in *ICASSP*, 2023.

[45] L. Gillick and S. J. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *ICASSP*, 1989.

[46] T. Ko *et al.*, "A study on data augmentation of reverberant speech for robust speech recognition," in *ICASSP*, 2017.

[47] E. A. Habets, "Room impulse response generator," *Technische Universiteit Eindhoven, Tech. Rep*, 2006.

[48] I.-T. Recommendation, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *Rec. ITU-T P. 862*, 2001.

[49] C. H. Taal *et al.*, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *TASLP*, 2011.