Interspeech 2024
1-5 September 2024, Kos, Greece

Unified Multi-Talker ASR with and without Target-speaker Enrollment

Ryo Masumura, Naoki Makishima, Tomohiro Tanaka, Mana Ihori, Naotaka Kawata, Shota Orihashi, Kazutoshi Shinoda, Taiga Yamane, Saki Mizuno, Keita Suzuki, Satoshi Suzuki, Nobukatsu Hojo, Takafumi Moriya, Atsushi Ando

NTT Corporation, Japan
ryo.masumura@ntt.com

Abstract

This paper proposes a novel multi-talker automatic speech recognition (MT-ASR) system that can perform both a target-speaker enrollment-driven process and a target-speaker-free process in a unified modeling framework. In previous studies, these two MT-ASR forms were independently modeled with unshareable parameters. However, the independent modeling cannot mutually utilize knowledge trained with different tasks. Our key idea for bridging the gap between the two forms is to introduce modeling that can regard the target-speaker-free process as the target-speaker enrollment-driven process enrolled with no target-speaker information. Therefore, our method constructs a unified autoregressive model with a removable target-speaker encoder, and its shareable model parameters are trained jointly using training datasets with and without target-speaker enrollment. Experiments demonstrated that our unified modeling significantly outperforms the independent modeling in both MT-ASR forms.
Index Terms: unified multi-talker ASR, target-speaker enrollment, unified model architecture, joint training

1. Introduction

Recognizing multi-talker overlapped monaural speech signals during conversations is a crucial automatic speech recognition (ASR) technology [1,2]. Two main forms are used in recognizing such speech signals: one is a target-speaker enrollment-driven multi-talker ASR (MT-ASR) process that only transcribes a target speaker's speech by enrolling the target speaker's information, and the other is a target-speaker-free MT-ASR process that transcribes multiple speakers' speech simultaneously without using any additional information. The former is useful for executing personalized speech commands in a cocktail-party situation or recording personalized life-logs, and the latter is useful for making meeting minutes or for the front-end of talking robots being used in public. This paper aims to jointly model these two MT-ASR forms in an unified modeling framework.

These two forms have been independently implemented using neural network based end-to-end modeling methods [3–14] that do not require signal-level speech separation [1,2]. The main advantage of the end-to-end modeling methods is to perform overall optimization for recognizing multi-talker overlapped monaural speech signals as transcriptions. In fact, these two forms are independently modeled using different architectures because they have to handle different numbers of inputs and outputs. Previous studies have introduced modeling for the target-speaker enrollment-driven ASR process. It combines a target-speaker enrollment module with autoregressive modeling [3, 4] or neural transducer modeling [5]. For the target-speaker-free ASR process, initial end-to-end modeling studies introduced modeling with multiple output branches, in which each branch generates a transcription for one speaker by considering all possible permutations of speakers [6–12]. In addition, a promising approach is autoregressive modeling in which transcriptions of multiple speakers are recursively generated from one output branch [13, 14].

However, the independent modeling cannot mutually utilize knowledge trained with different tasks, although these two forms both handle multi-talker overlapped monaural speech signals. In practice, collecting a large-scale dataset for training the target-speaker enrollment-driven MT-ASR is difficult because of the need to prepare multiple target-speaker speech sets that are different from the overlapped speech. On the other hand, dataset for training the target-speaker-free MT-ASR is easy to collect because target-speaker speech does not need to be prepared. We consider that observing various patterns of multi-talker overlapped monaural speech signals is valuable for improving the target-speaker enrollment-driven MT-ASR model. In addition, training the ability to pick out a target speaker is considered to be useful for improving the target-speaker-free MT-ASR model to distinguish between speakers. Thus, utilizing both training datasets simultaneously should yield performance improvements in both MT-ASR forms, but no study has investigated their integration.

In this paper, we propose unified MT-ASR modeling that can perform both the target-speaker enrollment-driven process and the target-speaker-free process. Our key idea for bridging the gap between the two forms is to introduce modeling that can regard the target-speaker-free process as the target-speaker enrollment-driven process enrolled with no target-speaker information. Our method takes the no target-speaker information into consideration by constructing a unified autoregressive model with a removable target-speaker encoder. The encoder is constrained not to assign any speaker characteristics to an all-ones vector. This enables us to ignore the all-ones vector in an element-wise multiplication operation introduced for the target-speaker enrollment (detailed in Section 4). The model parameters in our unified model can be trained jointly using training datasets with and without a target speaker. Therefore, it is expected that our unified modeling will outperform the independent modeling in both MT-ASR forms. In experiments using simulated multi-talker overlapped speech datasets, we show the effectiveness of the proposed method in both MT-ASR tasks.

2. Related Work

This study is related to methods that jointly model a MT-ASR task with other related tasks in target-speaker enrollment-driven MT-ASR and target-speaker-free MT-ASR. Joint modeling of transcribing target speaker's speech and non-target speakers' speech has been proposed in target-speaker enrollment-driven

10.21437/Interspeech.2024-1062

MT-ASR [15]. In target-speaker-free MT-ASR, some studies have been conducted on joint modeling with related estimation tasks. Joint modeling with gender and estimation has been studied to take into account speaker attributes [14], and joint modeling with time-stamp estimation was proposed to determine when utterances are made [16]. In addition, joint modeling with speaker identification has been examined to identify speakers beyond their utterance boundaries [17–19]. In contrast, our method jointly models target-speaker enrollment-driven MT-ASR and target-speaker-free MT-ASR in a unified model architecture. To the best of our knowledge, this paper is the first presenting a study unifying two main MT-ASR modeling methods.

3. Preliminaries

This section describes target-speaker enrollment-driven MT-ASR and target-speaker-free MT-ASR systems based on autoregressive modeling. These two systems can handle monaural multi-talker overlapped speech signals. The former generates only the target speaker's spoken text, and the latter generates all speakers' spoken text.

Target-speaker enrollment-driven MT-ASR: Target-speaker enrollment-driven MT-ASR based on autoregressive modeling [3, 4] predicts the generation probability of a target speaker's spoken text $\boldsymbol{W} = \{w_1, \cdots, w_N\}$ from monaural multi-talker overlapped speech $\boldsymbol{X}$ and from the target speaker's enrollment speech $\boldsymbol{E}$, where $w_n \in \mathcal{V}$ is the n-th token in the spoken text, N is the number of tokens in the spoken text, and $\mathcal{V}$ is the vocabulary set. In autoregressive modeling, the generation probability of $\boldsymbol{W}$ is defined as

$$P(\boldsymbol{W}|\boldsymbol{X}, \boldsymbol{E}; \boldsymbol{\Theta}_{\text{enroll}}) = \prod_{n=1}^{N} P(w_n|w_{1:n-1}, \boldsymbol{X}, \boldsymbol{E}; \boldsymbol{\Theta}_{\text{enroll}}), \tag{1}$$

where $\boldsymbol{\Theta}_{\text{enroll}}$ represents the trainable model parameter sets and $w_{1:n-1} = \{w_1, \cdots, w_{n-1}\}$. The loss function to optimize the model parameter sets is defined as

$$\mathcal{L}(\boldsymbol{\Theta}_{\text{enroll}}) = -\sum_{(\boldsymbol{X}, \boldsymbol{E}, \boldsymbol{W}) \in \mathcal{D}_{\text{enroll}}} \log P(\boldsymbol{W}|\boldsymbol{X}, \boldsymbol{E}; \boldsymbol{\Theta}_{\text{enroll}}), \tag{2}$$

where $\mathcal{D}_{\text{enroll}}$ represents a paired dataset of input multi-talker speech, the target speaker's enrollment speech, and the target speaker's spoken text. Note that spoken text becomes empty when the target speaker is not included in the multi-talker overlapped speech.

Target-speaker-free MT-ASR: Target-speaker-free MT-ASR based on autoregressive modeling [13, 14] predicts the generation probability of all speakers' spoken text $\boldsymbol{W}^{1:T} = \{\boldsymbol{W}^1, \cdots, \boldsymbol{W}^T\}$ from monaural multi-talker overlapped speech $\boldsymbol{X}$, where $\boldsymbol{W}^t = \{w_1^t, \cdots, w_{N^t}^t\}$ is the t-th speaker's spoken text, N^t is the number of tokens in the t-th speaker's spoken text, and T is the number of speakers in the multi-talker overlapped speech. Multiple spoken texts are serialized into a single token sequence to handle all of the speakers' spoken text in one autoregressive model. Thus, the generation probability of $\boldsymbol{W}^{1:T}$ is defined as

$$\begin{aligned} P(\boldsymbol{W}^{1:T}|\boldsymbol{X}; \boldsymbol{\Theta}_{\text{free}}) &= P(\boldsymbol{S}|\boldsymbol{X}; \boldsymbol{\Theta}_{\text{free}}) \\ &= \prod_{n=1}^{|\boldsymbol{S}|} P(s_n|s_{1:n-1}, \boldsymbol{X}; \boldsymbol{\Theta}_{\text{free}}), \end{aligned} \tag{3}$$

where $\boldsymbol{\Theta}_{\text{free}}$ represents the trainable model parameter sets, $\boldsymbol{S} = \{s_1, \cdots, s_{|\boldsymbol{S}|}\}$ is the serialized token sequence, and $s_n \in \{\mathcal{V} \cup \mathcal{O}\}$ is the n-th token in the serialized token sequence. $\mathcal{O} = \{[\texttt{sep}], [\texttt{eos}]\}$ represents the special token set, where $[\texttt{sep}]$ represents the change in speaker and $[\texttt{eos}]$ represents the end-of-sentence. Multiple permutations occur in the order of multiple spoken texts $\boldsymbol{W}^{1:T}$, so they are sorted by their start times, and this process is called first-in, first-out. When the speaker index t is ordered by the start time, the serialized token sequence is represented as

$$\begin{aligned} \boldsymbol{S} = \{&w_1^1, \cdots, w_{N^1}^1, [\texttt{sep}], w_1^2, \cdots, w_{N^2}^2, \\ &\cdots, w_{N^{T-1}}^{T-1}, [\texttt{sep}], w_1^T, \cdots, w_{N^T}^T, [\texttt{eos}]\}. \end{aligned} \tag{4}$$

Thus, the serialized token sequence is composed by concatenating multiple spoken texts while inserting $[\texttt{sep}]$ between them and $[\texttt{eos}]$ at the end of the entire sequence.

The loss function to optimize the model parameter sets is defined as

$$\mathcal{L}(\boldsymbol{\Theta}_{\text{free}}) = -\sum_{(\boldsymbol{X}, \boldsymbol{S}) \in \mathcal{D}_{\text{free}}} \log P(\boldsymbol{S}|\boldsymbol{X}; \boldsymbol{\Theta}_{\text{free}}), \tag{5}$$

where $\mathcal{D}_{\text{free}}$ represents a paired dataset of the serialized token sequence and the multi-talker overlapped speech.

4. Unified Multi-Talker ASR

This paper proposes unified MT-ASR modeling that can perform both the target-speaker enrollment-driven process and the target-speaker-free process. We introduce an autoregressive model that can compute both $P(w_n|w_{1:n-1}, \boldsymbol{X}, \boldsymbol{E})$, and $P(s_n|s_{1:n-1}, \boldsymbol{X})$ in a unified model architecture with a shared parameter set. Our key idea is to introduce modeling that can regard the target-speaker-free process as the target-speaker enrollment-driven process enrolled with no target-speaker information. Our model architecture is composed of a target-speaker encoder, a shared overlapped speech encoder, and a shared text decoder. The unified model can be trained jointly using training datasets with and without a target speaker; these are $\mathcal{D}_{\text{enroll}}$ and $\mathcal{D}_{\text{free}}$.

4.1. Unified model architecture

Figure 1 shows unified MT-ASR modeling with and without target-speaker enrollment. This modeling performs both a target-speaker enrollment-driven process and a target-speaker-free process.

Target-speaker encoder: The target-speaker encoder converts input acoustic features $\boldsymbol{E}$ into a target-speaker vector. The target-speaker $\boldsymbol{e}$ vector is extracted from

$$\boldsymbol{V}_{\text{co}} = \text{ConvolutionPooling}(\boldsymbol{E}; \boldsymbol{\theta}_{\text{ts}}^{\text{co}}), \tag{6}$$

$$\boldsymbol{V}_{\text{tr}} = \text{TransformerEnc}(\boldsymbol{V}_{\text{co}}; \boldsymbol{\theta}_{\text{ts}}^{\text{tr}}), \tag{7}$$

$$\boldsymbol{e}_{\text{ap}} = \text{AttentivePooling}(\boldsymbol{V}_{\text{tr}}; \boldsymbol{\theta}_{\text{ts}}^{\text{ap}}), \tag{8}$$

$$\boldsymbol{e} = \text{Linear}(\boldsymbol{e}_{\text{ap}}; \boldsymbol{\theta}_{\text{ts}}^{\text{lin}}), \tag{9}$$

where $\{\boldsymbol{\theta}_{\text{ts}}^{\text{co}}, \boldsymbol{\theta}_{\text{ts}}^{\text{tr}}, \boldsymbol{\theta}_{\text{ts}}^{\text{ap}}, \boldsymbol{\theta}_{\text{ts}}^{\text{lin}}\} = \boldsymbol{\Theta}_{\text{ts}}$ are the trainable parameters of the target-speaker encoder. $\text{ConvolutionPooling}()$ is a function composed of convolution layers and pooling layers, $\text{AddPosition}()$ is a function that adds a continuous vector in which position information is embedded, $\text{TransformerEnc}()$ is a function of the transformer encoder blocks that consist of multi-head self-attention layers and position-wise feed-forward networks [20], and $\text{Linear}()$ is a linear transformation layer.

Shared overlapped speech encoder: The role of the shared overlapped speech encoder changes depending on whether or

not there is a target speaker. It converts acoustic features of the multi-talker overlapped speech $\boldsymbol{X}$ into hidden representations $\boldsymbol{Z}$ while considering whether or not the target-speaker vector $\boldsymbol{e}$ exists. The output hidden representations $\boldsymbol{Z}$ is computed from

$$\begin{aligned} \boldsymbol{Z}^{\mathrm{co}} &= \mathrm{ConvolutionPooling}(\boldsymbol{X}; \boldsymbol{\theta}^{\mathrm{co}}_{\mathrm{enc}}), && (10) \\ \boldsymbol{Z}^{\mathrm{po}} &= \mathrm{AddPosition}(\boldsymbol{Z}^{\mathrm{co}}), && (11) \\ \boldsymbol{Z}^{\mathrm{mul}} &= \mathrm{ElementWiseMultiply}(\boldsymbol{Z}^{\mathrm{po}}, \bar{\boldsymbol{e}}), && (12) \\ \boldsymbol{Z} &= \mathrm{TransformerEnc}(\boldsymbol{Z}^{\mathrm{mul}}; \boldsymbol{\theta}^{\mathrm{tr}}_{\mathrm{enc}}), && (13) \end{aligned}$$

where $\{\boldsymbol{\theta}^{\mathrm{co}}_{\mathrm{enc}}, \boldsymbol{\theta}^{\mathrm{tr}}_{\mathrm{enc}}\} = \boldsymbol{\Theta}_{\mathrm{enc}}$ are the trainable parameters of the shared overlapped speech encoder. $\mathrm{AddPosition}()$ is a function that adds a continuous vector in which position information is embedded, and $\mathrm{ElementWiseMultiply}()$ is a function of the element-wise multiplication (Hadamard product) between two vector representations [21]. In Eq. (12), $\bar{\boldsymbol{e}}$ is defined as

$$\bar{\boldsymbol{e}} = \begin{cases} \boldsymbol{e} & \text{if target-speaker exists,} \\ \mathbf{1} & \text{else,} \end{cases} \quad (14)$$

where $\mathbf{1}$ is the all-ones vector. Thus, the target speaker vector extracted from the target-speaker encoder is used when the target speaker exists, while the all-ones vector is assigned without using the target-speaker encoder when the target speaker does not exist. In fact, in the latter case, the element-wise multiplication can be ignored because the multiplication with the all-ones vector does not change input vector representations.

Shared text decoder: The shared text decoder computes the generation probability of the n-th token from its preceding $n-1$ tokens within an utterance and from the vector representations $\boldsymbol{Z}$ extracted from the shared overlapped speech encoder. Thus, the shared text decoder handles both tokens about a target speaker's transcription and those about multiple concatenated transcriptions uttered by multiple speakers, so we redefine the preceding $n-1$ tokens as $q_{1:n-1}$ to treat both transcriptions in a unified framework. In this case, the generation probability of the n-th token is computed from

$$\begin{aligned} \boldsymbol{O}^{\mathrm{em}}_{1:n-1} &= \mathrm{Embedding}(q_{1:n-1}; \boldsymbol{\theta}^{\mathrm{em}}_{\mathrm{dec}}), && (15) \\ \boldsymbol{O}^{\mathrm{po}}_{1:n-1} &= \mathrm{AddPosition}(\boldsymbol{O}^{\mathrm{em}}_{1:n-1}), && (16) \\ \bar{\boldsymbol{o}}_n &= \mathrm{TransformerDec}(\boldsymbol{O}^{\mathrm{po}}_{1:n-1}, \boldsymbol{Z}; \boldsymbol{\theta}^{\mathrm{tr}}_{\mathrm{dec}}), && (17) \\ \boldsymbol{o}_n &= \mathrm{Softmax}(\bar{\boldsymbol{o}}_n; \boldsymbol{\theta}^{\mathrm{so}}_{\mathrm{dec}}), && (18) \end{aligned}$$

where $\{\boldsymbol{\theta}^{\mathrm{em}}_{\mathrm{dec}}, \boldsymbol{\theta}^{\mathrm{tr}}_{\mathrm{dec}}, \boldsymbol{\theta}^{\mathrm{so}}_{\mathrm{dec}}\} = \boldsymbol{\Theta}_{\mathrm{dec}}$ are the trainable parameters of the shared text decoder. $\mathrm{TransformerDec}()$ is a set of transformer decoder blocks that consist of masked multi-head self-attention layers, multi-head source-target attention layers, and position-wise feed-forward networks. $\mathrm{Softmax}()$ is a softmax layer with a linear transformation. The n-th output represents

$$\boldsymbol{o}_n = \begin{cases} P(w_n | w_{1:n-1}, \boldsymbol{X}, \boldsymbol{E}) & \text{if target-speaker exists,} \\ P(s_n | s_{1:n-1}, \boldsymbol{X}) & \text{else.} \end{cases} \quad (19)$$

Thus, the shared text decoder automatically switches operations of outputting a target-speaker's transcription or multiple concatenated transcriptions uttered from multiple speakers by capturing information embedded in $\boldsymbol{Z}$.

4.2. Joint optimization

The proposed unified autoregressive model can be trained jointly using training datasets with and without a target speaker. The

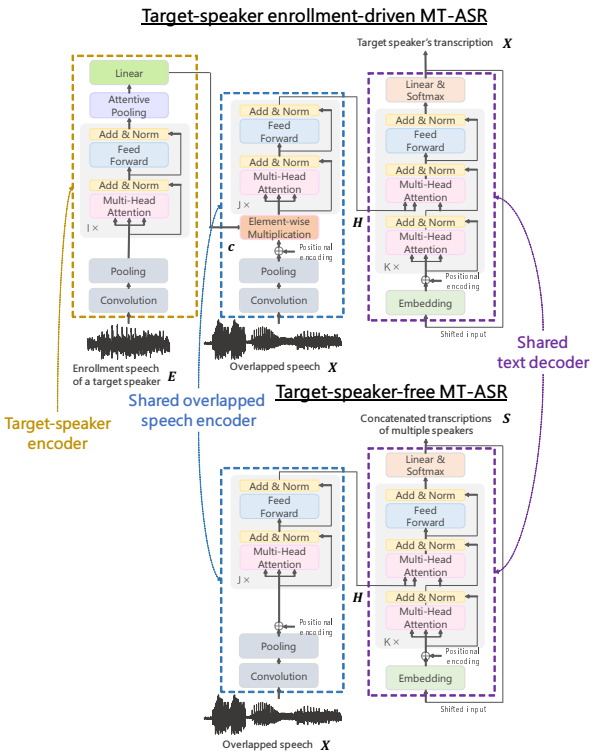

Figure 1: *Architecture of Unified Multi-Talker ASR.*

loss function to optimize the model parameter sets is defined as

$$\begin{aligned} &\mathcal{L}(\boldsymbol{\Theta}_{\mathrm{ts}}, \boldsymbol{\Theta}_{\mathrm{enc}}, \boldsymbol{\Theta}_{\mathrm{dec}}) = \\ &- \sum_{(\boldsymbol{X}, \boldsymbol{E}, \boldsymbol{W}) \in \mathcal{D}_{\mathrm{enroll}}} \log P(\boldsymbol{W} | \boldsymbol{X}, \boldsymbol{E}; \boldsymbol{\Theta}_{\mathrm{ts}}, \boldsymbol{\Theta}_{\mathrm{enc}}, \boldsymbol{\Theta}_{\mathrm{dec}}) \\ &- \sum_{(\boldsymbol{X}, \boldsymbol{S}) \in \mathcal{D}_{\mathrm{free}}} \log P(\boldsymbol{S} | \boldsymbol{X}; \boldsymbol{\Theta}_{\mathrm{enc}}, \boldsymbol{\Theta}_{\mathrm{dec}}). \quad (20) \end{aligned}$$

In mini-batch training for optimization, mini-batches are randomly chosen from either $\mathcal{D}_{\mathrm{enroll}}$ or $\mathcal{D}_{\mathrm{free}}$. In this optimization, we pre-train the target-speaker encoder with a speaker-recognition task and freeze $\boldsymbol{\theta}^{\mathrm{co}}_{\mathrm{ts}}, \boldsymbol{\theta}^{\mathrm{tr}}_{\mathrm{ts}}, \boldsymbol{\theta}^{\mathrm{ap}}_{\mathrm{ts}}$. In other words, the trainable parameters except for $\boldsymbol{\theta}^{\mathrm{lin}}_{\mathrm{ts}}$ are jointly optimized using both training datasets. Note that we can also perform fine-tuning against individual task after this joint optimization.

5. Experiments

In the experiments, we used the corpus of spontaneous Japanese [22], which involves single-talker training dataset (518.4 hours; 1,430 speakers) and a speaker-open single-talker test dataset (1.9 hours). These datasets were individually used for generating multi-talker overlapped speech dataset by mixing multiple audio signals as a monaural signal. We first randomly chose multiple audio signals so as not to select the same speakers. We set the number of speakers in the mixed signals as two or three. The original volume of each utterance was kept unchanged when mixing the audio signals, resulting in an average signal-to-interference ratio of about 0 dB. The start times of the individual utterances differed by 0.5 s or longer for the delay applied to each utterance.

Training datasets: For training of MT-ASR systems, we prepared about 2,500 hours in the training dataset for the target-speaker enrollment-driven MT-ASR and 3,500 hours in the training dataset for the target-speaker-free MT-ASR. Each of the train-

Table 1: *Evaluation results in terms of character error rate [%] on two MT-ASR tasks.*

System	*Task 1: target-speaker enrollment-driven task*		*Task 2: target-speaker-free task*		
	Single-talker 1	*Multi-talker 1*	*Single-talker 2*	*Multi-talker 2*	Avg.
Single-talker ASR	(98.80)	(92.58)	5.23	38.03	(58.66)
Target-speaker enrollment-driven MT-ASR	6.95	12.54	-	-	-
Target-speaker-free MT-ASR	(97.53)	(98.74)	4.66	10.16	(50.27)
Unified MT-ASR (Joint optimization)	**6.14**	**11.48**	**4.52**	**9.64**	**7.95**
Unified MT-ASR (+ fine-tuning on *Task 1*)	**6.03**	**11.10**	4.80	11.38	8.33
Unified MT-ASR (+ fine-tuning on *Task 2*)	9.42	14.62	**4.43**	**9.24**	9.43

ing datasets included not only multi-talker input speech but also single-talker input speech. For the target-speaker enrollment-driven task, we also prepared enrollment speech for cases where the target speaker was included in the input speech and for cases where the target speaker was not included in the input speech.

Test datasets: For evaluation on *target-speaker enrollment-driven task (Task 1)*, we prepared single-talker test dataset (*single-talker 1*; 3.8 hours) and multi-talker test dataset (*multi-talker 1*; 2.7 hours). The single-talker test dataset involved two cases: a different utterance of the target speaker was used as the speaker enrollment, and an utterance of a different speaker from the target speaker was used as the speaker enrollment. For the multi-talker test dataset, an utterance spoken by one of the talkers was used as the enrollment. In addition, for evaluation on *target-speaker-free task (task 2)*, we prepared enrollment speech-less single-talker test dataset (*single-talker 2*; 1.9 hours) and multi-talker test dataset (*multi-talker 2*; 2.7 hours).

5.1. Setups

For evaluation, we constructed a single-talker ASR system, which was trained from the single-talker training dataset without enrollment speech, a target-speaker-enrollment-driven MT-ASR system, a target-speaker-free MT-ASR system, and three unified MT-ASR systems. One unified MT-ASR system was constructed via joint optimization. Other two unified MT-ASR systems were performed fine-tuning on individual task after performing the joint optimization.

Model configurations: We introduced the model structure shown in Figure 1 for each system. The transformer blocks were designed for these systems under the following conditions: the dimensions of the output continuous representations were set to 512, the dimensions of the inner outputs were set to 2,048, and the number of heads in the multi-head attentions was set to 4. The Swish activation was used in the nonlinear transformational functions. We used 80 log mel-scale filterbank coefficients as acoustic features for both the target speaker and the overlapped speech encoders. The frame shift was 10 ms. The acoustic features passed two convolution and max pooling layers with a stride of two, so we down-sampled them to $1/4$ along with the time axis. After these layers, we stacked 8-layer transformer encoder blocks. For the text decoder, we used 512-dimensional character embeddings, where the vocabulary size was set to 3,262. We also stacked 6-layer transformer decoder blocks.

Training: For the training, we used the RAdam optimizer [23]. For the systems that used target-speaker enrollment, we trained the target-speaker encoder using the VoxCeleb2 dataset [24] with an ArcFace criterion [25], and then the overlapped speech encoder and text decoder were trained while freezing the parameters in the target-speaker encoder except for its output linear layer. We set the mini-batch size to 64 utterances and the dropout rate in the transformer blocks to 0.1. We introduced label smoothing, where its smoothing parameter was set to 0.1. In addition, we applied SpecAugment [26]. Each model training was performed on single NVIDIA A100 80G GPU.

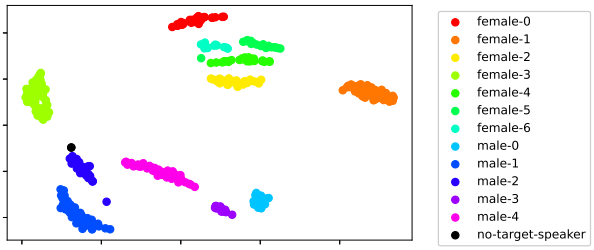

Figure 2: *Visualizations of target-speaker vectors using t-SNE. Axes labels were omitted because axes do not have a particular meaning.*

5.2. Results

Table 1 shows our evaluation of two MT-ASR tasks in terms of the character error rate (CER). Note that we ignored enrollment speech when evaluating target-speaker enrollment-driven-task using single-talker ASR or target-speaker-free MT-ASR systems.

The results show that our unified MT-ASR system outperformed the target-speaker enrollment-driven MT-ASR system and the target-speaker-free MT-ASR system in both evaluation tasks. This indicates that joint optimization of the shared model parameters using training datasets with and without target-speaker enrollment speech is effective in improving MT-ASR ability. In each test dataset, statistically significant performance improvements ($p < 0.01$) were achieved by unified MT-ASR with joint optimization. In addition, additional performance improvements were attained on either task by performing fine-tuning while averaged performance on both tasks was decreased. Furthermore, Figure 2 shows the t-SNE visualizations [27] of target-speaker vectors and a vector that represents no target speaker (all-ones vector), which were extracted after joint optimization. This shows that the vector that represents no target speaker was separated from each target-speaker set. This indicates that joint optimization that can function with or without a target speaker was correctly operated. These results demonstrate that our method effectively handles both MT-ASR tasks.

6. Conclusions

We presented unified MT-ASR modeling that can perform both the target-speaker enrollment-driven process and the target-speaker-free process. The key strength of the proposed method is its ability to construct a unified model architecture with shared model parameters using training datasets with and without target-speaker enrollment speech. The experimental results demonstrated that our unified modeling outperforms the independent modeling in both MT-ASR forms.

7. References

[1] Tae Jin Park, Naoyuki Kanda, Dimitrios Dimitriadis, Kyu J. Han, Shinji Watanabe, and Shrikanth Narayanan, "A review of speaker diarization: Recent advances with deep learning," *Computer Speech & Language*, vol. 72, pp. 101317, 2022.

[2] Katerina Zmolikova, Marc Delcroix, Tsubasa Ochiai, Keisuke Kinoshita, Jan Cernocky, and Dong Yu, "Neural target speech extraction: An overview," *IEEE Signal Processing Magazine*, vol. 40, pp. 8–29, 2023.

[3] Marc Delcroix, Shinji Watanabe, Tsubasa Ochiai, Keisuke Kinoshita, Shigeki Karita, Atsunori Ogawa, and Tomohiro Nakatani, "End-to-end speakerbeam for single channel target speech recognition," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 451–455, 2019.

[4] Pavel Denisov and Ngoc Thang Vu, "End-to-end multi-speaker speech recognition using speaker embeddings and transfer learning," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 4425–4429, 2019.

[5] Takafumi Moriya, Hiroshi Sato, Tsubasa Ochiai, Marc Delcroix, and Takahiro Shinozaki, "Streaming target-speaker ASR with neural transducer," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 2673–2677, 2022.

[6] Dong Yu, Xuankai Chang, and Yanmin Qian, "Recognizing multi-talker speech with permutation invariant training," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 2456–2460, 2017.

[7] Xuankai Chang, Yanmin Qian, Kai Yu, and Shinji Watanabe, "End-to-end monaural multi-speaker asr system without pretraining," *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 6256–6260, 2019.

[8] Hiroshi Seki, Takaaki Hori, Shinji Watanabe, Jonathan Le Roux, and John R. Hershey, "A purely end-to-end system for multi-speaker speech recognition," *In Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 2620–2630, 2018.

[9] Shane Settle, Jonathan Le Roux, Takaaki Hori, and Shinji Watanabe adn John R. Hershey, "End-to-end multi-speaker speech recognition," *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 4819–4823, 2018.

[10] Xuankai Chang, Wangyou Zhang, Yanmin Qian, Jonathan Le Roux, and Shinji Watanabe, "End-to-end multi-speaker speech recognition with transformer," *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 6129–6133, 2020.

[11] Ilya Sklyar, Anna Piunova, and Yulan Liu, "Streaming mult-speaker ASR with RNN-T," *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 6903–6907, 2021.

[12] Anshuman Tripathi, Han Lu, and Hasim Sak, "End-to-end multi-talker overlapping speech recognition," *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 6124–6128, 2020.

[13] Naoyuki Kanda, Yashesh Gaur, Xiaofei Wang, Zhong Meng, and Takuya Yoshioka, "Serialized output training for end-to-end overlapped speech recognition," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 2797–2801, 2020.

[14] Ryo Masumura, Daiki Okamura, Naoki Makishima, Mana Ihori, Akihiko Takashima, Tomohiro Tanaka, and Shota Orihashi, "Unified autoregressive modeling for joint end-to-end multi-talker overlapped speech recognition and speaker attribute estimation," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 2591–2595, 2021.

[15] Ryo Masumura, Naoki Makishima, Taiga Yamane, Yoshihiko Yamazaki, Saki Mizuno, Mana Ihori, Mihiro Uchida, Keita Suzuki, Hiroshi Sato, Tomohiro Tanaka, Akihiko Takashima, Satoshi Suzuki, Takafumi Moriya, Nobukatsu Hojo, and Atsushi Ando, "End-to-end joint target and non-target speakers ASR," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 2903–2907, 2023.

[16] Naoki Makishima, Keita Suzuki, Satoshi Suzuki, Atsushi Ando, and Ryo Masumura, "Joint autoregressive modeling of end-to-end multi-talker overlapped speech recognition and utterance-level timestamp prediction," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 2913–2917, 2023.

[17] Naoyuki Kanda, Yashesh Gaur, Xiaofei Wang, Zhong Meng, Zhuo Chen, Tianyan Zhou, and Takuya Yoshioka, "Joint speaker counting, speech recognition, and speaker identification for overlapped speech of any number of speakers," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 36–40, 2020.

[18] Fan Yu, Zhihao Du, Shiliang Zhang, Yuxiao Lin, and Lei Xie, "A comparative study on speaker-attributed automatic speech recognition in multi-party meetings," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 560–564, 2022.

[19] Guru Prakash Arumugam, Shuo-Yiin Chang, Tara N. Sainath, Rohit Prabhavalkar, and Quan Wang, "Improved long-form speech recognition by jointly modeling the primary and non-primary speakers," *In Proc. Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–8, 2023.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *In Proc. Advances in Neural Information Processing Systems (NIPS)*, pp. 5998–6008, 2017.

[21] Marc Delcroix, Katerina Zmolikova, Tsubasa Ochiai, Keisuke Kinoshita, Shoko Araki, and Tomohiro Nakatani, "Compact network for speakerbeam target speaker extraction," *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 6965–6969, 2018.

[22] Kikuo Maekawa, Hanae Koiso, Sadaoki Furui, and Hitoshi Isahara, "Spontaneous speech corpus of Japanese," *In proc. International Conference on Language Resources and Evaluation (LREC)*, pp. 947–952, 2000.

[23] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han, "On the variance of the adaptive learning rate and beyond," *In Proc. International Conference on Learning Representations (ICLR)*, 2020.

[24] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "VoxCeleb2: Deep speaker recognition," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 1086–1090, 2018.

[25] Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," *In Proc. IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR),*, pp. 4690–4699, 2018.

[26] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 2613–2617, 2019.

[27] Laurens van der Maaten and Geoffrey Hinton, "Visualizing data using t-SNE," *Journal of machine learning research*, vol. 9, pp. 2579–2605, 2008.