Capturing Dynamic Identity Features for Speaker-Adaptive Visual Speech Recognition

Sara Kashiwagi*, Keitaro Tanaka*, and Shigeo Morishima†
* Waseda University, Tokyo, Japan
E-mail: sara.kashiwagi@moegi.waseda.jp Tel/Fax: +81-3-5286-3510
† Waseda Research Institute for Science and Engineering, Tokyo, Japan

***Abstract*—This paper describes a multi-task learning method to improve speaker adaptation in visual speech recognition (VSR). VSR models are highly sensitive to variations in lip movements, resulting in degraded accuracy for speakers not encountered during training. A typical solution is to fine-tune pre-trained models with minimal data from the target speaker, but this often faces overfitting to the specific speech content of those samples. Effective speaker adaptation requires VSR models to learn both static and dynamic aspects of individual lip features independently of speech content. However, the dynamic features are time-variant and intertwined with similarly time-variant content information. To address this issue, we introduce an additional task into the fine-tuning process that encourages the model to focus on acquiring disentangled dynamic identity features. Specifically, we apply temporal transformations to the latent visual representations and input them together with the original ones into a dynamic identity discriminator. The discriminator determines whether each entry is original or transformed, where both sets of representations share the same static identity features and speech content, thereby promoting the desired speaker adaptation. Our evaluation demonstrates that the proposed method improves recognition accuracy for all speakers across two different datasets: public online speech videos and our private recordings using a smartphone camera.**

I. Introduction

Visual speech recognition (VSR), also known as lipreading, aims to recognize speech content solely based on lip movements [1]. VSR has attracted attention as an alternative to traditional automatic speech recognition (ASR), particularly in noisy environments or situations where speaking aloud is impractical. Additionally, VSR can be a valuable communication tool for those with hearing or speech impairments. In practical applications, VSR has already been implemented in systems known as silent speech interfaces (SSIs) [2], which enable users to input text without vocalization. To ensure high accuracy in these systems, it is crucial to develop models that are tailored to individual users.

One of the major challenges in VSR is the decreased recognition accuracy for speakers not included in the training data. For example, the word error rate (WER) of a VSR model [3] is 18.0% for speakers seen during training but rises to 30.5% for unseen speakers. This performance drop occurs due to the model's sensitivity to the diverse lip movements of different individuals. The most effective approach to adapt VSR models to unseen speakers is to perform parameter re-optimization (i.e., fine-tuning) on the entire model using a substantial amount of data from these speakers. However, in real-world scenarios, gathering sufficient data for this fine-tuning is challenging, as manual data collection places a significant burden on users.

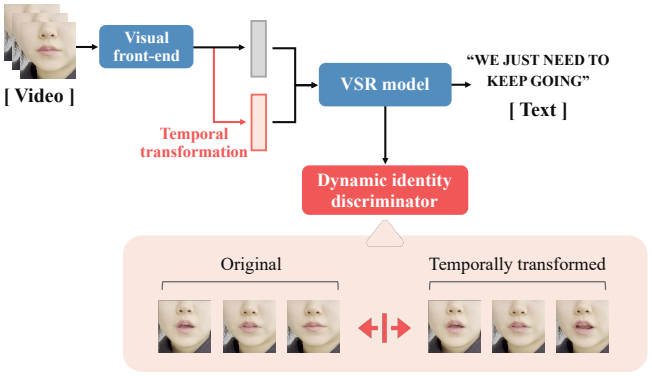

Fig. 1. Our multi-task learning approach employs temporal transformation and a dynamic identity discriminator to adapt VSR models to speaker-specific lip movements independently of speech content.

One remedy to this issue is applying parameter-efficient fine-tuning methods, particularly low-rank adaptation (LoRA) [4]. These methods leverage the extensive knowledge of pre-trained models and require fewer training samples compared to full parameter fine-tuning. While LoRA has proven effective in VSR [5] to some extent, it can lead to undesired learning behaviors when only minimal data from the target speaker is available. In speech videos, the speaker-dependent identity and speaker-independent content information are visually entangled, making it challenging to extract the information required for the target task [6]. Consequently, while LoRA can effectively capture the target speaker's identity with sufficient additional training samples, it tends to overfit the speech content of these samples under minimal data conditions.

In this paper, we present a novel multi-task learning approach that compels the VSR model to focus on the dynamic aspects of lip features (see Fig. 1). Generally, identity information includes two aspects: static (spatial) features related to lip appearance or shape and dynamic (temporal) features related to speaking speed or rhythm [5]. For effective speaker adaptation, the model needs to learn both features independently of speech content. Since the static identity features are time-invariant and unaffected by time-variant content information, the model can naturally acquire them during fine-tuning. Conversely, the dynamic identity features are time-variant and often emerge

alongside content information, such as the duration of pronunciation or the linking of sounds. Therefore, it is necessary to help the model acquire disentangled dynamic identity features in parallel with predicting speech content.

To perform effective speaker adaptation while preserving the robustness of pre-trained models against diverse speech content, we introduce an additional task into the fine-tuning process with generalized low-rank adaptation (GLoRA) [7]. Specifically, our method updates trainable low-rank decomposition matrices such that the model simultaneously recognizes dynamic identity features and produces accurate VSR results. The additional task incorporates temporal transformation and dynamic identity discrimination. The temporal transformation introduces local variations into the original dynamic identity features by randomly adding or removing frames of the latent visual representations with a certain probability. The dynamic identity discriminator receives the embeddings of both the original and transformed latent visual representations and determines whether each embedding entry is derived from the original or transformed representations. Since both sets of representations share the same static identity features and speech content, the model must focus on the target speaker's unique dynamic identity features to accomplish the additional task, thereby promoting the desired speaker adaptation.

Our multi-task learning approach for speaker adaptation prevents VSR models from overfitting speech content without requiring additional data collection. It facilitates efficient learning of speaker-specific lip features and promotes model specialization through targeted adjustments rather than relying on data augmentation or network architecture changes. As a result, our approach can be applied to various existing VSR models. Experimental results show that our method consistently improves accuracy for unseen speakers, both in a publicly available dataset and a custom dataset collected using a smartphone camera to simulate SSI scenarios. Notably, for some speakers, our proposed model trained with 80 samples even outperforms the ordinarily fine-tuned model with 100 samples.

II. Related Work

This section reviews existing research on model adaptation for VSR to handle speakers absent from the training data. We also explore the components of lip feature representation and their effects on speech recognition accuracy.

A. Speaker Adaptation in VSR

Speaker adaptation methods for VSR models have primarily focused on parameter re-optimization. Early research typically adjusted all model parameters using training data that included multiple target speakers. For example, some studies [6], [8] presented a multi-task learning approach that integrated the speaker identification task with VSR. Another approach by Kandala et al. [9] proposed utilizing speaker-specific latent identity vectors as additional inputs to the hidden layers of a VSR model. However, these approaches resulted in significant computational and data collection burdens. To address these issues, researchers have developed parameter-efficient fine-tuning methods that utilize samples from only a single target speaker. Examples include techniques such as input video transformation [10], prompt tuning [11], and low-rank adaptation (LoRA) [5]. These approaches enable speaker adaptation with a small amount of data by leveraging the knowledge embedded in existing pre-trained models.

Despite these advancements, there remains a significant lack of publicly available VSR datasets with adequate speaker labels. As a result, previous research on speaker adaptation has had to rely on datasets that share a limited vocabulary across the training, validation, and test data, such as the LRW dataset [12] and the GRID dataset [13]. Moreover, these studies have focused on classifying input videos into word classes present in the training data. Reflecting current progress in ASR, recent VSR models have shifted towards predicting sentences with unrestricted vocabularies through sub-word classification [14]. In this study, we advance speaker adaptation from word classification to sentence prediction by incorporating sub-word context through connectionist temporal classification (CTC) and attention mechanisms [15]. We also use two datasets with unrestricted vocabularies for our experiments: the publicly available VoxCeleb2 dataset [16] and a custom dataset developed to simulate real-world VSR applications.

B. Lip Feature Representation

Lip feature representation in VSR contains both identity and content information. This nature complicates the process of extracting and applying the information required for specific objectives, especially when the number of training samples is limited. For example, identity information can interfere with text prediction in VSR models, and content information is irrelevant for speaker adaptation. Existing studies on speaker adaptation have primarily addressed this issue by decoupling identity information from content information [6], [8].

Furthermore, He et al. [5] observed that identity information comprises both static (spatial) and dynamic (temporal) features. Static features are related to physiological aspects such as lip appearance and shape, while dynamic features are associated with behavioral aspects such as speaking speed and rhythm. In VSR, variations in lip opening widths caused by different voicing methods have led to performance gaps [17]. Specifically, larger lip openings in unvoiced speech resulted in an 8.5% decrease in accuracy compared to voiced speech, indicating that static identity features substantially affect VSR performance. Additionally, in both ASR and VSR, Pandey and Arif [18] demonstrated the relationship between speaking speed and recognition accuracy. Their results consistently show that native English speakers have lower accuracy than non-native speakers, attributed to their faster speaking speed. This performance decline is caused by shorter pronunciation durations and the linking of adjacent sounds, highlighting the importance of dynamic identity features in recognition tasks [19]. These findings suggest that adapting to both the static and dynamic identity features of the target speaker can contribute to improved recognition accuracy.

III. Proposed Method

In this study, we apply a parameter-efficient fine-tuning method to sub-word-based sentence prediction models. First, we explain the model architecture and the fine-tuning method. Then, an additional task is introduced to mitigate undesired overfitting to the content information.

A. End-to-End VSR Model

As the end-to-end model for predicting sub-word sequences from speech videos, we adopt a CTC/Attention hybrid model based on the conformer architecture [15], as shown in Fig. 2. This model employs an encoder-decoder architecture, with the encoder consisting of front-end and back-end networks. The input video is cropped to a 96×96 bounding box of the lip region using landmarks detected by MediaPipe [20]. The visual front-end comprises a 2D ResNet-18, where the initial layer is replaced with a 3D convolutional layer to extract both spatial and temporal features as latent visual representations from the input. The back-end is a 12-layer conformer encoder that learns temporal dependencies among the latent visual representations and outputs 768-dimensional embeddings per frame.

To leverage both CTC and attention mechanisms, the decoder is divided into two parallel fully connected (FC) layers. One processes the raw embeddings, while the other handles the 256-dimensional outputs from a 6-layer transformer decoder. The ground truth text labels consist of 5,000 sub-word classes generated by SentencePiece [21], and each FC layer produces a probability distribution over these classes. Finally, beam search with a width of 40 is applied to the weighted sum of the output probabilities from the CTC and attention branches, generating the predicted sub-word sequence, i.e., the sentence. The loss function $\mathcal{L}_{\mathrm{VSR}}$, which combines the CTC loss $p_{\mathrm{CTC}}(\mathbf{y}|\mathbf{x})$ and the cross-entropy loss $p_{\mathrm{CE}}(\mathbf{y}|\mathbf{x})$, is employed for training the model. $\mathcal{L}_{\mathrm{VSR}}$ is expressed as follows:

$$\mathcal{L}_{\mathrm{VSR}} = \alpha \log p_{\mathrm{CTC}}(\mathbf{y}|\mathbf{x}) + (1-\alpha) \log p_{\mathrm{CE}}(\mathbf{y}|\mathbf{x}), \quad (1)$$

where $\alpha = 0.1$ is used in the evaluation experiments.

B. Generalized Low-Rank Adaptation (GLoRA)

Given the above-mentioned VSR model pre-trained on large-scale datasets, we consider fine-tuning it with only minimal data from the target speaker. Specifically, we aim to facilitate the model's learning of speaker-specific lip features while preserving the linguistic knowledge of the pre-trained model. For this purpose, we apply generalized low-rank adaptation (GLoRA) [7] to the attention layers of the 12-layer conformer encoder, keeping all other parameters fixed.

GLoRA is an advanced version of LoRA [4], designed for universal parameter-efficient fine-tuning. Let $\boldsymbol{x} \in \mathbb{R}^{H\times 1}$ and $f : \mathbb{R}^{H\times 1} \to \mathbb{R}^{H\times 1}$ be the input and a linear transformation, where H denotes the input dimension. The unified equation representing all trainable parameters is described as follows:

$$f(\boldsymbol{x}) = (\boldsymbol{W}_0 + \boldsymbol{W}_0\boldsymbol{A} + \boldsymbol{B})\boldsymbol{x} + \boldsymbol{W}_0\boldsymbol{C} + \boldsymbol{b}_0 \odot \boldsymbol{D} + \boldsymbol{E} + \boldsymbol{b}_0, \quad (2)$$

where $\odot$ denotes the element-wise product. $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{H\times H}$, $\boldsymbol{C}, \boldsymbol{D}$ and $\boldsymbol{E} \in \mathbb{R}^{H\times 1}$ are the trainable parameters, while the pre-trained weights $\boldsymbol{W}_0 \in \mathbb{R}^{H\times H}$ and bias $\boldsymbol{b}_0 \in \mathbb{R}^{H\times 1}$ remain frozen throughout the fine-tuning process. $\boldsymbol{A}$ scales the pre-trained weight, decomposed into $\boldsymbol{A}_{\mathrm{d}} \in \mathbb{R}^{H\times r}$ and $\boldsymbol{A}_{\mathrm{u}} \in \mathbb{R}^{r\times H}$, where r represents the rank of LoRA. $\boldsymbol{B}$ scales the input and shifts the pre-trained weight, decomposed into $\boldsymbol{B}_{\mathrm{d}} \in \mathbb{R}^{H\times r}$ and $\boldsymbol{B}_{\mathrm{u}} \in \mathbb{R}^{r\times H}$. $\boldsymbol{C}$ acts as the layer-wise manipulation, decomposed into $\boldsymbol{C}_{\mathrm{d}} \in \mathbb{R}^{H\times r}$ and $\boldsymbol{C}_{\mathrm{u}} \in \mathbb{R}^{r\times 1}$. $\boldsymbol{D}$ and $\boldsymbol{E}$ scale and shift the pre-trained bias. We use a random Gaussian initialization for $\boldsymbol{A}_{\mathrm{u}}$, $\boldsymbol{B}_{\mathrm{u}}$, and $\boldsymbol{C}_{\mathrm{u}}$, and zero initialization for $\boldsymbol{A}_{\mathrm{d}}$, $\boldsymbol{B}_{\mathrm{d}}$, and $\boldsymbol{C}_{\mathrm{d}}$, ensuring $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$ are zeros at the beginning of the training, as well as $\boldsymbol{D}$ and $\boldsymbol{E}$.

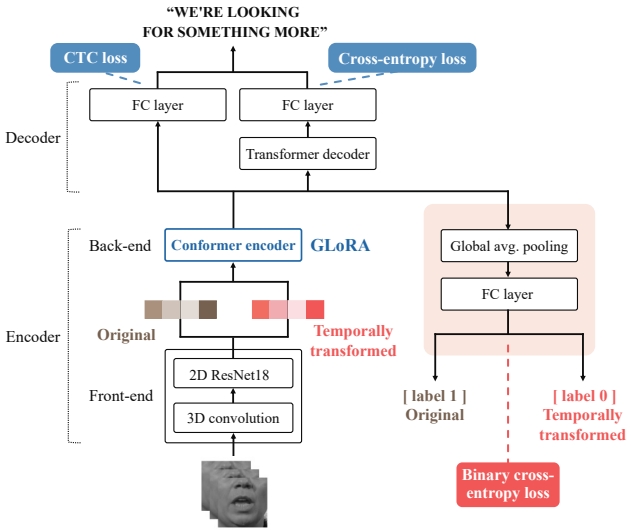

Fig. 2. The overall architecture of the proposed multi-task learning approach. We incorporate a discrimination task that distinguishes between the original and temporally transformed latent visual representations into the output from the conformer encoder, where GLoRA is applied.

C. Multi-Task Learning

Although the model specialized for individual speakers needs to learn both static and dynamic identity features independently of speech content, the dynamic identity features are time-variant and can overlap with similarly time-variant content information. To facilitate the model's learning of disentangled dynamic identity features in parallel with predicting speech content, we introduce an additional task into the fine-tuning process. This task involves temporal transformation and dynamic identity discrimination, as illustrated in Fig. 2.

The temporal transformation introduces local variations into the original dynamic identity features. As illustrated in Fig. 3, this is achieved by randomly adding or removing frames of the latent visual representations produced by the visual front-end, with a probability of 0.2. This process alters the sequence shape from $T \times 512$ to $T' \times 512$, where T and T' represent the number of frames before and after the transformation, respectively. Subsequently, the sequence is realigned to the original shape of $T \times 512$ by either zero-padding or cutting the end of the sequence. By using both the original and transformed versions of each sequence during training, the number of training sequences is doubled from N to $2N$, where N is the number of original data inputs.

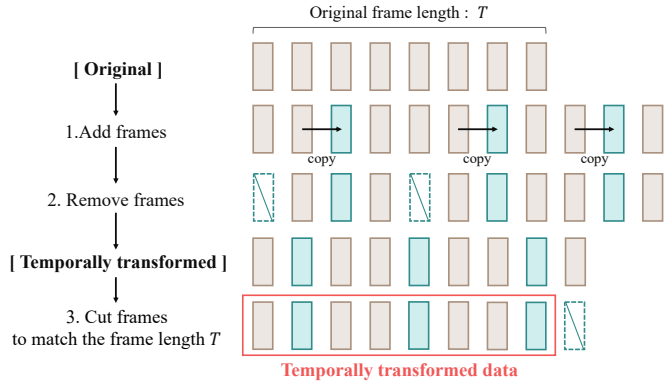

Fig. 3. Example of the temporal transformation process applied to the latent visual representations obtained from the visual front-end.

We then apply a discriminator to the output from the 12th layer of the conformer encoder to enable the model to simultaneously recognize dynamic identity features and generate accurate VSR results by updating GLoRA parameters. The discriminator consists of a global average pooling layer along the temporal axis and an FC layer. The global average pooling layer creates a 768-dimensional embedding for each sample to represent global information about the dynamic identity features. The FC layer takes this embedding and determines whether each embedding entry is derived from the original or transformed representations. Minimizing the binary cross-entropy loss $\mathcal{L}_{\mathrm{BCE}}$ compels the VSR model to focus on the target speaker's unique dynamic identity features, as both sets of representations share common static identity features and speech content. Given the ground truth labels y_i, where $y_i = 1$ for original data and $y_i = 0$ for transformed data, and the predicted probabilities $\hat{y}_i$, $\mathcal{L}_{\mathrm{BCE}}$ can be formulated as

$$\mathcal{L}_{\mathrm{BCE}} = -\frac{1}{2N}\sum_{i=1}^{2N}\left[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right]. \quad (3)$$

The overall loss in our multi-task learning method is given by:

$$\mathcal{L} = \mathcal{L}_{\mathrm{VSR}} + \mathcal{L}_{\mathrm{BCE}}. \quad (4)$$

IV. Evaluation

This section describes experiments conducted to evaluate the performance of our proposed method for speaker adaptation.

A. Dataset

In our experiments, we used two datasets with unrestricted vocabularies: the VoxCeleb2 dataset [16] and a custom dataset. VoxCeleb2 is a publicly available dataset collected from online videos, originally designed for the speaker recognition task. Therefore, it includes speaker identity labels but lacks ground truth text labels. Here, we used its audio data to automatically generate transcriptions for the text labels, leveraging Whisper models [22]. For the speaker adaptation settings, we selected the five speakers with the most utterances from the test data (S1-S5). The speaker IDs and the size of the training, validation, and test sets per speaker are shown in Table I.

TABLE I
SPEAKER INFORMATION IN THE VOXCELEB2 DATASET

Speaker	ID (`folder_name`)	#Train	#Valid	#Test
S1	id05816 (`SCbhADNyfwg`)	70	30	105
S2	id07874 (`_3D5upGq24U`)	70	30	58
S3	id06692 (`9vs0zAHfI0M`)	56	24	62
S4	id08696 (`Cg9tYEiOCQA`)	56	24	54
S5	id04232 (`tCiPy0q5588`)	42	18	45

To gather the custom dataset, we developed a tailored app using a smartphone camera to simulate real-world VSR conditions. In the recording interface, the text to be read is displayed at the top of the screen, and video is recorded while a button at the bottom is pressed. Five participants (S6-S10), who are native or have equivalent English proficiency, took part in the data collection. We used 300 speech texts from the LRS3 test data [23], with each participant reading each text once. In this setup, S6-S10 all read the same texts, while S1-S5 in VoxCeleb2 each utter different sentences. Out of these 300 samples, 200 were fixed as the test set. To evaluate the impact of adaptation data quantity on performance, we created three different sets from the remaining 100 samples: 100 samples (70 for training and 30 for validation), 80 samples (56 and 24), and 60 samples (42 and 18). The experiments were conducted five times for each set, with a new random split of the training and validation data each time.

B. Training Settings

In this study, to leverage the knowledge from large-scale datasets, we utilized two pre-trained models provided by Ma et al. [24]. For the VoxCeleb2 evaluation, we used the model trained on 438 hours of the LRS3 dataset. For the custom dataset evaluation, we used the model trained on 1,759 hours of the combined LRS3 and VoxCeleb2 datasets. In all experiments, we trained the model for 100 epochs using the AdamW optimizer [25] with a learning rate of 1.0×10^{-5} and a maximum of 1,800 frames per batch. We thoroughly explored parameters and selected the configurations that achieved the best performance. For the rank of GLoRA, we tested $r = 4$ and 8 based on the findings of [26] and set it to 8. In the inference phase, the model was created by averaging the weights of the three best-performing checkpoints based on validation loss.

C. Evaluation Metrics

For evaluation, we used word error rate (WER) for the model outputs. We calculated WER as follows:

$$\mathrm{WER} = \frac{W_S + W_D + W_I}{W_N}, \quad (5)$$

where W_N is the total number of words in the ground truth. W_S, W_I, and W_D represent the number of words substituted for wrong classifications, inserted for those not to be present, and deleted for those to be present, respectively. A lower score indicates higher recognition accuracy.

D. Experimental Results

We conducted experiments with a total of 10 speakers using two different datasets to evaluate our proposed method. The

TABLE II
PERFORMANCE COMPARISON OF DIFFERENT METHODS FOR SPEAKERS IN THE VOXCELEB2 DATASET. THE BASELINE REFERS TO THE PRE-TRAINED MODEL, FT REPRESENTS FINE-TUNING WITH GLORA, AND MT DENOTES THE PROPOSED MULTI-TASK LEARNING METHOD

Methods	WER (%)				
	S1	S2	S3	S4	S5
Baseline	63.50	66.69	47.76	76.67	63.59
FT	61.74	61.75	44.54	74.76	64.10
+ MT	**61.17**	**57.55**	**44.19**	**73.95**	**60.78**

TABLE III
PERFORMANCE COMPARISON OF DIFFERENT METHODS FOR SPEAKERS IN THE CUSTOM DATASET

#Data	Methods	WER (%)				
		S6	S7	S8	S9	S10
0	Baseline	31.20	16.41	21.26	71.30	58.45
60	FT	30.32	16.90	20.39	68.55	56.78
60	+ MT	**29.09**	16.69	**20.02**	**64.49**	**55.88**
80	FT	28.89	15.72	20.45	60.60	53.18
80	+ MT	**27.31**	**15.51**	**19.64**	**60.11**	**52.45**
100	FT	28.23	15.59	19.81	59.79	52.21
100	+ MT	**26.27**	**15.07**	**18.85**	**58.29**	**52.14**

average WER over five experiments is presented in Table II for the VoxCeleb2 dataset and in Table III for the custom dataset. Table II shows that our proposed method improved accuracy for all speakers. Specifically, the average WER among the five speakers (S1-S5) decreased by 1.83% compared to the regular fine-tuning results and by 4.09% compared to the baseline through our multi-task learning method. For S5, although fine-tuning resulted in lower accuracy than the baseline, our approach effectively mitigated overfitting and achieved an almost 3% decrease in WER.

Table III compares the performance of the fine-tuning and multi-task learning methods across different sample sizes, showing that the multi-task learning method consistently outperformed the regular fine-tuning method. Using 100, 80, and 60 samples for training and validation, the average WER among the five speakers (S6-S10) decreased by 1.00%, 0.76%, and 1.35% compared to the fine-tuning results. The WER reductions compared to the baseline results were 5.60%, 4.72%, and 2.49%, respectively. Furthermore, for S6-S8, the multi-task learning method with 80 samples achieved better accuracy than the fine-tuning method even when using 100 samples. These results demonstrate that our approach promotes more efficient utilization of the target speaker data.

Next, we discuss the variations in recognition accuracy from two perspectives: environmental differences between the VoxCeleb2 and custom datasets and individual differences among speakers within the custom dataset. In the VoxCeleb2 evaluation, both the pre-trained dataset (LRS3) and the fine-tuning dataset (VoxCeleb2) consist of public speech sourced from the Internet. In contrast, the custom dataset evaluation involves significant environmental differences, as the fine-tuning dataset was collected using a smartphone camera in private settings, differing from the pre-trained datasets (LRS3 and VoxCeleb2). Our multi-task learning approach proves effective for both datasets, demonstrating robustness to various environments. Focusing on Table III, despite identical speech content across different speakers, notable performance variations were observed due to individual differences in lip movements. Additionally, speakers with poorer baseline performance exhibited greater accuracy improvements through the multi-task learning method. These results show that our approach can mitigate undesired learning behaviors and adapt the VSR model to the diverse lip features of individual speakers.

TABLE IV
ABLATION STUDY RESULTS FOR SPEAKERS IN THE CUSTOM DATASET. TT DENOTES THE TEMPORAL TRANSFORMATION PROCESS

Methods	WER (%)				
	S6	S7	S8	S9	S10
Baseline	31.20	16.41	21.26	71.30	58.45
FT	28.23	15.59	19.81	59.79	52.21
+ TT	27.12	16.11	19.59	59.29	53.13
+ TT + MT (6th layer)	**26.10**	15.21	18.90	58.69	52.37
+ TT + MT (12th layer)	26.27	**15.07**	**18.85**	**58.29**	**52.14**

Table IV presents the results of the ablation study conducted on the custom dataset using 100 samples for training and validation. We assessed accuracy changes under the following conditions: (1) applying only the temporal transformation process without the discriminator, (2) applying the discriminator to the output from the 6th layer of the conformer encoder, and (3) applying the discriminator to the output from the 12th layer of the conformer encoder (as shown in Table III). The effect of the temporal transformation varied among speakers. It improved accuracy for S6, S8, and S9 compared to the fine-tuning results, but S7 and S10 experienced a decline. For these latter speakers, using temporally transformed data without the discriminator might have introduced noise into the learning process of the target speaker's identity features. Applying the discriminator to the 6th layer consistently achieved higher accuracy than using only the temporal transformation process. There was no significant difference between applying the discriminator to the 6th and 12th layers; still, a slight improvement in accuracy was observed when applying it to the 12th layer, excluding S1.

V. CONCLUSION

We proposed a novel speaker adaptation method to improve accuracy for speakers not included in the training data using only minimal data. Specifically, we presented a multi-task learning approach to suppress overfitting to the specific speech content and promote efficient learning of disentangled dynamic identity features. Through extensive experiments with various speakers, we demonstrated that our proposed method consistently outperformed conventional fine-tuning methods across datasets from two different scenarios, even under challenging conditions. Our future work includes applying the proposed method to various existing VSR models and examining its effectiveness for speakers with more diverse lip features.

Acknowledgment

This work is supported in part by JSPS KAKENHI Nos. 21H05054, 24H00742, and 24H00748.

References

[1] P. Ma, S. Petridis, and M. Pantic, "Visual speech recognition for multiple languages in the wild," *Nature Machine Intelligence*, vol. 4, no. 11, pp. 930–939, 2022.

[2] Z. Su, S. Fang, and J. Rekimoto, "LipLearner: Customizable silent speech interactions on mobile devices," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–21.

[3] B. Shi, W.-N. Hsu, K. Lakhotia, and A. Mohamed, "Learning audio-visual speech representation by masked multimodal cluster prediction," *International Conference on Learning Representations (ICLR)*, 2022.

[4] E. J. Hu, Y. Shen, P. Wallis, *et al.*, "LoRA: Low-rank adaptation of large language models," *International Conference on Learning Representations (ICLR)*, 2022.

[5] Y. He, L. Yang, H. Wang, Y. Zhu, and S. Wang, "Speaker-adaptive lipreading via spatio-temporal information learning," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2024, pp. 10 411–10 415.

[6] Q. Zhang, S. Wang, and G. Chen, "Speaker-independent lipreading by disentangled representation learning," in *IEEE International Conference on Image Processing (ICIP)*, 2021, pp. 2493–2497.

[7] A. Chavan, Z. Liu, D. Gupta, E. Xing, and Z. Shen, "One-for-All: Generalized LoRA for parameter-efficient fine-tuning," *arXiv preprint arXiv:2306.07967*, 2023.

[8] M. Wand and J. Schmidhuber, "Improving speaker-independent lipreading with domain-adversarial training," in *Interspeech*, 2017, pp. 3662–3666.

[9] P. A. Kandala, A. Thanda, D. K. Margam, *et al.*, "Speaker adaptation for lip-reading using visual identity vectors," in *Interspeech*, 2019, pp. 2758–2762.

[10] M. Kim, H. Kim, and Y. M. Ro, "Speaker-adaptive lip reading with user-dependent padding," in *European Conference on Computer Vision (ECCV)*, 2022, pp. 576–593.

[11] M. Kim, H.-I. Kim, and Y. M. Ro, "Prompt tuning of deep neural networks for speaker-adaptive visual speech recognition," *arXiv preprint arXiv:2302.08102*, 2023.

[12] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Asian Conference on Computer Vision (ACCV)*, 2016, pp. 6447–6456.

[13] M. Cooke, J. Barker, S. P. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120 5 Pt 1, pp. 2421–4, 2006.

[14] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6447–6456.

[15] P. Ma, S. Petridis, and M. Pantic, "End-to-end audio-visual speech recognition with conformers," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2021, pp. 7613–7617.

[16] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Interspeech*, 2018, pp. 1086–1090.

[17] S. Petridis, J. Shen, D. Cetin, and M. Pantic, "Visual-only recognition of normal, whispered and silent speech," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018, pp. 6219–6223.

[18] L. Pandey and A. S. Arif, "Effects of speaking rate on speech and silent speech recognition," in *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–8.

[19] L.-F. Lai and N. Holliday, "Exploring sources of racial bias in automatic speech recognition through the lens of rhythmic variation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2023, pp. 1284–1288.

[20] C. Lugaresi, J. Tang, H. Nash, *et al.*, "MediaPipe: A framework for building perception pipelines," in *Third Workshop on Computer Vision for AR/VR at IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[21] T. Kudo, "Subword regularization: Improving neural network translation models with multiple subword candidates," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018, pp. 66–75.

[22] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning (ICML)*, 2023, pp. 28 492–28 518.

[23] T. Afouras, J. S. Chung, and A. Zisserman, "LRS3-TED: A large-scale dataset for visual speech recognition," *arXiv preprint arXiv:1809.00496*, 2018.

[24] P. Ma, A. Haliassos, A. Fernandez-Lopez, H. Chen, S. Petridis, and M. Pantic, "Auto-AVSR: Audio-visual speech recognition with automatic labels," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[25] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *International Conference on Learning Representations (ICLR)*, 2019.

[26] A. Baby, G. Joseph, and S. Singh, "Robust speaker personalisation using generalized low-rank adaptation for automatic speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2024, pp. 11 381–11 385.