Learning Separable Hidden Unit Contributions for Speaker-Adaptive Lip-Reading

Songtao Luo[1,2]
luosongtao18@mails.ucas.ac.cn
Shuang Yang[1,2]
shuang.yang@ict.ac.cn
Shiguang Shan [1,2]
sgshan@ict.ac.cn
Xilin Chen[1,2]
xlchen@ict.ac.cn

[1] Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS,
Beijing, 100190, China

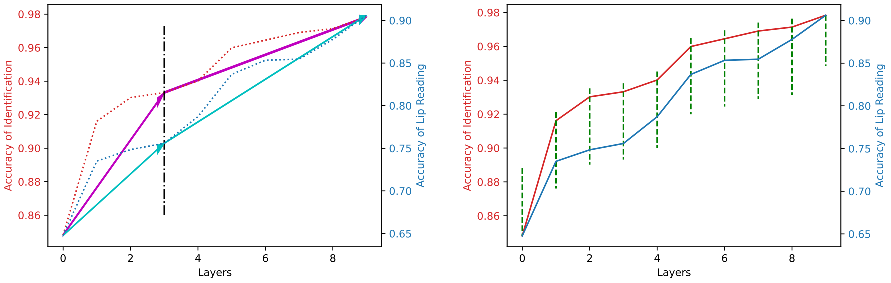
[2] University of Chinese Academy of Sciences,
Beijing, 100049, China

Abstract

In this paper, we propose a novel method for speaker adaptation in lip reading, motivated by two observations. Firstly, a speaker's own characteristics can always be portrayed well by his/her few facial images or even a single image with shallow networks, while the fine-grained dynamic features associated with speech content expressed by the talking face always need deep sequential networks to represent accurately. Therefore, we treat the shallow and deep layers differently for speaker adaptive lip reading. Secondly, we observe that a speaker's unique characteristics (e.g. prominent oral cavity and mandible) have varied effects on lip reading performance for different words and pronunciations, necessitating adaptive enhancement or suppression of the features for robust lip reading. Based on these two observations, we propose to take advantage of the speaker's own characteristics to automatically learn separable hidden unit contributions with different targets for shallow layers and deep layers respectively. For shallow layers where features related to the speaker's characteristics are stronger than the speech content related features, we introduce speaker-adaptive features to learn for enhancing the speech content features. For deep layers where both the speaker's features and the speech content features are all expressed well, we introduce the speaker-adaptive features to learn for suppressing the speech content irrelevant noise for robust lip reading. Our approach consistently outperforms existing methods, as confirmed by comprehensive analysis and comparison across different settings. Besides the evaluation on the popular LRW-ID and GRID datasets, we also release a new dataset for evaluation, CAS-VSR-S68, to further assess the performance in an extreme setting where just a few speakers are available but the speech content covers a large and diversified range. The results demonstrated our method's superiority on this challenging dataset as well.

1 Introduction

Lip reading, or Visual Speech Recognition (VSR), is a challenging task that aims to interpret the spoken content by analyzing visual cues of a speaker's lip or face movements. Thanks to



(a) Overall Growth Trend (b) Layer-by-layer Growth Trend

Figure 1: Accuracy of Lip Reading and Identification Using the Output at Different Layers

the emergence of large-scale lip reading datasets [1, 2, 9, 10, 49] and pre-training methods [6, 16, 30, 38], the domain of lip reading has achieved significant progress recently. Currently, the performance of lip reading has reached a level where it can rival the performance of audio-based speech recognition models from four years ago[8]. However, lip reading still encounters various challenges, especially the diverse speaking styles and facial appearance of different speakers, which severely limits the practical application of lip reading in the real world.

By observing a speaker's characteristics, we find that most of the speaker's characteristics, such as his/her facial structure, mouth shape, skin texture, skin tone, beard, glasses, markings, and other facial traits, are usually static and would not be significantly affected by pronouncing different words. This property enables the corresponding speaker to be easily be distinguished by checking the speaker's few facial images or even a single image with shallow networks. However, when we are speaking, our facial region, especially the lip region, is always in a state of motion during the whole speaking process. The dynamics involved in the process of speaking are typically fine-grained spatio-temporal changes, which require deep sequential networks to obtain good representations.

On the other hand, although most of the speaker's characteristics, are static and not affected by pronouncing different speech words, there exist some attributes which can enhance or weaken the facial dynamic information during speech production. For example, for a speaker with lips that turn outward, compared to others, the lip movement for plosive sounds will be weakened (due to the relatively small lip movement distance), while the lip movement for fricative sounds will be enhanced (due to relatively large lip movement distance). Therefore, we propose to leverage the speaker's characteristics for learning to enhance and suppress separate hidden unit contributions in the shallow layers and deep layers respectively.

In addition to the aforementioned observation, we also visualize features at different levels of a general lip reading model [40, 41] and use them for speaker identification and lip reading tasks, respectively. As shown in Figure 1, the comparison of accuracy of shallow and deep layers, as well as the difference in performance improvement as the network depth increases, further confirms the above observation.

Based on the above analysis, we propose a novel speaker adaptive approach for lip reading by learning separable hidden unit contributions. We enhance the content-dependent features in shallow layers and suppress content-independent features in deeper layers, respectively. By adaptively enhancing content-dependent features with the speaker's features, we guide the model to focus on capturing content-dependent features which would be transferred to deeper layers to benefit the final lip reading task. Conversely, in deep layers where content-dependent features have obtained good representations, we introduce the speaker's feature to adaptively suppress the content-independent features, allowing the model to rely more on speech content-dependent features for robust lip reading.

Our contributions can be summarized as follows: 1) We present new observations and analyses on lip-reading tasks for unseen speakers, focusing on two aspects: the distinction between speaker characteristics and speech content features at shallow and deep levels, and the dynamic role of speaker characteristics in recognizing different words. 2) We propose a new speaker adaptive lip reading method by taking advantage of the speaker's characteristics to learn for adaptively enhancing and suppressing separable hidden unit contributions for robust lip reading. 3) For evaluation of unseen speaker's tasks in extreme settings, we release a new lip reading benchmark, CAS-VSR-S68, which involves only a few speakers but a diversified speech content range. Experiments on both the public dataset and our new benchmark show the advantages of our method.

2 Related Work

Lip Reading. With the flourishing development of deep learning [17, 28], remarkable progress has been made in lip reading technology. At present, most lip reading models rely on the end-to-end deep learning framework, which can be broadly divided into visual feature extraction front-ends and language content decoding back-ends[35]. Early researchers endeavored to enhance the visual feature and language feature modeling capabilities by modifying the neural network structure[1, 4, 40, 41] or extracting knowledge from other modalities, such as audio[3, 29] and deformation flow[45]. Other researchers have created extensive audio-visual datasets from online videos to train models applicable to real-world scenarios[1, 2, 8, 9, 10, 49]. Recently, researchers have become increasingly interested in utilizing massive amounts of unlabeled data for self-supervised pre-training [5, 6, 16, 30, 38].

However, similar to automatic speech recognition (ASR), lip reading models always encounter performance degradation when dealing with unseen speakers [4, 52]. The emerging topic of speaker adaptation in lip reading has gradually attracted attention in recent years, and some notable works have been conducted in this area [21, 22, 24, 25, 48, 50]. In comparison to their methods, we designed a new speaker adaptive method from a novel perspective, based on the analysis of the speaker's characteristics and their effect on lip reading. Our method achieved superior performance, both with and without extra adaptation data.

Speaker Adaptation. A prevalent issue in audio-based Automatic Speech Recognition (ASR) is the significant degradation of recognition performance when testing conditions differ from the training conditions [7]. Speaker adaptation methods aim to address this problem. Some of these methods aim to extract more representative feature embeddings that capture speaker characteristics, including i-vectors [36], d-vectors [44], x-vectors [39], r-vectors [23], and l-vectors [34]. Regularization-like approaches, such as meta-learning [26, 27] or limiting the distance between the adaptation model and the speaker-independent model [31, 33], are also commonly used for speaker adaptation to mitigate overfitting. In some cases, data augmentation methods are used to alleviate data imbalance issues in datasets[18].

LHUC (Learning Hidden Unit Contribution) is a speaker adaptation method in which a separate linear layer is added to a pre-trained neural network to adjust the contribution of the hidden units for a specific speaker[43]. This method has been successful in adapting models to different speakers in ASR tasks and has been adopted in many works[13, 14, 19, 20, 42, 46, 47]. Inspired by the LHUC approach, we have taken into account the differences between ASR and lip reading tasks, and developed a novel speaker adaptive method for lip

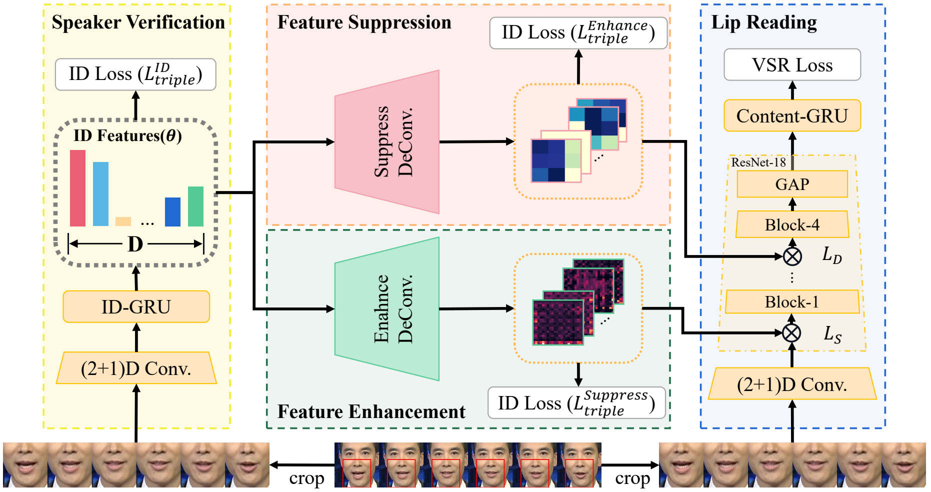


Figure 2: The Overall Architecture of Our Proposed Method.

reading based on our new observation and analysis of the speaker's characteristics and their effect on the lip reading task.

3 Our Proposed Method

Based on the previous observation and analysis, we propose a speaker-adaptive method that utilizes speaker-dependent features to enhance and suppress features at shallow and deep layers of the lip reading network, respectively. The overall architecture includes four modules, as shown in Figure 2: the speaker verification module, the feature enhancement module, the feature suppression module, and the lip reading module. The speaker verification module, represented by the yellow box on the left, aims to learn the speaker's characteristics. These characteristics are then passed to the feature enhancement module (lower middle green box) and the feature suppression module (upper middle pink box) to respectively generate speaker-dependent enhancement and suppression weights. These weights are then applied to the shallow and deep layers of the lip reading module (blue box on the left) respectively to adaptively enhance content-dependent features and suppress content-independent features.

3.1 Model Architecture

We will introduce the four modules of our method in this part respectively.

Speaker Verification Module (Yellow Box). To learn speaker-dependent features for lip reading, we use speaker verification as a monitoring and guiding task during our model's learning process. Traditional face recognition models are not capable of modeling dynamic speaker-dependent features over time, such as speaking style, which can provide helpful auxiliary information for lip reading. Therefore, we propose a speaker verification module that crops each frame in the talking face video to the lip region and uses a (2+1)D convolutional network to model the spatial and temporal information. Then the speaker-dependent feature θ is generated with a single layer of GRU which is capable of modeling the temporal characteristics of the speaker. The obtained speaker-dependent feature finally serves as input to the subsequent feature enhancement and suppression modules.

Lip Reading Module (Blue Box). We adopt a popular lip reading network[12] as the backbone, which consists of a frontend network and a backend network. The frontend network includes a multi-layer (2+1)D convolutional neural network and a 2D ResNet-18, which models local temporal and spatial information in the sequence and aims to extract fine-grained visual features at each time step. The feature enhancement and suppression modules separately generate speaker adaptive weights, which are element-wise multiplied with the shallow and deep layer features of ResNet-18, respectively. The backend network consists of a 3-layer bidirectional GRU that models global temporal and linguistic information and generates content-dependent features for lip reading.

Feature Enhancement/Suppression Module (Green/Pink Box). The feature enhancement module is used for shallow layers where the speech content features are less prominent, in order to adaptively take advantage of the speaker's characteristics to enhance the speech content features. It generates enhancement weights based on the speaker's features θ to guide the model to focus on content-dependent features. Then these weights are applied to the shallow layers L_S (the first layer of the 2D ResNet-18) in the lip reading module. Specifically, we introduce an enhancement deconvolutional network to upsample the speaker's features θ to the final enhancements weights of the same size and channel number as the target shallow layer L_S of the lip reading module. The enhancement weights are generated with the activation function $\sigma_{enhance}(h) = 1 + |\text{LeakyReLU}(h)|$ after enhancement deconvolutional layers. As a result, the output values are bounded to the range of [1,+∞), ensuring that content features that are originally weak in the shallow layers of the lip reading module are only enhanced adaptively, rather than being suppressed. This approach is beneficial for achieving robust lip reading features to feed subsequent stages.

For deep layers where the speech content features have been obtained well, the feature suppression module is introduced to take the speaker's characteristics to generate suppression weights to adaptively suppress irrelevant noise for robust lip reading. This module also takes the speaker's feature θ as input but generates the suppression weights through another deconvolutional network as shown in Figure 2. These weights are then applied to the deep layers L_D(14th layer of the 2D ResNet-18) of the lip reading module. We apply an activation function of $\sigma_{suppress}(h) = 1 - |\tanh(h)|$ after the suppression deconvolutional layers to obtain values within the range of (0,1]. The operation of adaptive suppression aims to eliminate only the irrelevant noise, rather than content-dependent features, to minimize the risk of model overfitting.

3.2 Optimization

Four losses of the architecture are shown in the white solid box in Figure 2.

ID Loss for Speaker Verification. To ensure the effective extraction of speaker characteristics by the speaker verification module, we introduce the triplet loss [37]. This helps the model map ID features to a generalizable feature space while avoiding the use of an excessively large linear layer. Specifically, during the sample selection process, we choose a positive sample A'_{ID} of the same speaker as the anchor video A_{ID}, and a negative sample B_{ID} of a different speaker. Then, we simultaneously input the three samples into the speaker verification module to extract the speaker's features θ, denoted as $\theta_{A_{ID}}$, $\theta_{A'_{ID}}$, and $\theta_{B_{ID}}$, respectively. The optimization of this loss function is as:

$$L^{ID}_{triple}(\theta_{A_{ID}}, \theta_{A'_{ID}}, \theta_{B_{ID}}) = \max(d(\theta_{A_{ID}} - \theta_{A'_{ID}}) - d(\theta_{A_{ID}} - \theta_{B_{ID}}) + \alpha_t, 0) \tag{1}$$

where α_t is a margin between positive and negative samples of ID features, and the Euclidean distance is used as the distance function, i.e., $d(A,B) = ||A - B||^2$.

ID Loss for Feature Enhancement/Suppression Module. To avoid the model collapsing to trivial solutions, where all generated enhancement and suppression weights are equal to 1, we add speaker-level supervision to the generated weights to ensure that they exhibit differences across speakers. Similar to the loss function of the speaker verification module, we first input the speaker's feature θ extracted by the speaker verification module into the feature enhancement network to obtain the corresponding enhancement weights $\theta^{Enhance}$. The optimized loss can be represented as

$$L_{triple}^{Enhance}(A_{ID}, A'_{ID}, B_{ID}) = \max(d(\theta_{A_{ID}}^{Enhance} - \theta_{A'_{ID}}^{Enhance}) - d(\theta_{A_{ID}}^{Enhance} - \theta_{B_{ID}}^{Enhance}) + \alpha_E, 0) \tag{2}$$

where α_E represents the margin that distinguishes the weights generated for positive and negative samples in the enhancement process.

For the suppression weights, we use a similar approach to calculate the optimization loss:

$$L_{triple}^{Suppress}(A_{ID}, A'_{ID}, B_{ID}) = \max(d(\theta_{A_{ID}}^{Suppress} - \theta_{A'_{ID}}^{Suppress}) - d(\theta_{A_{ID}}^{Suppress} - \theta_{B_{ID}}^{Suppress}) + \alpha_S, 0) \tag{3}$$

where α_S represents the margin that distinguishes the weights generated for positive and negative samples in the suppression process. The choice of α_S was determined based on the proportion of the mean and variance of the two types of weights since the generated enhancement and suppression weights have corresponding value ranges. We used the Euclidean distance as the distance measure d.

VSR Loss. We have carried out evaluations at both the word-level and sentence-level for lip-reading tasks. For word-level lip reading task, we use the cross-entropy loss for training, represented as

$$L_{CE}^{VSR} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{|V|} y_{i,j}\log(\hat{y}_{i,j}) \tag{4}$$

where N is the total number of training samples, $|V|$ is the vocabulary size, $y_{i,j}$ is the ground truth label, and $\hat{y}_{i,j}$ is the predicted probability of the j-th word for the i-th sample. For sentence-level lip reading, we used the CTC loss [15] for optimization, represented as

$$L_{CTC}^{VSR} = -\log \sum_{\pi \in \mathcal{B}^{-1}(y)} p(\pi|x) \tag{5}$$

where $\mathcal{B}^{-1}(y)$ denotes the set of all possible alignments of the label sequence y and $p(\pi|x)$ is the probability of the alignment π given the predicted input sequence x.

4 Experiments

4.1 Datasets

GRID[11] corpus comprises 1000 utterances spoken by 33 speakers, with both audio and visual data captured simultaneously. We use the widely-adopted split of unseen speakers as in LipNet[4], where the 1st, 2nd, 20th, and 22nd speakers serve as the test set while the remaining speakers are used for training. When evaluating in the setting of providing adaptation data, we randomly divide the four speakers of the original test set of GRID into adaptation and evaluation sets for each speaker[24].

LRW-ID[24] is a redivision of the LRW dataset[10]. LRW is a large-scale dataset for lip reading, which consists of 500 different words spoken by more than 10 thousand speakers in the wild. Kim et al. [24] use face recognition technology to annotate the identity of speakers and redivide the original training and testing sets to evaluate speaker-adaptive lip reading methods. Specifically, LRW-ID selects data of 20 speakers as validation (adaptation) and test sets, and the remaining data is used for training. We adopt the same division for evaluation.

CAS-VSR-S68 is a new dataset for evaluation in the extreme setting of unseen speakers' lip reading. The data was collected from news broadcast programs aired between 2009 and 2019, covering a wide range of topics and speakers. The video clips and corresponding

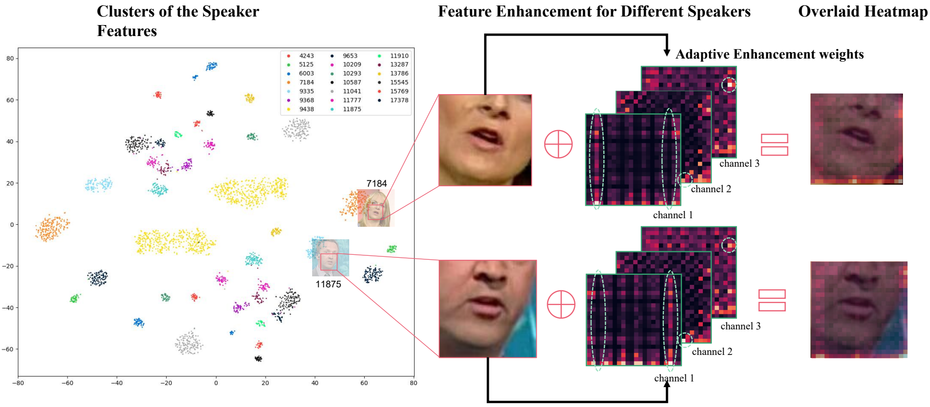


Figure 3: Visualization of the Generated Enhancement Weights

text annotations were extracted from the broadcast clips of the hosts, with a resolution of 256*256 pixels and sentence lengths varying from one character to over 20 characters. The dataset has a total duration of approximately 68 hours and includes over 3800 commonly used Chinese characters, making it a challenging dataset for lip reading tasks. The dataset contains data from 11 hosts, with 10 hosts' data used for training and the remaining host's data randomly split into adaptation and testing sets.

The detailed settings for each dataset are given in the supplementary materials.

4.2 Results

4.2.1 Qualitative Analysis

Discriminative Speaker's Features. Figure 3 shows the visualization of speaker features θ for each speaker in the LRW-ID dataset using t-SNE dimensionality reduction.We performed clustering on the reduced speakers' feature θ, and each cluster is generally gathered together, effectively capturing the unique characteristics of each speaker.

Adaptive Enhancement Weights. We produce heat map visualizations to represent the enhancement weights for randomly selected samples, as shown in the middle of Figure 3. The enhancement weights for the same channel display significant differences across different speakers. For instance, the green circle in the figure highlights the significant contrast between the two randomly sampled speakers. The same region may be significantly enhanced for one speaker, while the enhancement in the same region for another speaker may not be obvious.

Visual Analysis of Enhanced Features. We compute the average of the feature enhancement weights along the channel dimension and overlay them on the corresponding speaker's input image with some transparency. This enables us to analyze which part of the information on the face would the feature enhancement weights intend to focus on. As displayed on the right-hand side of Figure 3, the feature enhancement weights are noticeably concentrated in the regions close to the image edges. This is because lip reading networks usually focus on the central lip region of the image, neglecting helpful information from other regions such as the chin, cheeks, and nose. In fact, it has been shown that these areas contain speech-related information indeed[51]. Our proposed method is able to enhance the model's attention to areas beyond the lips, therefore enabling more effective recognition of the speaker's speech content.

Table 1: Ablation Study of Loss Functions

Method	$\mathbf{L^{ID}_{triple}}$	$\mathbf{L^{Enh}_{triple}}$&$\mathbf{L^{Sup}_{triple}}$	$\mathbf{L^{VSR}_{CE}}$	**Acc(%)**
Baseline	-	-	✓	87.25
Ours	×	×	✓	87.73
	✓	×	✓	87.74
	×	✓	✓	87.75
	✓	✓	✓	**87.91**

Table 2: Ablation Study of Modules

Method	**Acc(%)**
Baseline	87.25
Enhance Only	87.83
Suppress Only	87.81
Proposed	**87.91**

Table 3: Performance on LRW-ID with Reduced Data and Speaker Diversity

	Sample size	**Number of speakers**	**Acc(%) Baseline**	**Acc(%) Ours**	**Perf.Drop Baseline**	**Perf.Drop Ours**
a	480378	17560	87.25	87.91	-	-
b	383788	17551	85.26	86.11	↓ 2.28%	↓ **2.05%**
c	388577	1047	85.4	86.36	↓ 2.12%	↓ **1.76%**
d	386681	1000	85.2	86.27	↓ 2.35%	↓ **1.87%**
e	354807	500	84.22	84.88	↓ 3.47%	↓ **3.45%**
f	298369	200	81.37	82.7	↓ 6.74%	↓ **5.93%**
g	246108	100	77.46	78.5	↓ 11.22%	↓ **10.70%**

The suppression weight also exhibits a similar property to the enhancement weight, and detailed analyses are provided in the supplementary materials.

4.2.2 Quantitative Analysis

Ablation Study. The ablation study in Table 1 shows the effectiveness of each loss in our method. L^{ID}_{triple} constrains the speaker verification module to extract each speaker's own features correctly. $L^{Enhance}_{triple}$ and $L^{Suppress}_{triple}$ are introduced to prevent the collapse of the generated weights. Therefore, removing either loss results in consistent performance degradation.

The ablation experiments in Table 2 show the effectiveness of each module. The enhance and suppress modules improve the model's ability to use speaker-dependent information to handle unseen speakers. When combined, they further improve the model's ability to handle unseen speakers, demonstrating the effectiveness of our method.

Reduce Speaker Diversity. Usually, the lip reading model's performance on unseen speakers would decline as the sample size decreases or speaker diversity reduces for training. As shown in Table 3, our proposed method effectively alleviates this degradation in performance caused by both the reduction in speaker diversity and sample size in the training set.

Table 4: Comparison with Other Methods on LRW-ID and CAS-VSR-S68

Adapt min.	**LRW-ID(ACC%)**					**CAS-VSR-S68(CER%)**	
	User-padding[24]	Prompt Tuning[25]	DCTCN [32]	Baseline	Proposed Method	Baseline	Proposed Method
0	85.85	87.54	86.75	87.25	**87.91**	19.61	19.37
1	87.06	88.53	-	88.52	**89.21**	21.53	20.69
3	87.61	89.45	-	89.48	**89.88**	18.65	18.55
5	87.91	89.99	-	89.96	**90.45**	17.55	**16.72**

Notably, our method demonstrates significant mitigation of performance loss in scenarios where the sample size is relatively close but the number of speakers is small, as observed in experiments **b** and **c**, indicating that our method utilizes speaker information sufficiently. Furthermore, experiments from **d** to **g** demonstrate that our model generally mitigates performance degradation as the speaker diversity decreases, even in scenarios with only 100 speakers remaining.

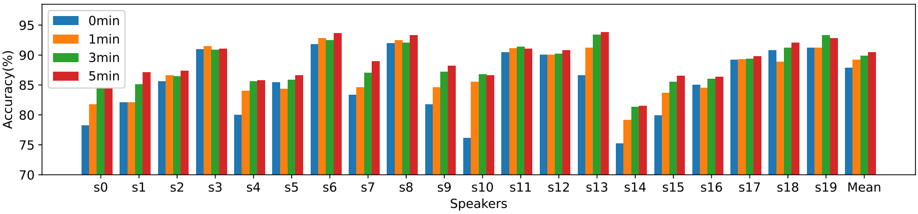


Figure 4: Adaptation Rresult Using Different Amount of Adaptation Data on LRW-ID

Table 5: Comparison Results on GRID with Other Methods without any Adaptation Data

Method	WER(%)
WAS[10]**	14.6
LipNet[4]**	11.4
TM-seq2seq[1]**	11.7
User-padding[24]	11.12
User-padding[24]*	7.2
Prompt Tuning[25]	12.04
TVSR-Net[48]	9.1
DVSR-Net[50]	7.8
Visual i-vector[21]	7.3
Baseline (ours)	10.62
Proposed Method	9.59
Proposed Method*	**6.99**

* Using our method in the manner as [24]

** Results reproduced by [48, 50]

Table 6: Adaptation Result on GRID Dataset

Method	Adapt min.	WER(%)
User Padding[24]	0	11.12
	1	6.8
	3	6.05
	5	5.67
Prompt Tuning[25]	0	12.04
	1	5.53
	3	4.31
	5	3.8
Proposed Method	0	9.59
	1	5.61
	3	4.6
	5	**3.59**

Without Adaptation Data. Our proposed lip reading method performs well in scenarios where there is no adaptation data for unseen speakers. From Table 4, it can be seen that our proposed method not only achieves an overall improvement compared to the LRW-ID baseline but also outperforms the previous works. The comparison between the baseline and our proposed method in Table 5 demonstrates that even when the baseline performance is already relatively high, our method still achieves a performance improvement of about relativelt 9.70% on GRID. By applying the unsupervised speaker adaptation technique proposed in [24], our model achieves a significant improvement in WER, reducing it from 9.59% to 6.99%, which corresponds to an additional performance gain of approximately 27.1%. We also observed performance gains on the CAS-VSR-S68 dataset, which represents an extreme case characterized by limited speaker diversity and a broad scope of speech content. This is a challenging dataset, with only 1-minute short adaptation data, the performance decreased instead of increasing, which fully illustrates the extreme situation and challenges faced by this dataset. In summary, our method shows its ability on effective utilization of speaker-dependent information to improve the model's ability to recognize unseen speakers, regardless of the speaker diversity of the dataset, as observed in LRW-ID (20k+ speakers), GRID (29 speakers), and CAS-VSR-S68 (9 speakers).

With Limited Adaptation Data. In addition to the model's excellent performance in the absence of any adaptation data, our method can also effectively utilize limited speaker adaptation data (<5min) to significantly improve the model's performance, outperforming other methods. As shown in Figure 4, the model's performance generally increases as the amount of adaptation data increases. Combining with Tables 4 and 6, our method shows a significant advantage on the LRW-ID and GRID datasets, achieving an accuracy of 90.45% and a WER of 3.59% respectively with only about 5 minutes of training data.

5 Conclusion

This paper proposes a novel speaker adaptive method for lip reading of unseen speakers, leveraging the speaker's own characteristics to learn separable hidden unit contributions. The approach outperforms existing methods on public popular datasets and our challenging dataset CAS-VSR-S68 with a few speakers but diverse speech content. Our method provides a new idea and solution for robust visual speech recognition.

6 Acknowledgements

This work is partially supported by National Natural Science Foundation of China (No. 62276247, 62076250). We thank Yuanhang Zhang for his contributions to the collection, download, and processing of the CAS-VSR-S68 dataset.

References

[1] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[2] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Lrs3-ted: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*, 2018.

[3] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Asr is all you need: Cross-modal distillation for lip reading. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2143–2147. IEEE, 2020.

[4] Yannis M Assael, Brendan Shillingford, Shimon Whiteson, and Nando De Freitas. Lipnet: End-to-end sentence-level lipreading. *arXiv preprint arXiv:1611.01599*, 2016.

[5] Alexei Baevski, Arun Babu, Wei-Ning Hsu, and Michael Auli. Efficient self-supervised learning with contextualized target representations for vision, speech and language. *arXiv preprint arXiv:2212.07525*, 2022.

[6] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*, pages 1298–1312. PMLR, 2022.

[7] Peter Bell, Joachim Fainberg, Ondrej Klejch, Jinyu Li, Steve Renals, and Pawel Swietojanski. Adaptation algorithms for neural network-based speech recognition: An overview. *IEEE Open Journal of Signal Processing*, 2:33–66, 2020.

[8] Oscar Chang, Hank Liao, Dmitriy Serdyuk, Ankit Shah, and Olivier Siohan. Conformers are all you need for visual speech recogntion. *arXiv preprint arXiv:2302.10915*, 2023.

[9] J. S. Chung and A. Zisserman. Lip reading in the wild. In *Asian Conference on Computer Vision*, 2016.

[10] 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*, 2017.

[11] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120(5):2421–2424, 2006.

[12] Dalu Feng, Shuang Yang, Shiguang Shan, and Xilin Chen. Learn an effective lip reading model without pains. *arXiv preprint arXiv:2011.07557*, 2020.

[13] Mengzhe Geng, Xurong Xie, Rongfeng Su, Jianwei Yu, Zi Ye, Xunying Liu, and Helen Meng. On-the-fly feature based speaker adaptation for dysarthric and elderly speech recognition. *arXiv preprint arXiv:2203.14593*, 2022.

[14] Mengzhe Geng, Xurong Xie, Zi Ye, Tianzi Wang, Guinan Li, Shujie Hu, Xunying Liu, and Helen Meng. Speaker adaptation using spectro-temporal deep features for dysarthric and elderly speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:2597–2611, 2022.

[15] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376, 2006.

[16] Alexandros Haliassos, Pingchuan Ma, Rodrigo Mira, Stavros Petridis, and Maja Pantic. Jointly learning visual and auditory speech representations from raw data. *arXiv preprint arXiv:2212.06246*, 2022.

[17] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97, 2012.

[18] Ehsan Hosseini-Asl, Yingbo Zhou, Caiming Xiong, and Richard Socher. Augmented cyclic adversarial learning for low resource domain adaptation. *arXiv preprint arXiv:1807.00374*, 2018.

[19] Zengrui Jin, Mengzhe Geng, Jiajun Deng, Tianzi Wang, Shujie Hu, Guinan Li, and Xunying Liu. Personalized Adversarial Data Augmentation for Dysarthric and Elderly Speech Recognition, June 2022. URL http://arxiv.org/abs/2205.06445. arXiv:2205.06445 [cs, eess].

[20] Zengrui Jin, Xurong Xie, Mengzhe Geng, Tianzi Wang, Shujie Hu, Jiajun Deng, Guinan Li, and Xunying Liu. Adversarial data augmentation using vae-gan for disordered speech recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[21] Pujitha Appan Kandala, Abhinav Thanda, Dilip Kumar Margam, Rohith Chandrashekar Aralikatti, Tanay Sharma, Sharad Roy, and Shankar M Venkatesan. Speaker adaptation for lip-reading using visual identity vectors. In *INTERSPEECH*, pages 2758–2762, 2019.

[22] Pujitha Appan Kandala, Abhinav Thanda, Dilip Kumar Margam, Rohith Chandrashekar Aralikatti, Tanay Sharma, Sharad Roy, and Shankar M. Venkatesan. Speaker Adaptation for Lip-Reading Using Visual Identity Vectors. In *Interspeech 2019*, pages 2758–2762. ISCA, September 2019. doi: 10.21437/Interspeech.2019-3237. URL https://www.isca-speech.org/archive/interspeech_2019/kandala19_interspeech.html.

[23] Yuri Y Khokhlov, Alexander Zatvornitskiy, Ivan Medennikov, Ivan Sorokin, Tatiana Prisyach, Aleksei Romanenko, Anton Mitrofanov, Vladimir Bataev, Andrei Andrusenko, Mariya Korenevskaya, et al. R-vectors: New technique for adaptation to room acoustics. In *INTERSPEECH*, pages 1243–1247, 2019.

[24] Minsu Kim, Hyunjun Kim, and Yong Man Ro. Speaker-adaptive lip reading with user-dependent padding. In *European Conference on Computer Vision*, pages 576–593. Springer, 2022.

[25] Minsu Kim, Hyung-Il Kim, and Yong Man Ro. Prompt tuning of deep neural networks for speaker-adaptive visual speech recognition. *arXiv preprint arXiv:2302.08102*, 2023.

[26] Ondej Klejch, Joachim Fainberg, and Peter Bell. Learning to adapt: a meta-learning approach for speaker adaptation, August 2018. URL http://arxiv.org/abs/1808.10239. arXiv:1808.10239 [cs].

[27] Ondej Klejch, Joachim Fainberg, Peter Bell, and Steve Renals. Speaker Adaptive Training using Model Agnostic Meta-Learning, October 2019. URL http://arxiv.org/abs/1910.10605. arXiv:1910.10605 [cs, eess].

[28] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[29] Wei Li, Sicheng Wang, Ming Lei, Sabato Marco Siniscalchi, and Chin-Hui Lee. Improving audio-visual speech recognition performance with cross-modal student-teacher training. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6560–6564. IEEE, 2019.

[30] Jiachen Lian, Alexei Baevski, Wei-Ning Hsu, and Michael Auli. Av-data2vec: Self-supervised learning of audio-visual speech representations with contextualized target representations. *arXiv preprint arXiv:2302.06419*, 2023.

[31] Hank Liao. Speaker adaptation of context dependent deep neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7947–7951. IEEE, 2013.

[32] Pingchuan Ma, Yujiang Wang, Stavros Petridis, Jie Shen, and Maja Pantic. Training strategies for improved lip-reading. In *ICASSP*, pages 8472–8476. IEEE, 2022.

[33] Zhong Meng, Jinyu Li, and Yifan Gong. Adversarial speaker adaptation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5721–5725. IEEE, 2019.

[34] Zhong Meng, Hu Hu, Jinyu Li, Changliang Liu, Yan Huang, Yifan Gong, and Chin-Hui Lee. L-vector: Neural label embedding for domain adaptation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7389–7393. IEEE, 2020.

[35] Marzieh Oghbaie, Arian Sabaghi, Kooshan Hashemifard, and Mohammad Akbari. Advances and challenges in deep lip reading. *arXiv preprint arXiv:2110.07879*, 2021.

[36] George Saon, Hagen Soltau, David Nahamoo, and Michael Picheny. Speaker adaptation of neural network acoustic models using i-vectors. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 55–59. IEEE, 2013.

[37] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[38] Bowen Shi, Wei-Ning Hsu, and Abdelrahman Mohamed. Robust self-supervised audio-visual speech recognition. *arXiv preprint arXiv:2201.01763*, 2022.

[39] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5329–5333. IEEE, 2018.

[40] Themos Stafylakis and Georgios Tzimiropoulos. Combining residual networks with lstms for lipreading. *arXiv preprint arXiv:1703.04105*, 2017.

[41] Themos Stafylakis, Muhammad Haris Khan, and Georgios Tzimiropoulos. Pushing the boundaries of audiovisual word recognition using residual networks and lstms. *Computer Vision and Image Understanding*, 176:22–32, 2018.

[42] Pawel Swietojanski and Steve Renais. Sat-lhuc: Speaker adaptive training for learning hidden unit contributions. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5010–5014. IEEE, 2016.

[43] Pawel Swietojanski and Steve Renals. Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 171–176. IEEE, 2014.

[44] Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez Moreno, and Javier Gonzalez-Dominguez. Deep neural networks for small footprint text-dependent speaker verification. In *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4052–4056. IEEE, 2014.

[45] Jingyun Xiao, Shuang Yang, Yuanhang Zhang, Shiguang Shan, and Xilin Chen. Deformation flow based two-stream network for lip reading. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 364–370. IEEE, 2020.

[46] Xurong Xie, Xunying Liu, Tan Lee, Shoukang Hu, and Lan Wang. Blhuc: Bayesian learning of hidden unit contributions for deep neural network speaker adaptation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5711–5715. IEEE, 2019.

[47] Xurong Xie, Rukiye Ruzi, Xunying Liu, and Lan Wang. Variational auto-encoder based variability encoding for dysarthric speech recognition. *arXiv preprint arXiv:2201.09422*, 2022.

[48] Chenzhao Yang, Shilin Wang, Xingxuan Zhang, and Yun Zhu. Speaker-Independent Lipreading With Limited Data. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 2181–2185, 2020. doi: 10.1109/ICIP40778.2020.9190780. ISSN: 2381-8549.

[49] Shuang Yang, Yuanhang Zhang, Dalu Feng, Mingmin Yang, Chenhao Wang, Jingyun Xiao, Keyu Long, Shiguang Shan, and Xilin Chen. Lrw-1000: A naturally-distributed large-scale benchmark for lip reading in the wild. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–8. IEEE, 2019.

[50] Qun Zhang, Shilin Wang, and Gongliang Chen. Speaker-independent lipreading by disentangled representation learning. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 2493–2497. IEEE, 2021.

[51] Yuanhang Zhang, Shuang Yang, Jingyun Xiao, Shiguang Shan, and Xilin Chen. Can we read speech beyond the lips? rethinking roi selection for deep visual speech recognition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 356–363. IEEE, 2020.

[52] Yingzhu Zhao, Chongjia Ni, Cheung-Chi Leung, Shafiq Joty, Eng Siong Chng, and Bin Ma. A unified speaker adaptation approach for asr. *arXiv preprint arXiv:2110.08545*, 2021.