# Supplementary Materials: Learning Separable Hidden Unit Contributions for Speaker-Adaptive Lip Reading

BMVC 2023 Submission # 146

In this supplementary material, we provide additional insights and analyses of our method for lip reading. Specifically, Section 1 illustrates the distinction between speaker-dependent and content-dependent features extracted by lip reading models. Section 2 presents more experimental results, including more detailed quantitative results and qualitative visualizations, which further prove the effectiveness of our proposed approach.

## 1 Illustration of the Features Extracted by Lip Reading Models

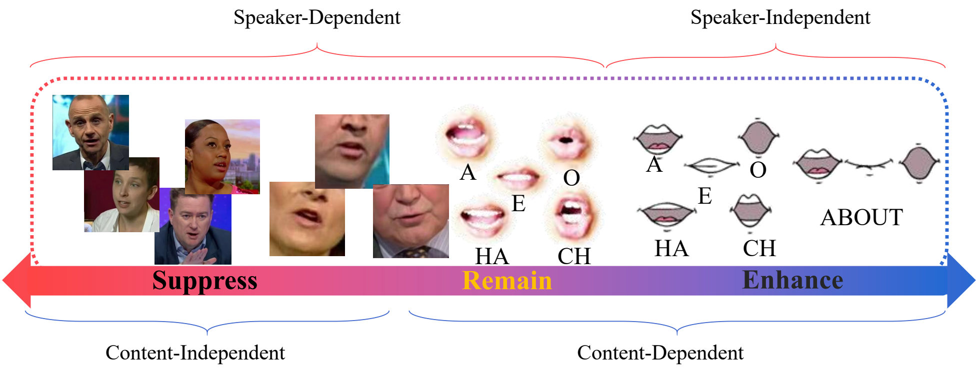


Figure 1: Illustration of the relationships between the speaker-dependent and the content-dependent features.

Given any lip reading model, we can roughly divide the features extracted by this model into two types according to different criteria: speaker-dependent and speaker-independent features, or content-dependent and content-independent features. The features under these two criteria focus on expressing different properties when give a talking face video. We show these two types as the two ends with different colors in Figure 1.

**Speaker-dependent features** primarily capture the unique characteristics of the speaker and are always reflected by the speaker's static facial traits, such as mouth shape, skin texture, skin tone, beard, markings, and also a few dynamic traits corresponding with the speaker's speaking style. These features encode the speaker's individuality and remain relatively constant when speaking different words or utterances. They significantly contribute to the overall process of identifying and differentiating the speaker from other individuals.

**Content-dependent features** are closely related to the specific spoken content and mainly focus on the fine-grained spatio-temporal changes in the facial region, especially the lip region, during the speaking process. They are more sensitive to the specific words being pronounced, which provides the basis for the lip reading task.

As discussed in the main submission, there exists an interesting phenomenon regarding the performance of shallow-layer and deep-layer features for speaker identification and lip reading tasks. The accuracy of speaker identification using shallow-layer features is already high, and as the layers go deeper in the network, the accuracy of speaker identification experiences a rapid increase. However, when utilizing the same shallow-layer features for lip reading, the recognition accuracy is relatively low, and the rate of improvement in accuracy is much slower compared to speaker identification. Intriguingly, the rate of increase in lip reading task accuracy is generally higher than that of speaker classification accuracy in the deep layers of the network. This phenomenon highlights the distinctive nature of our method to learn separable hidden unit contributions for shallow and deep layers respectively.

# 2 More Detailed Experiments

## 2.1 Training Details

We employ a three-step training approach to learn our model as shown in Figure 2 in the main submission, to learn the contradictory targets of the enhancement and suppression module. Firstly, we train the left speaker verification module with $L^{ID}_{triple}$ and the right lip reading modules with $L^{VSR}_{CE}$ separately. Then, we introduce the feature enhancement module together with the learned speaker verification module and the lip reading module to continue the training process. Finally, we freeze the feature enhancement module and the speaker verification module to introduce the suppression module to continue training until convergence.

## 2.2 Experimental Setup

**LRW-ID:** We utilized the Adam[5] optimizer with a maximum learning rate of $8.125 \times 10^{-4}$ and a batch size of 130. The input size was $29 \times 96 \times 96$ (T, W, H), where T represents the number of frames. Data augmentation techniques included horizontal flipping and random cropping to $88 \times 88$. The fine-tuning phase with adaptation data involved the Adam optimizer with a maximum learning rate of $6.25 \times 10^{-5}$ and a batch size of 200.

**GRID:** The Adam optimizer with a maximum learning rate of $1.5 \times 10^{-4}$ and a batch size of 32 was employed. The input size was set to $29 \times 96 \times 96$ (T, W, H), and data augmentation techniques included horizontal flipping and random cropping to $88 \times 88$. Fine-tuning with adaptation data was performed using the Adam optimizer with a maximum learning rate of $7.25 \times 10^{-5}$ and a batch size of 32.

**CAS-VSR-S60h:** The Adam optimizer with a maximum learning rate of $8.125 \times 10^{-4}$ and a dynamic batch strategy was used. The maximum input frame count was set to 300, and the input size was $T \times 96 \times 96$ (T, W, H). Data augmentation included horizontal flipping with a 0.5 probability. For fine-tuning with adaptation data, the Adam optimizer with a maximum learning rate of $6.25 \times 10^{-5}$ and a batch size of 1 was used. Similar to the previous datasets, 1 minute, 3 minutes, and 5 minutes of data from the adaptation set were randomly selected for full model fine-tuning.

## 2.3 Results

### 2.3.1 More Quantitative Results

Table 1: Comparison on GRID with Other Methods without Any Adaptation Data

| **Method** | **Test Speaker** | | | | **Mean(WER)** |
|---|---|---|---|---|---|
| | S1 | S2 | S20 | S22 | |
| LipNet (reproduce)[1] | 22.13 | 10.42 | 11.83 | 6.73 | 13.6 |
| User-padding[3] | 17.04 | 9.02 | 10.33 | 8.13 | 11.12 |
| User-padding[3]* | - | - | - | - | 7.2 |
| Prompt Tuning[4] | 16.4 | 9.42 | 11.23 | 11.57 | 12.04 |
| DVSR-Net[6] | - | - | - | - | 9.1 |
| TVSR-Net[7] | - | - | - | - | 7.8 |
| Visual i-vector[2] | - | - | - | - | 7.3 |
| Baseline (ours) | 19.60 | 10.96 | 7.26 | 4.65 | 10.62 |
| Proposed Method | 17.96 | 9.20 | 6.46 | 4.72 | 9.59 |
| Proposed Method* | 13.01 | 5.63 | 5.86 | 3.45 | **6.99** |

* Test in the manner as [3]

Table 2: Using Different Quantities of Adaptation Data on GRID

| **Method** | **Adapt min.** | **Test Speaker** | | | | **Mean(WER)** |
|---|---|---|---|---|---|---|
| | | S1 | S2 | S20 | S22 | |
| User Padding[3] | 0 | 17.04 | 9.02 | 10.33 | 8.13 | 11.12 |
| | 1 | 10.65 | 4.2 | 7.77 | 4.59 | 6.8 |
| | 3 | 9.35 | 3.75 | 6.88 | 4.27 | 6.05 |
| | 5 | 8.78 | 3.45 | 6.49 | 3.99 | 5.67 |
| Prompt Tuning[4] | 0 | 16.40 | 9.42 | 11.23 | 11.57 | 12.04 |
| | 1 | 7.91 | 3.81 | 6.07 | 4.43 | 5.53 |
| | 3 | 6.43 | **2.14** | 5.63 | 3.07 | 4.31 |
| | 5 | 5.08 | 2.24 | 5.13 | 2.8 | 3.8 |
| Proposed Method | 0 | 17.96 | 9.20 | 6.46 | 4.72 | 9.59 |
| | 1 | 9.22 | 4.53 | 5.59 | 3.11 | 5.61 |
| | 3 | 6.52 | 3.2 | 6.12 | **2.54** | 4.60 |
| | 5 | **4.78** | 2.53 | **4.38** | 2.68 | **3.59** |

**Detailed Results on GRID.** In our main submission, we evaluated the effectiveness of our method on the GRID dataset by measuring the Word Error Rate (WER), both with and without adaptation data. In this section, we provide a detailed analysis of the experimental results for each speaker and compare our method with other approaches.

Table 1 clearly shows that our method consistently outperforms the comparison methods in terms of WER across all speakers, even in the absence of adaptation data. Additionally, our method exhibits a notable overall performance improvement compared to the baseline. We specifically observe significant improvements for speakers S1 and S2, who initially had higher WER. Moreover, Table 2 demonstrates consistent improvements achieved across different speakers when adaptation data is available. Remarkably, with a mere 5 minutes of adaptation data, the WER for Speaker 1 significantly decreases from 17.96 to 4.78, repre-

Table 3: Character Error Rate (CER) on CAS-VSR-S68h with Adaptation Data

| Adapt min | Baseline | Proposed Method |
|---|---|---|
| 0 | 44.93 | 43.24 |
| 1 | 38.63 | 37.38 |
| 3 | 36.37 | 35.64 |
| 5 | 33.79 | **33.17** |

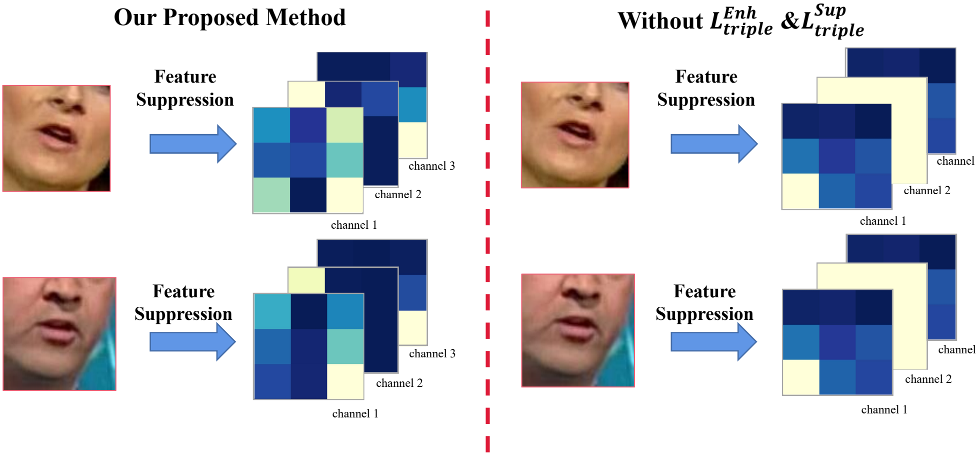


Figure 2: Visualization of Features Generated by Suppression Module

senting an impressive reduction of approximately 73.4%.

**Detailed Results on CAS-VSR-S68h.** To further validate the effectiveness of our method, we conducted additional experiments using a different test set. In the main submission, we utilized the data of a male news anchor, Gang Qiang, as the test set. In addition to that, we also evaluated our method on a separate test set consisting of data from a female news anchor, Li Ruiying, as shown in Table 3.

It is worth noting that due to the limited amount of female data in the training set, the baseline performance on the female news anchor was relatively lower compared to the male news anchor. However, our proposed method consistently outperformed the baseline across different adaptation settings, demonstrating its effectiveness in improving lip reading performance. Furthermore, we did not observe the unusual performance decrease when using only 1-minute short adaptation data, as mentioned in the main submission. This suggests that the extreme situation and challenges faced by the CAS-VSR-S68h dataset may have different underlying factors that require further investigation in future research.

Overall, the additional experiments provide further evidence of the effectiveness of our method in improving lip reading performance. They highlight the importance of considering speaker diversity and addressing the challenges posed by different speakers in the dataset.

#### 2.3.2 More Qualitative Results

**Further Analysis of Alblation Study.** As shown in Figure 2, the enhancement or suppression modules would collapse to become indistinguishable for different speakers without $L^{Enhance}_{triple}$ and $L^{Suppress}_{triple}$. This emphasizes the necessity of $L^{Enhance}_{triple}$ and $L^{Suppress}_{triple}$ to ensure the enhancement and suppression module effectively capture and differentiate the characteristics of individual speakers.

**Visualization Analysis of Feature Suppression.** In our main submission, we primarily pre-

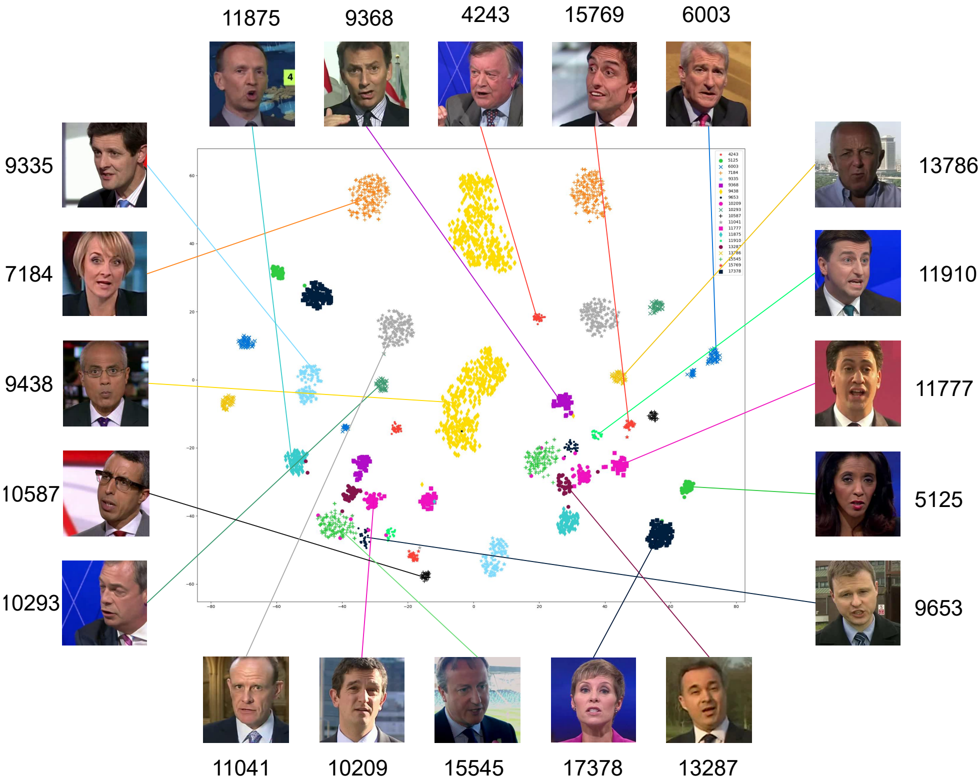


Figure 3: Clustering Visualization of the Learned ID Features Obtained through t-SNE Dimensionality Reduction on the LRW-ID Dataset.
Due to the similarity in colors of the figures presented in the main text, which made them less distinguishable, we have made modifications to the legend. We performed t-SNE dimensionality reduction on the same set of samples to obtain a clearer visualization. The clustering results in the revised figure show some shifting compared to the clusters mentioned in the original text.

sented the visualization of enhancement weights. However, in this supplementary material, we provide the visualization of suppression weights, which exhibit consistent behavior with the enhancement weights. As shown in left side of Figure 2, the visualization of suppression weights demonstrates similar patterns to the enhancement weights.

Specifically, when visualizing the enhancement weights for the same channel, we observe significant differences across different speakers. Similarly, this pattern becomes even more pronounced when examining the suppression weights( Three channels are randomly selected from the set of 64 channels as examples). In some cases, a specific region may undergo significant suppression for one speaker, while the suppression in the same region for another speaker may not be as prominent.

This consistent behavior between the visualization of enhancement and suppression weights further supports the effectiveness of our approach. It indicates that the model effectively learns to enhance content-dependent information and suppress content-independent information in a discriminative manner.

**Visualization Analysis of Speaker Features.** In order to gain a clearer understanding of the speaker features extracted by our model, we conducted a visualization analysis using t-SNE dimensionality reduction. Specifically, we visualized the speaker features for each speaker in the LRW-ID dataset, and we also associated each cluster with the appearance of speakers in the LRW-ID test set, as shown in Figure 3.

# References

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

[2] 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.

[3] 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.

[4] 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.

[5] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[6] 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.

[7] 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.