Supplementary Material of Audio-Guided Self-supervised Learning for Disentangled Visual Speech Representations

1 Datasets

Our main experiments are performed on two popular and challenging datasets: LRW and LRS2-BBC datasets, which are widely used for audio-visual speech-related tasks. Moreover, we also performed an evaluation on the setting of low-resource language with the GLips dataset.

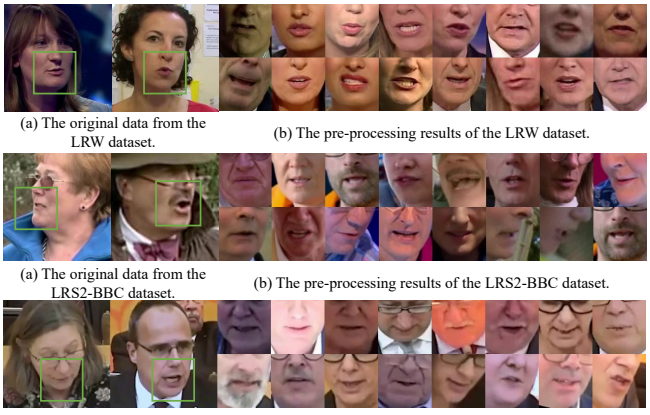
LRW dataset [1] is one of the most widely used word-level datasets for audio-visual speech-related tasks [2, 3, 4]. It contains more than a thousand speakers and covers a large variation of imaging conditions, such as head pose, illumination, gender, and so on. On this dataset, we crop each video using a fixed rectangle with coordinates of $(x_1, x_2, y_1, y_2) = (115, 210, 79, 174)$ to obtain a region of size 96×96 which will be taken as the input to feed our model, where (x_1, y_1) and (x_2, y_2) represent the coordinates of the upper-left corner and the lower-right corner of the rectangle, respectively. The results are as shown in images of the first row of Fig. 1

LRS2-BBC dataset [5] is one of the largest publicly available datasets for sentence-level audio-visual speech analysis. It covers a wide range of visual conditions and also many types of the video source, such as interviews, news, talk show, and so on. In this work, we align faces in each video to a template of size 256×256 by using similarity transformations and then crop the central 96×96 region as the input of each model. The results are as shown in images of the second row of Fig. 1.

GLips dataset [6] is a multimodal audio-visual speech dataset for the low-resource German language. It only consists of nearly half the number of videos compared to the LRW dataset. For this dataset, we only perform an evaluation to show our method's robustness to low-resource languages. We align faces in each video to a template of size 256×256 by using similarity transformations and then crop the central 96×96 region as the input of each model. The results are as shown in images of the last row of Fig. 1

2 Experimentalal Settings

For a clear comparison, we adopt the default training and testing splits on each dataset without introducing any extra data. For our two-branch network, we adopt the Adam optimizer, set the learning rate to 3×10^{-4}, and train it for 5 epochs on all the datasets.



(a) The original data from the LRW dataset. (b) The pre-processing results of the LRW dataset.

(a) The original data from the LRS2-BBC dataset. (b) The pre-processing results of the LRS2-BBC dataset.

(a) The original data from the GLips dataset. (b) The pre-processing results of the GLips dataset.

Fig. 1 Illustration of the data in different datasets. The specific region used in the experiments are shown as the green boxes.

When performing knowledge distillation from the model $\mathcal{M}_{F^S_{adj}}$ to the original video based VSR model $\mathcal{M}_{original}$, we optimize the model $\mathcal{M}_{original}$ using both the lip reading loss L_{VSR} and the knowledge distillation loss L_{KD}. Specifically, L_{VSR} is the visual speech recognition loss which usually takes the cross-entropy loss or the Connectionist Temporal Classification (CTC) loss. L_{KD} is Kullback-Leibler (KL) divergence-based loss for knowledge distillation. A weight coefficient is introduced to balance these two targets and the final loss function for $\mathcal{M}_{original}$ is as follows:

$$L_{\mathcal{M}_V} = \alpha L_{VSR} + (1 - \alpha) L_{KD}, \tag{1}$$

In this work, we set $\alpha = 0.5$ for simplicity.

3 Qualitative Results

In this part, we visualize the results of our method to show the effectiveness of our work for visual speech representation learning. We fixed the source frame I_i as the first frame in each video, i.e. $i = 1$, and only changed the target frame I_j and the corresponding audio signal a_j with $j \in \{1, 2, 3, ..., T_v\}$ as the input of the speech-relevant and speech-irrelevant branch.

The four rows in Fig. 2 are the sequence of the original video frames (a), the learned speech-relevant deformation flow $F^S_{1\to j}$ from I_1 to I_j (b), the warped frame I'^S_j based

Fig. 2 Results of our method on LRW when taking the first frame as the fixed source frame.

on $F^S_{1\to j}$ and I_1 (c), the learned speech-irrelevant deformation flow $F^{\bar{S}}_{1\to j}$ from I_1 to I_j (d), and the warped frame $I'^{\bar{S}}_j$ based on $F^{\bar{S}}_{1\to j}$ and I_1 (e) respectively.

By comparison of Fig. 2 (c) with Fig. 2 (a), we can see that the frames I'^S warped by speech-relevant deformation flow F^S only reflect lip movements while other parts (e.g. pose) remain the same as the source frame I_1. By comparison of Fig. 2 (e) with Fig. 2 (a), we can see that the frames $I'^{\bar{S}}$ warped by speech-irrelevant deformation flow $F^{\bar{S}}$ preserve the same speech status as the source frame I_1, while speech-irrelevant status, like the pose status, matches the target frames I_j. These results indicate that our proposed method can indeed disentangle the speech-relevant and speech-irrelevant information in videos.

4 Quantitative Results

4.1 Evaluation of the obtained visual speech representation

We further evaluate our method by taking the first frame as the fixed source frame of each video to generate the whole deformation flow sequence. This is a hard case for the model to acquire the speech cues because the target frame may be very different from the fixed source frame in several aspects beyond the speech content, leading to too much irrelevant changes between the source and target. Specifically, we fixed the branch $\mathbb{B}_S$ and feed the first frame of each video as the source frame $I_{i=1}$ and the target audio signal a_j in the video to generate the speech-relevant deformation flow sequence F^S_{fixed} for each video sample, where $F^S_{fixed} = \{F^S_{1\to 2}, F^S_{1\to 3}, ..., F^S_{1\to T_v}\}$. Then, we take the generated sequence F^S_{fixed} of each video to train a VSR model $\mathcal{M}_{F^S_{fixed}}$ with the same pipeline as $\mathcal{M}_{F^S_{original}}$. The F^S_{fixed}-based VSR model $\mathcal{M}_{F^S_{fixed}}$ achieves an accuracy of 86.8% (2.3% higher than $\mathcal{M}_{original}$). The only difference between $\mathcal{M}_{F^S_{fixed}}$ and $\mathcal{M}_{F^S_{original}}$ is the input, the original video frames for $\mathcal{M}_{F^S_{fixed}}$ and F^S_{fixed} for $\mathcal{M}_{F^S_{fixed}}$. This result indicates that our proposed representation is more discriminative in representing speech than the original video frames.

	Data Type (Model)	Accuracy (%) ↑
(a)	Original Video V ($\mathcal{M}^g_{original}$)	35.7
(b)	Speech-relevant Deformation Flow F^S_{adj} ($\mathcal{M}^g_{F^S_{adj}}$)	36.6
(c)	Flow-distilled Lip Reading Model ($\mathcal{M}'^g_{original}$)	35.9

Table 1 Evaluation on the low-resource GLips dataset.

4.2 Evaluation of the role of the bottleneck module

We show the effectiveness of the bottleneck module in preventing the speech-irrelevant branch from learning high-frequency and fine-grained speech cues. Specifically, we set the dimension of the bottleneck d_{bottle} to different values to train the proposed two-branch network respectively. Then, we adopt the fixed source frame setting as section 4.1 to clearly show the effect of the bottleneck. Specifically, we take the generated speech-relevant representations F^S_{fixed} under different settings of the bottleneck as input to train different VSR models. The models' accuracies are 86.8%, 84.0%, and 83.5%, when d_{bottle} is set to $3, 12, 32$, respectively. The performance decreases as the dimension of the bottleneck module d_{bottle} increases. This result demonstrates that the speech-irrelevant branch $\mathbb{B}_{\bar{S}}$ with a too-wide bottleneck tends to 'purloin' speech-relevant information and in turn make the speech-relevant representation not discriminative enough, which has also been inferred as a similar phenomenon in [7]. By default, we set the dimension d_{bottle} to 3 to make the bottleneck narrow enough, while still capable of expressing speech-irrelevant information.

4.3 Evaluation on the low-resource language

Our approach is based on the intrinsic difference between the frequency of speech-relevant and speech-irrelevant facial movements and does not rely on language-specific knowledge. To evaluate our method's effectiveness in handling low-resource languages, we conducted additional studies on a public German lip reading dataset GLips [6].

First, we train the original frame-based lip reading model $\mathcal{M}^g_{original}$ based on the original RGB frames only. The model achieves an accuracy of 35.7%, as shown in Tab. 1 (a).

Then, we train our two-branch network to obtain the deformation flow F^S_{adj} based sequence, which is then taken as the input to train a lip reading model $\mathcal{M}^g_{F^S_{adj}}$. The process to obtain the F^S_{adj} based sequence is illustrated in the experimental section of the main text. As shown in Tab. 1 (b), the model achieves a much higher accuracy of 36.6%, demonstrating that our proposed representations are indeed discriminative

in representing speech information even on low-resource languages.

Finally, we performed knowledge distillation from the speech-relevant representation-based lip reading model $\mathcal{M}^{g}_{F^{S}_{adj}}$ (Tab. 1 (b)) to the normal lip reading model $\mathcal{M}'^{g}_{original}$ (Tab. 1 (a)). The final accuracy of lip reading increased from 35.7% to 35.9% under this simple distillation, as shown in Tab. 1 (c). This further improvement shows again that our proposed method benefits the downstream lip reading task even in low-resource languages with simple logit-level knowledge distillation.

References

[1] Chung J S, Zisserman A. Lip reading in the wild. In: Proceedings of the Asian conference on computer vision. 2016, 87–103

[2] Stafylakis T, Tzimiropoulos G. Combining residual networks with lstms for lipreading. In: Proceedings of the Annual Conference of the International Speech Communication Association. 2017, 3652–3656

[3] Zhao X, Yang S, Shan S, Chen X. Mutual information maximization for effective lip reading. In: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition. 2020, 420–427

[4] Xiao J, Yang S, Zhang Y, Shan S, Chen X. Deformation flow based two-stream network for lip reading. In: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition. 2020, 364–370

[5] Afouras T, Chung J S, Senior A, Vinyals O, Zisserman A. Deep audio-visual speech recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 44(12): 8717–8727

[6] Schwiebert G, Weber C, Qu L, Siqueira H, Wermter S. A multimodal german dataset for automatic lip reading systems and transfer learning. arXiv preprint arXiv:2202.13403, 2022

[7] Qian K, Zhang Y, Chang S, Yang X, Hasegawa-Johnson M. AUTOVC: Zero-shot voice style transfer with only autoencoder loss. In: International Conference on Machine Learning. 2019, 5210–5219