Landmark-Guided Cross-Speaker Lip Reading with Mutual Information Regularization

Linzhi Wu[1,3], **Xingyu Zhang**[2*], **Yakun Zhang**[2,3], **Changyan Zheng**[2,4], **Tiejun Liu**[1], **Liang Xie**[2,3], **Ye Yan**[1,2,3], **Erwei Yin**[2,3]
[1]School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu, China
[2]Defense Innovation Institute, Academy of Military Sciences, Beijing, China
[3]Tianjin Artificial Intelligence Innovation Center, Tianjin, China
[4]High-tech Institute of Qingzhou, Weifang, China
lindgew@std.uestc.edu.cn, zhangxingyu1994@126.com, ykzhang1222@126.com
echoaimaomao@163.com, liutiejun@uestc.edu.cn, xielnudt@gmail.com
yy_taiic@163.com, yinerwei1985@gmail.com

Abstract

Lip reading, the process of interpreting silent speech from visual lip movements, has gained rising attention for its wide range of realistic applications. Deep learning approaches greatly improve current lip reading systems. However, lip reading in cross-speaker scenarios where the speaker identity changes, poses a challenging problem due to inter-speaker variability. A well-trained lip reading system may perform poorly when handling a brand new speaker. To learn a speaker-robust lip reading model, a key insight is to reduce visual variations across speakers, avoiding the model overfitting to specific speakers. In this work, in view of both input visual clues and latent representations based on a hybrid CTC/attention architecture, we propose to exploit the lip landmark-guided fine-grained visual clues instead of frequently-used mouth-cropped images as input features, diminishing speaker-specific appearance characteristics. Furthermore, a max-min mutual information regularization approach is proposed to capture speaker-insensitive latent representations. Experimental evaluations on public lip reading datasets demonstrate the effectiveness of the proposed approach under the intra-speaker and inter-speaker conditions.

Keywords: Lip reading, Cross-speaker, Lip landmark, Mutual information regularization

1. Introduction

Lip reading, commonly known as visual speech recognition (VSR), aims to automatically recognize spoken text units through the speaker's lip movements of a silent video clip, and is widely used in various potential applications such as aiding individuals with hearing impairments, speech recognition in noisy environments, human-computer interaction (Chung and Zisserman, 2016; Afouras et al., 2018c; Yang et al., 2019; Afouras et al., 2018a; Rekik et al., 2015). Recently lip reading research has made great progress thanks to the advent of deep learning and the availability of large-scale annotated corpus. Particularly, the advanced neural models adapted from the fields of automatic speech recognition (ASR) and natural language processing (NLP) significantly boost the performance of lip reading (Assael et al., 2016; Chung et al., 2017; Martínez et al., 2020; Ma et al., 2021, 2022).

Despite its success, lip reading still suffers from a non-trivial problem for practical usage, namely the considerable variations between speakers (Almajai et al., 2016; Burton et al., 2018). Conventional lip reading systems trained on a limited set of speakers tend to recognize the lip movements of specific individuals, and are easily sensitive to speaker variance, making them more suitable for overlapped speakers appeared in the training set. However, different speakers usually have different lip appearance and shapes even when they say the same utterances, and those systems may be prone to overfit the visual variations of lip region, which results in degraded performance when adapting to a speaker never seen before (Huang et al., 2021; Xue et al., 2023). Hence, it is essential to develop a lip reading system that can be generalized across speakers in favor of real-world applications.

To improve the robustness and accuracy of a lip reading model when dealing with unseen speakers, one intuitive solution is to eliminate the visual variations across speakers as much as possible. Since facial landmarks are sparse geometric coordinate points, indicating the location of key facial areas, they are robust to the pixel-based visual appearance and could serve as speaker-independent clues (Morrone et al., 2019). In (Xue et al., 2023), besides the visual features extracted from lip images, the authors introduced the facial landmarks to suppress the speaker variance in lip shapes and movements, achieving effective performance gains. In addition to the visual clues, another trend is to encourage the lip reading model to learn speaker-independent but speech content related visual representations by various means (Wand and Schmid-

[*] Corresponding author.

LREC-COLING 2024, pages 10023–10033

huber, 2017; Yang et al., 2020; Huang et al., 2021; Zhang et al., 2021; Lu et al., 2022).

Existing studies mostly take mouth-centered crops as input, but the visual variations of lip shapes and appearance may be inevitably introduced. To handle speaker variations, we rethink both the input visual clues and intermediate latent representations in this work. For the visual clues, we make the most of the lip landmarks and explore the landmark-guided fine-grained visual features from three aspects. First, we consider the landmark-centered patches as they are not only key areas closely related to lip reading, but also facilitate reducing lip shape variance. In particular, we extract tubelets (*i.e.*, 3D patches) centered on landmarks in view of both spatial and temporal dimensions. Second, to build the geometric correlation between different patches within each frame, the relative distances between landmarks are used as positional information to complement the geometric features. Moreover, as lip motion trajectories tend to be speaker-independent, we obtain the lip motion features from landmark tracks across frames by calculating the inter-frame difference of landmark coordinates. The aforementioned visual features can be regarded as the front-end obtained features, which are then fed into a back-end conformer encoder to model global and local temporal relationships (Gulati et al., 2020). Finally, a hybrid CTC/attention architecture (Hori et al., 2017; Petridis et al., 2018) is utilized for target text prediction.

Although the fine-grained local visual features induced by lip landmarks are expected to reduce the visual appearance variance, redundant speaker-specific characteristics may still be preserved within some patches. Therefore, we propose to leverage a max-min mutual information (MI) regularization scheme to decouple the identity-related features and speech-related features, facilitating speaker-insensitive latent representations. More specifically, in order to dig out speaker identity-related information, a speaker identification module is additionally introduced. Then, we minimize the MI between the speech-related features extracted by the conformer encoder and identity-related features extracted by the speaker identification module, while maximizing the MI between the representations encoded by the front-end and back-end of a lip reading model. In conclusion, the major contributions of this paper are summarized as follows:

- We investigate the landmark-guided fine-grained visual clues tailored for the cross-speaker lip reading task, with better interpretability in contrast to the widely-used mouth-cropped images.

- We propose a max-min mutual information regularization approach to encourage the lip reading model to learn speaker-insensitive latent representations.

- Experiments and analysis performed on public sentence-level lip reading datasets demonstrate the effectiveness of the proposed approach in the cross-speaker setting.

2. Methodology

Figure 1 briefly illustrates the overall model framework built on the joint CTC/attention architecture. It consists of several key components, each of which will be detailed in the following subsections.

2.1. Model Architecture

Let $\mathbf{x} = [x_1, x_2, \cdots, x_T]$ be the visual input streams of length T drawn from a facial video clip, mapped into the target text sequence $\mathbf{y} = [y_1, y_2, \cdots, y_N]$ with N tokens by a lip reading model. Suppose K lip landmarks are detected for each video frame in the pre-processing stage.

2.1.1. Landmark-Guided Visual Front-end

The mouth-centered cropped regions are commonly used as the visual clues. Nevertheless, local subtle lip dynamics (*e.g.*, mouth contours) may fail to to be effectively captured (Sheng et al., 2022). In addition, the whole mouth-cropped images potentially contain much speaker-related information (*e.g.*, personal appearance traits), resulting in more significant visual appearance difference across speakers. Thus, we investigate the landmark-based visual clues to capture local fine-grained lip movements and meanwhile reduce visual variance among different speakers. More specifically, we exploit the 2D landmarks of lip region from three distinct perspectives as follows:

3D Patches For each frame at time-step t, we first extract a small square window (*patch*)[1] of size $w_t \times w_t$ centered at each landmark position p_t^i $(1 \leq i \leq K)$, and the landmark-centered patch describes this point by the surrounding spatial context. Considering the neighboring temporal context of lip movements between adjacent frames, we further extract the 3-dimensional patch (*tubelet*) achieved by a 3D convolutional module, producing the tubelet embedding $\mathbf{v}_t^i$. In other words, we can construct a tubelet of size $w_t \times w_t \times d$ (d means the depth of patch determined by the kernel size in time dimension) around each lip landmark at each time-step. Unlike the whole mouth-cropped images

[1] For the sake of simplicity and ensuring a symmetric spatial context, the square-shape window is given preference.

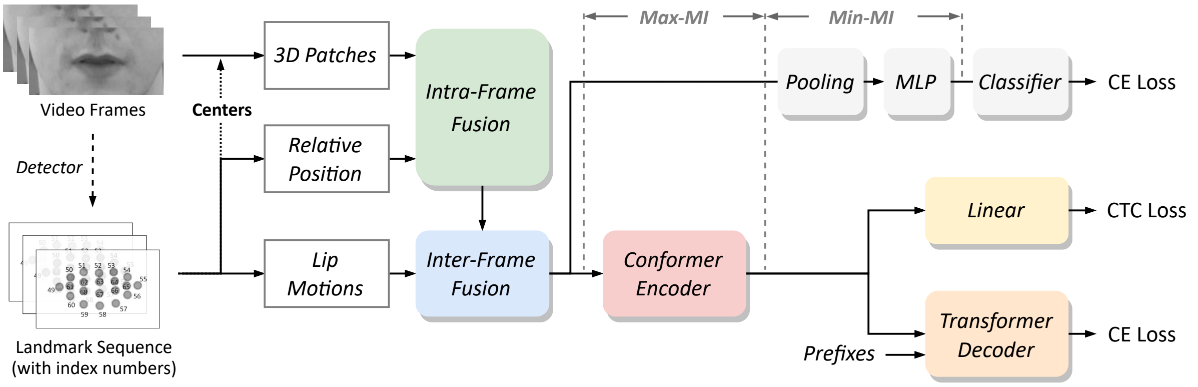


Figure 1: Illustration of the overall multi-task learning framework for cross-speaker lip reading. The model inputs are derived from the mouth-centered crops coupled with lip landmarks.

encoded by 3D convolution, the landmark-centered tubelet reduces the computational complexity of visual feature extraction. Moreover, less speaker identity-related information is retained.

Intra-Frame Relative Position As lip landmarks within a frame are not in fixed and regular position but distributed in a certain shape, the local patches alone may be not sufficient to learn good visual features. Thus, we consider the geometrical relationships between them within a frame by calculating the relative distance between any two landmark points. Specifically, we adopt a Multi-Layer Perceptron (MLP) layer to encode the coordinate differences between the i-th landmark and other landmarks at the t-th time-step, producing the relative positional vector defined as:

$$\mathbf{r}_t^i = \mathrm{MLP}(\{p_t^i - p_t^j\}_{i \neq j}), \quad (1 \leq j \leq K) \tag{1}$$

As a result, we have the position-aware visual feature at the i-th landmark: $\mathbf{u}_t^i = \mathbf{v}_t^i + \mathbf{r}_t^i$. In order to obtain the visual representations of the whole frame, we leverage a attention-weighted aggregation for the fine-grained features of all the landmarks at the t-step, allowing for the interaction between those tubelets. Concretely, a L_f-layer attentive encoder consumes the sequence of tubelet vectors $\mathbf{z}_t = [\mathbf{u}_t^1, \mathbf{u}_t^2, \cdots, \mathbf{u}_t^K]$ at the t-step. Each layer is composed of multi-head self-attention (MHSA) (Vaswani et al., 2017) and MLP blocks along with layer normalization (LN) as follows:

$$\begin{aligned} \mathbf{y}_t^{(l)} &= \mathrm{MHSA}(\mathrm{LN}(\mathbf{z}_t^{(l)})) + \mathbf{z}_t^{(l)}, \\ \mathbf{z}_t^{(l+1)} &= \mathrm{MLP}(\mathrm{LN}(\mathbf{y}_t^{(l)})) + \mathbf{y}_t^{(l)}, \\ &(l = 1 \cdots L_f) \end{aligned} \tag{2}$$

The outputs of the last layer ($\mathbf{z}_t^{(L_f)}$) followed by a global average pooling over all the landmarks produce the intra-frame visual features $\mathbf{f}_t$.

Inter-Frame Lip Motions We explicitly extract the inter-frame lip movement features derived from the lip landmark tracks. For the t-th time-step, we mainly consider the contour and geometric information involving lip dynamics, including the landmark's *x-y* coordinates; the height and width of outer and inner lip measured by Euclidean distance. Because these metrics may vary significantly when a speaker pronouncing. The motion vector can be obtained by computing the difference between two adjacent frames, *i.e.*, the current frame (t) simply subtracts the pre-frame ($t-1$). The motion vector of the first time-step can be set to zero. Here we use a 1-dimensional convolutional module to extract context-aware motion features $\mathbf{m}_t$. Furthermore, we combine the motion features with the orthogonal intra-frame visual features to generate the front-end visual representations: $\mathbf{h}_t = \mathbf{f}_t || \mathbf{m}_t$ $(t = 1 \cdots T)$ through a simple concatenation operation (||).

2.1.2. Conformer Back-end

In light of the sequential spatio-temporal properties of video data, the spatially dominant visual front-end mentioned above may fail to capture temporal dependencies between video frames effectively. Hence, we take advantage of the conformer encoder which integrates self-attention mechanisms and convolutional operations, to model global and local temporal dependencies across frames dynamically. The front-end visual features are passed through the conformer encoder with L_b sequentially stacked blocks with identical structure:

$$\begin{aligned} \mathbf{H}^{(0)} &= [\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_T], \\ \mathbf{H}^{(l)} &= \mathrm{ConformerBlock}(\mathbf{H}^{(l-1)}), \\ &(l = 1 \cdots L_b) \end{aligned} \tag{3}$$

Each conformer block with the macaron-like structure is composed of a set of stacked modules: a feed-forward module, a multi-head self-attention module, a convolution module and a second feed-forward module (Gulati et al., 2020).

Similar to ASR, the monotonic alignment property between input and target sequences is also supposed to be satisfied in VSR. To this end, the auxiliary CTC loss over the encoder outputs is applied to maximize the correct target alignments:

$$\mathcal{L}_{CTC} = -\log p_{\mathrm{CTC}}(\mathbf{y}|\mathbf{x}), \tag{4}$$

where $p_{\mathrm{CTC}}(\mathbf{y}|\mathbf{x}) \approx \prod_{t=1}^{T} p(y_t|\mathbf{x})$ based on conditional independence assumption between the predicted outputs.

2.1.3. Transformer Decoder

Autoregressively, a standard transformer decoder (Vaswani et al., 2017) is applied to receive the front-end encoded hidden representations and the prefixes of the target sequence to generate the next token of speech content. The decoder is comprised of a embedding layer followed by L_d stacked multi-head attention blocks. The sequence of prefixes is projected to embedding vectors, and then the absolute positional encoding is added. Each attention block consists of a masked multi-head self-attention module, an encoder-decoder multi-head attention module and a feed-forward module.

$$\begin{aligned} \mathbf{S}^{(0)} &= \mathrm{TE}(\hat{\mathbf{y}}) + \mathrm{PE}(\hat{\mathbf{y}}), \\ \mathbf{S}^{(l)} &= \mathrm{DecoderBlock}(\mathbf{S}^{(l-1)}, \mathbf{H}^{(L_b)}), \\ & (l = 1 \cdots L_d) \end{aligned} \tag{5}$$

where $\hat{\mathbf{y}}$ is the prefixes of target sequence, TE and PE denote the token embedding layer and positional encoding layer respectively. To generate the desired output sequence, the cross-entropy based training loss is defined to narrow the gap between the predicted sequence and the target sequence.

$$\mathcal{L}_{CE} = -\log p_{\mathrm{CE}}(\mathbf{y}|\mathbf{x}), \tag{6}$$

where $p_{\mathrm{CE}}(\mathbf{y}|\mathbf{x}) \approx \prod_{t=1}^{N} p(y_t|y_{<t}, \mathbf{x})$ based on the chain rule during the step-wise decoding process.

Finally, the training objective of the encoder-decoder architecture is calculated by a simple linear combination as follows:

$$\mathcal{L}_{VSR} = \lambda \mathcal{L}_{CTC} + (1-\lambda)\mathcal{L}_{CE}. \tag{7}$$

where the relative weight λ satisfies $0 \leq \lambda \leq 1$.

2.2. Speaker Identification

To allow for the subsequent decoupling of speaker-related features and speech content related features, the model should be able to distinguish different speakers from local lip appearance. Therefore, we introduce an additional speaker identification branch. The underlying speaker identity-related features are obtained via a MLP head that has a batch normalization layer, a ReLU activation function and a fully-connected (FC) layer. The speaker classifier with a FC layer is then used for speaker identification based on multi-class cross-entropy objective function.

$$\begin{aligned} \mathbf{h}^{ID} &= \mathrm{MLP}(\mathrm{GAP}(\mathbf{H}^{(0)})), \\ p^{ID} &= \mathrm{softmax}(\mathrm{FC}(\mathbf{h}^{ID})), \\ \mathcal{L}_{ID} &= -\sum_{c=0}^{C-1} y_c^{ID} \log p_c^{ID}, \end{aligned} \tag{8}$$

where GAP is the global average pooling layer used to temporally aggregate the front-end visual features (from Eq. 3), C refers to the number of all speakers, p_c^{ID} represents the class probability of the input sample belongs to speaker c while y_c^{ID} is the binary ground-truth label indicating whether this sample belongs to the speaker or not. The speaker identification branch is only used in the training phase.

2.3. Max-Min Mutual Information Regularization

Although the landmark-based visual clues potentially reduce the inter-speaker visual variations, there may still retain the redundant speaker-specific information within some patches. Therefore, we further exploit the speaker-insensitive features in latent representation space. To ensure the independence between speaker identity and speech content, we adopt the mutual information (MI) regularization method to facilitate learning of independent disentangled representations. Formally, the basic definition of MI between variables X and Y, establishing on the KL-divergence of their joint and marginal probability distributions:

$$\mathcal{I}(X,Y) = \mathrm{KL}\big(p(X,Y)||p(X)p(Y)\big), \tag{9}$$

For differentiable MI estimation, we introduce using the variational Contrastive Log-Ratio Upper Bound (vCLUB) (Cheng et al., 2020) as the upper bound of the desired MI $\mathcal{I}(X,Y)$, achieving MI minimization between latent representations.

$$\begin{aligned} \mathcal{I}_{\mathrm{vCLUB}}(X,Y) := \mathbb{E}_{p(X,Y)}\big[\log q_\phi(y|x)\big] - \\ \mathbb{E}_{p(X)}\mathbb{E}_{p(Y)}\big[\log q_\phi(y|x)\big], \end{aligned} \tag{10}$$

where the variational distribution $q_\phi(y|x)$ with parameter ϕ is applied to approximate the unknown conditional distribution $p(y|x)$.[2] Then, the training loss of minimizing the MI between the content-related features (Eq. 3) and the identity-related

[2] Note that $\mathcal{I}_{\mathrm{vCLUB}}(x,y)$ remains a MI upper bound when we have a good variational approximation $q_\phi(y|x)$ using a neural network.

(Eq. 8) features can be defined as $\mathcal{L}_{minMI} = \mathcal{I}_{\text{vCLUB}}(\mathbf{h}^{ID}, \mathbf{H}^{(L_b)})$.

Besides, we adopt a MI neural estimator based on the Jensen-Shannon divergence (Hjelm et al., 2019) as a lower bound of Eq. 9 to achieve MI maximization among latent representations.

$$\hat{\mathcal{I}}_{\theta}^{(\text{JSD})}(X,Y) := \mathbb{E}_{p(X,Y)}\big[-\log(1+e^{-\mathcal{F}_\theta(x,y)})\big] - \mathbb{E}_{p(X)p(Y)}\big[\log(1+e^{\mathcal{F}_\theta(x,y)})\big], \tag{11}$$

where $\mathcal{F}_\theta$ stands for a score function approximated by a MLP with learnable parameter θ. The front-end is encouraged to capture speaker irrelevant representations, accomplishing by maximizing the MI between features extracted from the front-end and back-end encoders. Thus, the training loss is defined as $\mathcal{L}_{maxMI} = -\hat{\mathcal{I}}_{\theta}^{(\text{JSD})}(\mathbf{H}^{(0)}, \mathbf{H}^{(L_b)})$ where $\mathbf{H}^{(0)}$ corresponds to the front-end visual features.

Consequently, the training objective of the MI estimators is to minimize the overall loss as follows:

$$\mathcal{L}_{MI} = \mathcal{L}_{minMI} + \mathcal{L}_{maxMI}. \tag{12}$$

2.4. Training and Decoding

Finally, to optimize the whole model parameters, we propose to utilize a two-stage training strategy:

Stage I We jointly train the VSR module and speaker identification module in a multi-task learning manner, until the speaker identification module converges.

$$\mathcal{L} = \mathcal{L}_{VSR} + \alpha_1 \mathcal{L}_{ID}. \tag{13}$$

Stage II We freeze the weights of the well-trained speaker identification module, and continue to train the VSR module along with the MI estimators.

$$\mathcal{L} = \mathcal{L}_{VSR} + \alpha_2 \mathcal{L}_{MI}. \tag{14}$$

where α_1 and α_2 are the weight coefficients. The first stage is to initialize the speaker identification module, ensuring that speaker identity-related features can be effectively extracted. While the second stage aims to encourage the speech content related features to be towards speaker-invariant. During inference, the transformer decoder is performed with a left-to-right beam search algorithm.

3. Experiments

3.1. Datasets and Evaluation

We conduct experiments on publicly available lip reading dataset GRID (Cooke et al., 2006). It is a popular sentence-level dataset, consisting of 34 speakers[3], and each speaker utters a set of 1000 sentences with fixed grammar. The duration of each recorded facial video clip is about 3 seconds, sampling 25 frames per second. Following Assael et al. (2016), we utilize the same unseen speaker split that four speakers (1, 2, 20 and 22) are used for testing and the rest for training. For the overlapped speaker setting, we randomly select 255 samples from each speaker for testing and the remaining samples for training. Detailed data statistics can be found in Table 1.

Setting	Subset	#Speaker	#Sentence
Overlap	Train	33	24408
	Test	33	8415
Unseen	Train	29	28837
	Test	4	3986

Table 1: Data statistics for the overlapped and unseen speaker settings.

To measure model performance, we use word error rate (WER) as evaluation protocol following previous literature (Assael et al., 2016; Chung et al., 2017). WER in percentage is calculated by comparing the number of substitutions (S), deletions (D), and insertions (I) required to transform the recognized output generated by a lip reading system into the reference transcription, divided by the total number of words in the reference transcription (N). Mathematically, the formula for calculating WER can be just defined as WER = (S + D + I) / N. Lower WER values indicate higher accuracy.

3.2. Implementation Details

For the pre-processing, we use the face alignment detector (Bulat and Tzimiropoulos, 2017) to detect and track 68 facial landmarks for each frame of video clips from the dataset[4]. We resize the original video into 360×288, and select all 20 lip landmarks aligned with the landmark point of the nose tip ($K = 20$). The basic size of each patch centered on a landmark point is 24×24. Moreover, all video frames are converted to grayscale and normalized by division by 255.

Table 2 shows the architecture of the 3D patch encoding module. For the conformer encoder, we use 3 blocks, hidden dim of 256, feed forward dim of 1024, 8 attention heads, and the kernel size of each depth-wise convolutional layer is set to 31. For the

[3] Data source: https://spandh.dcs.shef.ac.uk/gridcorpus. It is worth noting that the video data for speaker 21 is not available.

[4] Open-source toolkit: https://github.com/1adrianb/face-alignment. According to the same index numbers in consecutive frames, we can connect the corresponding landmark coordinates across frames and track those landmarks over time.

Layers	Filters	Output size
Conv3D	$5 \times 3 \times 3, 64$	$64 \times T \times \frac{H}{2} \times \frac{W}{2}$
MaxPool3D	$1 \times 3 \times 3, 64$	$64 \times T \times \frac{H}{4} \times \frac{W}{4}$
Conv2D	$3 \times 3, 128$	$T \times 128 \times \frac{H}{8} \times \frac{W}{8}$
Conv2D	$3 \times 3, 256$	$T \times 256 \times \frac{H}{16} \times \frac{W}{16}$
AvgPooling	–	$T \times 256$

Table 2: The 3D patch encoding module. Each convolution layer is followed by batch normalization and Swish activation function.

transformer decoder, the basic hyper-parameters are the same as in the conformer ($L_f = L_b = L_d = 3$). λ is set to 0.1 as suggested in (Ma et al., 2022), and the values of α_1, α_2 are empirically set to 0.2. During training, the Adam optimizer is used to update the learnable model parameters with a mini-batch size of 50. The initial learning rate is $3e^{-4}$, following a schedule strategy that increases linearly from 0 to the initial value and thereafter decreases with cosine annealing. In the testing phase, a beam search decoder is applied to the transformer decoder for character-level prediction with beam width 10, without using external language model.

Flexible Patch Size Compared with a fixed patch size, Beyer et al. (2023) have demonstrated the superiority of randomized patch sizes for a standard vision transformer (Dosovitskiy et al., 2021). Drawing inspiration from this, we try to dynamically change the size of landmark-centered patch at each iteration during training, which is implemented by randomly sampling a window size from a range of windows (e.g., $w \times w$, $w \in \{20, 22, 24, 26, 28, 30, 32\}$ with an interval of 2 pixels in this work).[5]

3.3. Ablation Analysis

We perform a series of ablation studies in the unseen speaker setting to better understand our method from different aspects. Results are shown in Table 3. First, performance drop can be observed when removing the relative position or lip motion features from the fine-grained visual clues of our pipeline. That verifies the importance of each part to enhance the visual features. Moreover, we ignore the temporal context information of the landmark-centered patch through replacing the tubelet with 2D patch, resulting in about 1.4% WER increase. Furthermore, we abandon the mutual information regularization terms from the framework, leading to consistent performance drops.

Here we consider adopting the commonly-used mouth-cropped images rather than the proposed fine-grained visual cues, and the results with performance degradation prove the benefits of our visual cues in improving the recognition performance of unseen speakers. The whole mouth regions and beyond may provide more complete and informative spatial context cues beneficial for lip reading (Zhang et al., 2020), but at the cost of preserving more speaker-specific characteristics that are not conducive to cross-speaker adaptation. Also, this motivates us to make a good trade-off between recognition accuracy and robustness when exploiting the visual features.

Method	**WER** (%)
Ours	10.21
w/o *RelPos*	10.69
w/o *Motion*	10.92
w/o *MI*	11.13
w/o *RelPos*&*Motion*	11.41
w/o *RelPos*&*MI*	11.65
w/o *Motion*&*MI*	11.77
w/o *RelPos*&*Motion*&*MI*	12.40
Replacing 3D patch with 2D patch	11.62
Using mouth-centered crops	11.50

Table 3: Ablation studies for unseen speakers. *RelPos*, *Motion* and *MI* mean relative positions among intra-frame landmarks, lip motion information, and mutual information regularization, respectively.

3.4. Comparison with Previous Methods

As shown in Table 4, we compare the proposed method with the previous competitive baselines for the overlapped and unseen speaker settings. We can observe that different lip reading methods indeed perform much better when handling those seen speakers. The comparison results also indicate that the proposed method can achieve performance on par with or exceeding those competitive methods in both settings, by attaining 1.83% WER and 10.21% WER in the seen and unseen speaker scenarios respectively. Since the core motivation of this work is not to pursue a new state-of-the-art, here we do not consider the mouth-cropped images like the previous methods as global visual cues having rich spatial information. Improved recognition performance may be further obtained through any feasible integration (Sheng et al., 2022; Xue et al., 2023) with the proposed fine-grained visual cues.

3.5. Effect of Patch Size

To analyze the effect of patch size on recognition performance, we further examine the landmark-centered patches with different sizes, as depicted in Fig. 2. Large patch size means more spatial contextual information around a landmark point, and

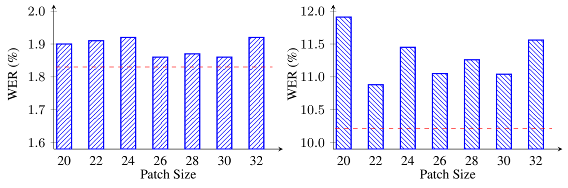
[5] Due to limited computational resources, we uniformly resize the large cropped patches to a proper resolution of 24×24, which is also used for model inference.

Method (Overlapped)	**WER** (%)
LipNet (Assael et al., 2016)	4.80
WAS (Chung et al., 2017)	3.00
LCANet (Xu et al., 2018)	2.90
DualLip (Chen et al., 2020)	2.71
CALLip (Huang et al., 2021)	2.48
LCSNet (Xue et al., 2022)	2.30
Ours	**1.83**
Method (Unseen)	**WER** (%)
LipNet (Assael et al., 2016)	14.2
WAS (Chung et al., 2017)	14.6
TM-seq2seq (Afouras et al., 2018a)	11.7
Motion&Content (Riva et al., 2020)	19.8
PCPG-seq2seq (Luo et al., 2020)	12.3
LCSNet (Xue et al., 2022)	11.6
Ours	**10.21**

Table 4: Performance comparison with previous competitive baseline methods.

vice versa. We observe that: (1) Patch size has a less impact on the performance of overlapped speakers compared to unseen speakers. It may be attributed to a fact that the local appearance around a landmark point is similar for a speaker who has already been seen. (2) For unseen speaker setting, smaller patches may provide less spatial contexts, while larger patches may lead to redundant visual information since lip landmarks are closely arranged. Thus, a moderate patch size ensures good recognition results for unseen speakers.

Instead of using a single patch size, we propose to utilize flexible patch size (FPS) for the lip-reading model training by patch sampling strategy. The comparison demonstrates that FPS can lead to better recognition performance (*i.e.*, dashed lines), especially for the unseen speaker setting. FPS can actually be regarded as a spatial context augmentation, and a lip reading model trained at FPS enables receive patches of variable scales in comparison to that trained at a single fixed patch size.



(a) Overlapped Speaker (b) Unseen Speaker

Figure 2: Performance comparison of different patch size (ranged from 20 to 32) in the overlapped and unseen speaker settings. The dashed line indicates the recognition performance using flexible patch size.

3.6. Attention Visualization

As mentioned in subsection 2.1.1, a multi-head attentive fusion module is used to aggregate the features of 3D patches centered on lip landmarks within a frame. We examine all 20 lip landmarks out of 68 facial landmarks, indexed from 49 to 68. Figure 3 presents the attention maps of a sampled video clip produced by the attentive fusion module. The weights are calculated by averaging over all the self-attention heads at all layers, with values suggesting the importance between the landmarks. We can observe that not all landmark-centered areas are non-trivial across frames, and the module pays more attention to the areas around the corners of the mouth (*e.g.*, landmark 49, 54, 55, 57, 61, etc). One possible reason is that the visual movements of those areas are relatively more evident at a local scale (*"view"*), when a speaker utters with his mouth open and closed. This finding may help us determine lip landmarks that need to be processed, leading to reduced computation overhead.

4. Related Work

4.1. Lip Reading

Lip reading technique is essentially the translation of lip movements related visual signals into corresponding transcribed text. For the visual signal input, aligned mouth regions of interest cropped by detected landmarks are the most commonly-used (Chung and Zisserman, 2016; Chung et al., 2017; Afouras et al., 2018a; Petridis et al., 2018; Ma et al., 2021, 2022; Xue et al., 2023). Using extraoral regions (*e.g.*, the upper face and cheeks) also helps boost recognition performance (Zhang et al., 2020). For the textual output, lip reading paradigms can be broadly categorized as word-level setting and sentence-level setting. The former aims to map a video frame sequence into isolated units with limited number (*e.g.*, digits, letters or words), which is extensively explored by early research efforts (Chung and Zisserman, 2016; Yang et al., 2019; Martínez et al., 2020; Zhao et al., 2020). The latter is challenging yet practical, mapping a video frame sequence into a spoken sentence. Typically, the model architecture of the two paradigms consists of 3D and 2D convolutional layers (*e.g.*, ResNet (He et al., 2016)) as the front-end, and sequential models as the back-end, such as RNN (Assael et al., 2016; Chung et al., 2017), TCN (Martínez et al., 2020) and Transformer (Afouras et al., 2018a,b). Unlike the word-level trained with simple classification loss, sentence-level lip reading models usually train with Connectionist Temporal Classification (CTC) (Graves et al., 2006) or sequence-to-sequence (Sutskever et al., 2014; Vaswani et al., 2017) fashion to achieve effective performance.

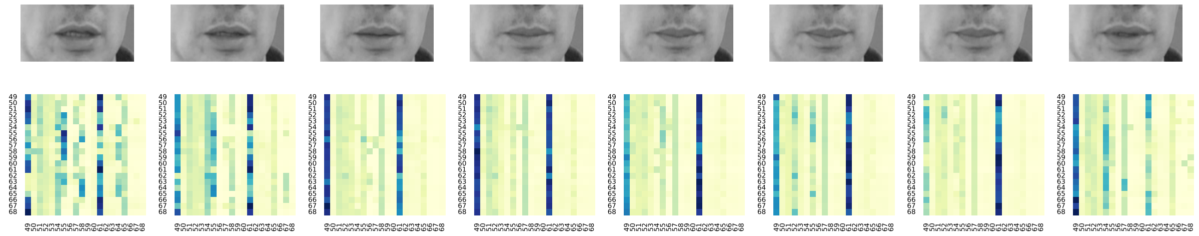


Figure 3: Attention weight maps between lip landmarks (indices from 49 to 68) from the attentive intra-frame fusion module. The weights are calculated by averaging over all the self-attention heads. The video clip used here is drawn from the test set. Darker colors indicate larger weight values.

Even with the great progress, existing lip reading systems are restricted in the limited number of speaker, leading to speaker dependency problems. Due to the visual variations of lip movements across speakers, these systems enable achieve considerable performance for overlapped speakers in training set, whereas obtain significant performance drops for unseen speakers. To eliminate the variations, on the one hand, enhancing the input visual clues may be a straightforward way. Riva et al. (2020) used the motion dynamics derived from simple adjacent-frame differences to improve the performance. Xue et al. (2023) used facial landmarks to complement the features extracted from lip images. On the other hand, several previous studies instead pay more attention to learning speaker-independent features that are robust to speaker identity information, such as adversarial training (Wand and Schmidhuber, 2017), disentangled representation learning (Zhang et al., 2021; Lu et al., 2022), and speaker normalization (Yang et al., 2020; Huang et al., 2021).

In this work, we focus on the sentence-level lip reading. Unlike the previous studies, considering that the mouth-centered crops might preserve more speaker-related features (*e.g.*, the beard or mole around mouth) irrelevant to speech content recognition, we explore the landmark-guided fine-grained visual clues to reduce visual appearance variance. Moreover, a mutual information regularization scheme is proposed to encourage both the front- and back-end of a lip reading model to learn speaker-insensitive latent representations.

4.2. Landmark-based Visual Features

Facial landmarks, referring to specific coordinate points on a person's face that are used to locate key facial areas (*e.g.*, eyes, eyebrows, nose and lip), have attracted increasing attention in visual speech-related fields over recent years. One advantage of landmark points is that they outline the overall shape of facial key areas in a sparse positional encoding way, and further geometric and contour features can be easily derived (Cetingul et al., 2006; Kumar et al., 2007; Zhou et al., 2011), effectively describing lip motion irrespective of speakers. Morrone et al. (2019) applied motion features based on facial landmarks to improve speech enhancement in a multi-speaker scenario, and proved that the advantages of motion-based features over position-based features. To fully exploit the characteristics of lip dynamics, Sheng et al. (2022) leveraged Graph Convolution Network (GCN) to model dynamic mouth contours and capture local subtle movements, improving recognition performance by enhancing visual feature representations. Since landmark-based features are less affected by visual variations caused by lip shapes and appearance, Xue et al. (2023) introduced the facial landmarks as complementary feature to the visual appearance of lip regions via a cross-modal fusion manner, eliminating biased visual variations between speakers and yield improved performance and robustness for unseen speakers. Motivated by the success of landmarks, in this work, we further investigate the lip-landmark guided visual clues for facilitating generalization to unseen speakers.

4.3. Mutual Information Regularization

Mutual information (MI) is typically employed as a measure of the amount of information that one random variable reveals about the other (Kinney and Atwal, 2014). It quantifies the dependence between two variables. In the context of (unsupervised) representation learning, through maximizing the MI, the model is enforced to capture meaningful dependencies or relevance between different feature representations, and vice versa. Actually, MI is hard to exactly calculate in the high-dimensional and continuous cases. Thus, various efficient neural estimation methods have been proposed over recent years as approximate solutions (Belghazi et al., 2018; van den Oord et al., 2018; Hjelm et al., 2019; Cheng et al., 2020). Krishna et al. (2019) improved the image-to-question generation model by maximizing the MI between the image, expected answer, and generated question. Zhu et al. (2018) tried to solve talking face generation generation problem by MI maximization between word distribution and other modal distribution. Similarly, to improve the lip reading performance, Zhao et al. (2020) utilized the global and local MI maximization constraints to extract discriminative features.

Different from previous works, we introduce a MI regularization term to learn informative representations for cross-speaker adaptation in lip reading. Instead of relying on sample pairs from the same or different speakers as model input (Yang et al., 2020; Zhang et al., 2021; Lu et al., 2022), we try to minimize the MI between speaker-dependent features and content-dependent features for the purpose of decoupling, while maximizing the MI between the front-end encoded features and the back-end encoded features.

5. Conclusion

In this paper, we provide insights into the cross-speaker lip reading task in terms of visual clues and latent representations, aiming to reduce visual appearance variations across speakers. On the basis of the hybrid CTC/attention architecture, we propose to exploit the landmark-guided fine-grained visual clues as model input features, while introducing the max-min mutual information regularization to learn speaker-insensitive latent representations via a two-stage optimizing scheme. The experimental results evaluated on the sentence-level lip reading demonstrate the effectiveness of the proposed approach.

6. Limitations

One potential drawback of softmax activation in the speaker identification module is that it fails to encourage cluster compactness and cannot ensure the similarity among samples within the same category. To address this problem, the AM-Softmax loss (Wang et al., 2018), an enhanced version of softmax, may help better learn speaker discriminative representations. In addition, the performance of the current lip reading system still has room for further advancement. One possible way to improve is to make the most of the mouth-cropped images and beyond as complementary information. We leave these for future research.

7. Acknowledgements

This work was supported in part by the grants from the National Natural Science Foundation of China under Grant 62332019 and 62076250, the National Key Research and Development Program of China (2023YFF1203900, 2023YFF1203903).

8. References

Triantafyllos Afouras, Joon Son Chung, Andrew W. Senior, Oriol Vinyals, and Andrew Zisserman. 2018a. Deep audio-visual speech recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(12):8717–8727.

Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. 2018b. Deep lip reading: A comparison of models and an online application. In *Interspeech 2018*, pages 3514–3518.

Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. 2018c. LRS3-TED: a large-scale dataset for visual speech recognition. *CoRR*, abs/1809.00496.

Ibrahim Almajai, Stephen J. Cox, Richard W. Harvey, and Yuxuan Lan. 2016. Improved speaker independent lip reading using speaker adaptive training and deep neural networks. In *ICASSP 2016*, pages 2722–2726.

Yannis M. Assael, Brendan Shillingford, Shimon Whiteson, and Nando de Freitas. 2016. Lipnet: Sentence-level lipreading. *CoRR*, abs/1611.01599.

Ishmael Belghazi, Sai Rajeswar, Aristide Baratin, R. Devon Hjelm, and Aaron C. Courville. 2018. MINE: mutual information neural estimation. *CoRR*, abs/1801.04062.

Lucas Beyer, Pavel Izmailov, Alexander Kolesnikov, Mathilde Caron, Simon Kornblith, Xiaohua Zhai, Matthias Minderer, Michael Tschannen, Ibrahim Alabdulmohsin, and Filip Pavetic. 2023. Flexivit: One model for all patch sizes. In *CVPR 2023*, pages 14496–14506.

Adrian Bulat and Georgios Tzimiropoulos. 2017. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *ICCV 2017*, pages 1021–1030.

Jake Burton, David Frank, Mahdi Saleh, Nassir Navab, and Helen L. Bear. 2018. The speaker-independent lipreading play-off; a survey of lipreading machines. In *IPAS 2018*, pages 125–130.

H Ertan Cetingul, Yücel Yemez, Engin Erzin, and A Murat Tekalp. 2006. Discriminative analysis of lip motion features for speaker identification and speech-reading. *IEEE Transactions on Image Processing*, 15(10):2879–2891.

Weicong Chen, Xu Tan, Yingce Xia, Tao Qin, Yu Wang, and Tie-Yan Liu. 2020. Duallip: A system for joint lip reading and generation. In *MM '20: The 28th ACM International Conference on Multimedia*, pages 1985–1993.

Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. 2020. CLUB: A contrastive log-ratio upper bound of mutual information. In *ICML 2020*, volume 119, pages 1779–1788.

Joon Son Chung, Andrew W. Senior, Oriol Vinyals, and Andrew Zisserman. 2017. Lip reading sentences in the wild. In *CVPR 2017*, pages 3444–3453.

Joon Son Chung and Andrew Zisserman. 2016. Lip reading in the wild. In *ACCV 2016*, volume 10112, pages 87–103.

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

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR 2021*.

Alex Graves, Santiago Fernández, Faustino J. Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *ICML 2006*, volume 148, pages 369–376.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition. In *Interspeech 2020*, pages 5036–5040.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR 2016*, pages 770–778.

R. Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Philip Bachman, Adam Trischler, and Yoshua Bengio. 2019. Learning deep representations by mutual information estimation and maximization. In *ICLR 2019*.

Takaaki Hori, Shinji Watanabe, and John Hershey. 2017. Joint CTC/attention decoding for end-to-end speech recognition. In *ACL 2017*, pages 518–529.

Yiyang Huang, Xuefeng Liang, and Chaowei Fang. 2021. Callip: Lipreading using contrastive and attribute learning. In *MM '21: ACM Multimedia Conference*, pages 2492–2500.

Minsu Kim, Hyunjun Kim, and Yong Man Ro. 2022. Speaker-adaptive lip reading with user-dependent padding. In *ECCV 2022*, volume 13696, pages 576–593.

Justin B Kinney and Gurinder S Atwal. 2014. Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences*, 111(9):3354–3359.

Ranjay Krishna, Michael S. Bernstein, and Li Fei-Fei. 2019. Information maximizing visual question generation. In *CVPR 2019*, pages 2008–2018.

Kshitiz Kumar, Tsuhan Chen, and Richard M. Stern. 2007. Profile view lip reading. In *ICASSP 2007*, pages 429–432.

Longbin Lu, Xuebin Xu, and Jun Fu. 2022. Siamese decoupling network for speaker-independent lipreading. *Journal of Electronic Imaging*, 31(3):033045–033045.

Mingshuang Luo, Shuang Yang, Shiguang Shan, and Xilin Chen. 2020. Pseudo-convolutional policy gradient for sequence-to-sequence lipreading. In *15th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2020*, pages 273–280.

Pingchuan Ma, Stavros Petridis, and Maja Pantic. 2021. End-to-end audio-visual speech recognition with conformers. In *ICASSP 2021*, pages 7613–7617.

Pingchuan Ma, Stavros Petridis, and Maja Pantic. 2022. Visual speech recognition for multiple languages in the wild. *Nat. Mac. Intell.*, 4(11):930–939.

Brais Martínez, Pingchuan Ma, Stavros Petridis, and Maja Pantic. 2020. Lipreading using temporal convolutional networks. In *ICASSP 2020*, pages 6319–6323.

Giovanni Morrone, Sonia Bergamaschi, Luca Pasa, Luciano Fadiga, Vadim Tikhanoff, and Leonardo Badino. 2019. Face landmark-based speaker-independent audio-visual speech enhancement in multi-talker environments. In *ICASSP 2019*, pages 6900–6904.

Stavros Petridis, Themos Stafylakis, Pingchuan Ma, Georgios Tzimiropoulos, and Maja Pantic. 2018. Audio-visual speech recognition with a hybrid ctc/attention architecture. In *2018 IEEE*

Spoken Language Technology Workshop, SLT 2018, pages 513–520.

Andrés Prados-Torreblanca, José Miguel Buenaposada, and Luis Baumela. 2022. Shape preserving facial landmarks with graph attention networks. In *BMVC 2022*, page 155.

Prajit Ramachandran, Barret Zoph, and Quoc V. Le. 2018. Searching for activation functions. In *ICLR 2018, Workshop Track Proceedings*.

Ahmed Rekik, Achraf Ben-Hamadou, and Walid Mahdi. 2015. Human machine interaction via visual speech spotting. In *ACIVS 2015*, volume 9386, pages 566–574.

Matteo Riva, Michael Wand, and Jürgen Schmidhuber. 2020. Motion dynamics improve speaker-independent lipreading. In *ICASSP 2020*, pages 4407–4411.

Timothy Israel Santos, Andrew Abel, Nick Wilson, and Yan Xu. 2021. Speaker-independent visual speech recognition with the inception V3 model. In *IEEE Spoken Language Technology Workshop, SLT 2021*, pages 613–620.

Changchong Sheng, Xinzhong Zhu, Huiying Xu, Matti Pietikäinen, and Li Liu. 2022. Adaptive semantic-spatio-temporal graph convolutional network for lip reading. *IEEE Trans. Multim.*, 24:3545–3557.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *NeurIPS 2014*, pages 3104–3112.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS 2017*, pages 5998–6008.

Michael Wand and Jürgen Schmidhuber. 2017. Improving speaker-independent lipreading with domain-adversarial training. In *Interspeech 2017*, pages 3662–3666.

Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu. 2018. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7):926–930.

Kai Xu, Dawei Li, Nick Cassimatis, and Xiaolong Wang. 2018. Lcanet: End-to-end lipreading with cascaded attention-ctc. In *13th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2018*, pages 548–555.

Feng Xue, Yu Li, Deyin Liu, Yincen Xie, Lin Wu, and Richang Hong. 2023. Lipformer: Learning to lipread unseen speakers based on visual-landmark transformers. *IEEE Transactions on Circuits and Systems for Video Technology*.

Feng Xue, Tian Yang, Kang Liu, Zikun Hong, Mingwei Cao, Dan Guo, and Richang Hong. 2022. Lcsnet: End-to-end lipreading with channel-aware feature selection. *ACM Trans. Multim. Comput. Commun. Appl.*, 19(1s):28:1–28:21.

Chenzhao Yang, Shilin Wang, Xingxuan Zhang, and Yun Zhu. 2020. Speaker-independent lipreading with limited data. In *ICIP 2020*, pages 2181–2185.

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

Qun Zhang, Shilin Wang, and Gongliang Chen. 2021. Speaker-independent lipreading by disentangled representation learning. In *ICIP 2021*, pages 2493–2497.

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

Xing Zhao, Shuang Yang, Shiguang Shan, and Xilin Chen. 2020. Mutual information maximization for effective lip reading. In *15th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2020*, pages 420–427.

Ya Zhao, Rui Xu, and Mingli Song. 2019. A cascade sequence-to-sequence model for chinese mandarin lip reading. In *MM Asia '19: ACM Multimedia Asia*, pages 32:1–32:6.

Ziheng Zhou, Guoying Zhao, and Matti Pietikäinen. 2011. Towards a practical lipreading system. In *CVPR 2011*, pages 137–144.

Hao Zhu, Aihua Zheng, Huaibo Huang, and Ran He. 2018. High-resolution talking face generation via mutual information approximation. *CoRR*, abs/1812.06589.