Can We Read Speech Beyond the Lips? 
Rethinking RoI Selection for Deep Visual Speech Recognition

Yuanhang Zhang\textsuperscript{1,2}, Shuang Yang\textsuperscript{1}, Jingyun Xiao\textsuperscript{1,2}, Shiguang Shan\textsuperscript{1,2}, Xiulin Chen\textsuperscript{1,2}

\textsuperscript{1} Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China
\textsuperscript{2} University of Chinese Academy of Sciences, Beijing 100049, China

Abstract—Recent advances in deep learning have heightened interest among researchers in the field of visual speech recognition (VSR). Currently, most existing methods equate VSR with automatic lip reading, which attempts to recognise speech by analysing lip motion. However, human experience and psychological studies suggest that we do not always fix our gaze at each other’s lips during a face-to-face conversation, but rather scan the whole face repetitively. This inspires us to revisit a fundamental yet somehow overlooked problem: can VSR models benefit from reading extraoral facial regions, i.e. beyond the lips? In this paper, we perform a comprehensive study to evaluate the effects of different facial regions with state-of-the-art VSR models, including the mouth, the whole face, the upper face, and even the cheeks. Experiments are conducted on both word-level and sentence-level benchmarks with different characteristics. We find that despite the complex variations of the data, incorporating information from extraoral facial regions, even the upper face, consistently benefits VSR performance. Furthermore, we introduce a simple yet effective method based on Cutout to learn more discriminative features for face-based VSR, hoping to maximise the utility of information encoded in different facial regions. Our experiments show obvious improvements over existing state-of-the-art methods that use only the lip region as inputs, a result we believe would probably provide the VSR community with some new and exciting insights.

I. INTRODUCTION

Visual speech recognition (VSR) is the task of recognising speech by analysing video sequences of people speaking. A robust VSR system has a variety of useful applications, such as silent speech interfaces [29], audio-visual speech recognition (AVSR) in noisy environments [1], face liveness detection, and so on. Its performance has progressed significantly over the past few years, thanks to several successful deep learning architectures, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs). It also benefits from the emergence of large-scale, in-the-wild audiovisual datasets, from which deep neural networks can automatically learn strong representations that outperform previous hand-crafted features.

Traditionally, the term “visual speech recognition” is used almost interchangeably with “lip reading” within the VSR community, since it is usually believed that lip shapes and lip motion most likely contain almost all the information correlated with speech. The information in other facial regions is considered by default weak, and not helpful for VSR.

This work was done by Yuanhang Zhang during his internship at Institute of Computing Technology, Chinese Academy of Sciences.

Can We Read Speech Beyond the Lips? 
Rethinking RoI Selection for Deep Visual Speech Recognition

Yuanhang Zhang\textsuperscript{1,2}, Shuang Yang\textsuperscript{1}, Jingyun Xiao\textsuperscript{1,2}, Shiguang Shan\textsuperscript{1,2}, Xiulin Chen\textsuperscript{1,2}

\textsuperscript{1} Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China
\textsuperscript{2} University of Chinese Academy of Sciences, Beijing 100049, China

Abstract—Recent advances in deep learning have heightened interest among researchers in the field of visual speech recognition (VSR). Currently, most existing methods equate VSR with automatic lip reading, which attempts to recognise speech by analysing lip motion. However, human experience and psychological studies suggest that we do not always fix our gaze at each other’s lips during a face-to-face conversation, but rather scan the whole face repetitively. This inspires us to revisit a fundamental yet somehow overlooked problem: can VSR models benefit from reading extraoral facial regions, i.e. beyond the lips? In this paper, we perform a comprehensive study to evaluate the effects of different facial regions with state-of-the-art VSR models, including the mouth, the whole face, the upper face, and even the cheeks. Experiments are conducted on both word-level and sentence-level benchmarks with different characteristics. We find that despite the complex variations of the data, incorporating information from extraoral facial regions, even the upper face, consistently benefits VSR performance. Furthermore, we introduce a simple yet effective method based on Cutout to learn more discriminative features for face-based VSR, hoping to maximise the utility of information encoded in different facial regions. Our experiments show obvious improvements over existing state-of-the-art methods that use only the lip region as inputs, a result we believe would probably provide the VSR community with some new and exciting insights.

I. INTRODUCTION

Visual speech recognition (VSR) is the task of recognising speech by analysing video sequences of people speaking. A robust VSR system has a variety of useful applications, such as silent speech interfaces [29], audio-visual speech recognition (AVSR) in noisy environments [1], face liveness detection, and so on. Its performance has progressed significantly over the past few years, thanks to several successful deep learning architectures, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs). It also benefits from the emergence of large-scale, in-the-wild audiovisual datasets, from which deep neural networks can automatically learn strong representations that outperform previous hand-crafted features.

Traditionally, the term “visual speech recognition” is used almost interchangeably with “lip reading” within the VSR community, since it is usually believed that lip shapes and lip motion most likely contain almost all the information correlated with speech. The information in other facial regions is considered by default weak, and not helpful for VSR.

This work was done by Yuanhang Zhang during his internship at Institute of Computing Technology, Chinese Academy of Sciences.
to handle pose variation? Answers to these questions are not evident or straightforward, and there has been no consensus or a universal guideline for choosing RoIs until now. In fact, RoIs are often chosen based on intuition and the researchers’ experience. The choices can be very different for different datasets and different methods, making transfer learning and cross-dataset evaluation difficult. Meanwhile, this specialised RoI selection step separates VSR from other face-related tasks, hindering further joint analysis. Using mouth inputs alone makes VSR an isolated problem, while including other facial regions opens up the possibility of various research in affective analysis and visual speech understanding. For example, joint modeling of non-verbal behaviour and verbal behaviour has been shown to be beneficial for learning adaptive word representations [32].

Finally, data encountered in real-world applications may not have the same luxury that some academic datasets enjoy, where data is biased towards frontal or near-frontal views, and high-resolution mouth crops are easily obtainable. Models trained on such well-behaved, cropped data may not perform as well when presented with the various challenges in practice. This severely limits the potential usage scenarios of VSR systems.

Motivated by these observations, we conduct a comparative study on input RoI selection, to (a) quantitatively estimate the contribution of different facial regions to the VSR task, and (b) determine whether state-of-the-art VSR networks, when presented with complex in-the-wild data, still benefit from additional clues within extraoral regions. Previous studies have explored different RoIs in the lower face [17], [24], and demonstrated the importance of suitable RoI coverage and selection. However, these attempts are limited to the lower face, and relatively small datasets with narrow vocabularies. For this study, we approach the problem with state-of-the-art, deep VSR models trained on large-scale “in-the-wild” VSR datasets, which depict many real-world variations, such as pose, lighting, scale, background clutter, makeup, expression, different speaking manners, etc. This is a much fairer reflection of real-world scenarios. Besides, we propose Cutout as an effective approach to encourage the model to utilise all facial regions. This allows researchers to sidestep the ambiguous choice of selecting appropriate RoIs, enhances the visual features, and increases the model’s robustness to mild occlusion within the mouth or other facial regions.

II. RELATED WORK

A. Visual Speech Recognition

Visual speech recognition (VSR), commonly referred to as automatic lip reading, is a classical problem in computer vision and has received increased interest over recent years. The combination of deep learning approaches and large-scale audiovisual datasets has been highly successful, achieving remarkable word recognition rates and even surpassing human performance. The deep network approach has become increasingly common and mainstream [9], [4], [35]. These methods have retracted the success of the CNN-LSTM-DNN (CLDNN) paradigm [22] in the automated speech recognition (ASR) community. However, though many works have reported state-of-the-art results on multiple challenging datasets, few have investigated the influence of input RoI selection, which can be a nuisance due to variations in mouth shapes, face geometry, and pose, as well as the relatively unknown effect of spatial context and extraoral regions. Unlike face recognition, where face frontalisation is a recognised need for pose invariance, and has been thoroughly investigated as an important part of the pipeline, VSR researchers tend to specify RoIs based on their own experience, and in a dataset-specific manner.

In some controlled-environment datasets, e.g. OuluVS [37] and GRID [11], the subjects remain relatively still while speaking. Lombard GRID [2] uses a head-mounted camera which the speakers face directly at for the frontal view, essentially fully removing head motion. The Lip Reading in the Wild dataset [9] is of short duration with small or no face scale variation within clips, and loose registration is enforced by aligning nose centers. A fixed, mouth-centered (and sometimes affine-transformed) rectangular RoI is preferable on these datasets, because the faces are usually stable, frontal or near-frontal, and of uniform sizes. In contrast, OuluVS2 [3] is a multi-view dataset, and RoIs are processed separately for each camera viewpoint. Chung and Zisserman [10] are the first to investigate large-scale deep lip reading in profile, and use an extended bounding box covering the whole face to account for significant pose variations. Yang et al. [35] determine the RoI size based on the distance between the nose and the mouth center and the width of the speaker’s lips, which also effectively handles pose variations.

There have been some previous attempts to address the RoI selection problem explicitly. The most relevant work is [17], which perform experiments with rectangular lower face regions with different spatial context and resolution in a connected digits recognition task. [24] experiments with optical flow features in non-rectangular RoIs after removing head motion, and obtain results better than rectangular RoIs. However, these work are limited by the amount of data used and model capacity, and only investigate the utility of the lower face. We adopt face inputs, and mimic real-world scenarios by experimenting on large-scale, in-the-wild VSR benchmarks that are a magnitude larger than those used in the above two papers.

B. Human Perception Studies

Our intuition that we do not fix our gaze at the speaker’s lip region when communicating with others is supported by a number of psychology studies on human gaze behaviour during visual and audio-visual speech perception [30], [31], [18]. It has been reported that human gaze patterns usually involve repetitive transitioning between the eyes and the mouth, even at high noise levels and when audio is absent [30]. Interestingly, [18] suggests that in a visual-only scenario, speechreading accuracy was related to the difficulty of the presented sentences and individual proficiency, but not...
to the proportion of gaze time at a specific part of the face. The roles of the upper face, which is less intuitive for VSR, have also been studied [13]. Studies in audiovisual speech perception show that head and eyebrow motion can help discriminate prosodic contrasts [12], which may be helpful for the VSR task. Moreover, in tonal languages like Mandarin Chinese, movements in the neck, head, and mouth might be related to the lexical tones of syllables [7]. In the context of deep-learning based VSR, recently [38] has shown that using video information from the cropped lip region together with pinyin (Chinese syllables) information yields 0.85% reduction in tone prediction error, supporting this hypothesis.

III. EXPLORING THE INFLUENCE OF ROIs ON VSR

In this section, we first introduce the deep VSR architectures used in this work, and four manually selected RoIs (some of which include extraoral regions) to be experimented on. Next, we introduce Cutout as a simple strategy to enhance face-based VSR. Finally, we introduce a few visualisation methods we use to diagnose the resulting models.

A. Model Architecture

**3D-ResNet18.** The 3D-ResNet18 backbone which holds the current state-of-the-art on the LRW dataset [27], [26] is used for all word-level experiments. It consists of a spatiotemporal convolution layer and a 18-layer 2D residual network which gradually reduces spatial dimensionality, and yields a 512-dimensional average-pooled vector for each frame. We use a 2-layer bidirectional Gated Recurrent Unit (Bi-GRU) with 1024 hidden units as the recurrent backend, and do not experiment with further regularisation such as Dropout and recurrent batch normalisation [26], as it deviates from the main objective of this paper.

**LipNet.** For sentence-level VSR, we use the LipNet architecture, which achieves state-of-the-art performance on the GRID corpus [4]. With only three spatiotemporal convolutional layers in the frontend, this lightweight model should be less prone to overfitting on such a small-scale dataset. We also use temporal augmentation and Dropout as recommended by the authors.

B. Description of Manually Selected RoIs

For fixed-size mouth crops, there are two dominant approaches: one uses fixed bounding box coordinates, and the other uses mouth-centered crops. To make this study comprehensive and self-complete, we experiment with both choices (although eventually we found no difference; see Table I). For face based models, we first crop and align the faces using detected eye and nose landmarks. This is because these parts undergo less non-rigid deformation, and this allows us to capture more significant motion in the lower face including the cheeks, the mouth and the jaw. At a low computational cost, this allows for more stable face tracks, and helps the network capture information in different spatial patches more consistently. However, it should be noted that face motion consists of rigid head movement (yaw, pitch, and roll) and non-rigid movement (e.g. frowning, raising eyebrows, and skin movement). Aligning the face retains yaw and pitch, but totally removes roll rotation. To this end, we also experiment with whole face and upper face cropped directly by fixed coordinates on LRW, since the faces are of uniform sizes and loosely aligned. The input sizes we choose are those commonly adopted in prior work, 112 × 112 and 100 × 100. We do not elaborate on the effects of spatial downsampling, since lower RoI resolution does not always yield poorer performance, as long as it remains above an acceptable level (e.g. 50 × 50 pixels) [5], [15], [35]. Besides, in any case datasets collected in-the-wild already contain inherent scale and image quality variations.

To crop the cheeks without revealing too much information in other regions, we crop a rectangular tile from the aligned face, which has no in-plane rotation. The vertical center of the RoI is set as the mean y coordinate of the 18th, 27th, 57th, and 59th landmark after transformation.

Finally, we do not experiment with the jaw and the lower neck. Although these parts are directly related to the vibrations of the vocal tract, these regions are not always visible, and such an RoI will be difficult to define.

C. Enhancing Face Based VSR with Cutout

Our motivation of using face as input is that it covers all other three sub-RoIs, which has the possibility of providing a strong “holistic” representation. If the network is able to remain robust against the speech-irrelevant variations that human faces present, including pose, lighting, makeup, etc., we can benefit from additional information in other facial regions all at once. However, a network has only limited capacity. On a fixed budget, the goals of achieving invariance to nuisance factors and keeping stronger discrimination capabilities are inherently complimentary, and the recognition performance may suffer if either is not sufficiently realised.

Indeed, our experiments will later show that the face models have already achieved performance comparable to lip based models. However, when we inspect the convolutional responses of vanilla face-based models (see Fig. 5) we find that the network tends to focus on the lip region and does not sufficiently utilise other parts of the face. Inspired by the observation that the vanilla face-based models able to achieve un-expected good performance, we try to enhance further the performance by asking the model to learn more discriminative features by utilising signals spread across the whole face. An intuitive idea is to create a strong patch-based ensemble, which has been shown feasible for facial expression recognition [19]. However, for VSR it would be
computationally expensive and impractical for deployment, since we are dealing with spatiotemporal data. Moreover, the redundancy between patches will burden optimisation.

Therefore, we propose to apply Cutout [14], a regularisation technique which has been popular with image CNNs. It augments the dataset with partially occluded versions of the samples in the dataset, and has already been successfully applied to image classification and object detection [39]. Our motivation is that this “adversarial erasing” process should help the model pay more attention to less significant motion signals related to speech in extraoral regions.

During training, a patch within the face region is randomly zeroed. Note that Cutout is applied to identical spatial positions across the video, so we expect the same facial region to be in roughly the same position after decent alignment. Over the course of the training process the model will encounter many masked versions of the same video, and eventually learn discriminative features for all parts of the face, which the model should be able to combine to its advantage during inference. Cutout fits seamlessly into the training process, and can be performed efficiently as a data augmentation step at no additional cost.

D. Interpreting Model Behaviour

To explore what our models are paying attention to throughout the video, we apply three visualisation techniques that can provide insights as to what the network focuses on for the task of visual speech recognition – similar to the use of gaze trackers in human psychology experiments.

Feature maps. Feature maps are filter responses to input images and outputs from previous layers, and can provide insight into intermediate representations learned by the model. Looking at responses from increasingly deeper layers can give us a sense of how the model combines low-level features into higher-level concepts that are useful for recognition.

Saliency maps. We use guided backpropagation saliency visualisation [25], which has been previously utilised to interpret lip reading [4], [9] and image classification models. Specifically, let \( x \in \mathbb{R}^{T \times H \times W \times C} \) be an input video, \( V \) be the alphabet or vocabulary, and \( \phi \) be the non-linear function that underlies the deep recognition network. For word-level classification, the network outputs a score for the \( i \)-th class in the vocabulary, \( p(i \mid x) = \phi(x)_i \), where \( 1 \leq i \leq |V| \). We compute its gradient with respect to the input \( x \) using guided backpropagation. Likewise, for sentence-level VSR, we compute the likelihood of the greedily decoded output \( y \in (V \cup \{\_\})^* \), which is \( \prod_i p(y_i \mid x) \), and derive it against the input video sequence \( x \) to obtain the saliency maps. Eventually we obtain a tensor the same size as the input, which depicts spatial regions the model bases its prediction on for different timesteps.

Spatiotemporal masking. This approach, adapted from the patch masking method for visualising the receptive fields of 2D ConvNets [36] has been used to visualise important space-time video patches in audio-visual speech enhancement models [16]. In our experiments, we mask each frame at the identical position for spatially aligned faces using a small 7 \( \times \) 7 patch in a sliding window fashion, and measure how overall accuracy is affected by computing performance drop, e.g. \( \Delta \text{accuracy} \). This process results in a heatmap depicting the contribution of different facial regions to recognition.

IV. EXPERIMENTS AND RESULTS

We train and evaluate on three VSR benchmarks, which cover tonal and atonal languages as well as in-the-wild and scripted speech: the Lip Reading in the Wild (LRW) dataset, the recent LRW-1000 dataset, and the GRID audiovisual corpus. In this section, we first briefly introduce the three datasets we use, and present some implementation details. Next, we compare recognition performance using the four manually selected RoIs in Sec. III-B. We highlight the benefits of incorporating extraoral regions, in particular by using aligned entire faces. Finally, we present results of our best performing model which combine Cutout with face inputs, and make a few useful remarks.

A. Datasets

LRW. LRW [9] is a challenging “in-the-wild” English word-level lip reading dataset derived from BBC news collections, with 500 classes and over 500,000 instances, of which 25,000 are reserved for testing.

LRW-1000. LRW-1000 [35] is a 1000-class, large-scale, naturally distributed Mandarin Chinese word-level lip reading dataset, also derived from TV broadcasts. With over 700,000 word instances, it is even more challenging, with significant pose, scale, background clutter, word length, and inter-speaker variations. Note that the whole face was not provided with LRW-1000 in the initial release. We will release face tracks for LRW-1000\(^1\) along with annotations that have undergone another round of manual cleaning as well as corresponding baselines in due course.

GRID. The GRID audiovisual corpus [11], released in 2006, is a popular benchmark for sentence-level VSR. It consists of video recordings from 34 speakers\(^2\), yielding 33,000 utterances. All sentences follow a fixed grammar.

B. Implementation Details

Data preprocessing. We detect faces and facial landmarks with the open-source SeetaFace2 toolkit [23], and align faces with similarity transformations using upper face landmarks [8], which are smoothed using a temporal Gaussian kernel of width 3. For LRW and LRW-1000, the faces are resized to 122 \( \times \) 122 and randomly cropped to 112 \( \times \) 112 during training. For GRID, the faces are resized to 100 \( \times \) 100.

Training details. All models are implemented with PyTorch and trained on 2 to 3 NVIDIA Titan X GPUs, each with 12GB memory. We use the Adam optimizer with default hyperparameters. For word-level VSR, we use three-stage training described in [27]. Weight decay is set to 0.0001, and

\(^1\)For now, we train and evaluate our models with the original word alignment and landmark annotations in [35] for fair comparison. Note that the original paper used ResNet-34, while this work uses ResNet-18 which is currently more popular due to fewer parameters and better performance.

\(^2\)One speaker’s data is unavailable due to technical reasons.
and train/validation/test partitioning provided with LRW are classification accuracy and Word Error Rate (WER), training, the same as previous state-of-the-art [4], [34]. The from each speaker for evaluation, and the remainder for sentence-level VSR, we use a fixed learning rate of 0.0001. Flipping, but no random cropping is used because of relatively accurate and stable landmarking results. We use the Data partitioning, and evaluation metrics. We use the train / validation / test partitioning with LRW and LRW-1000. For GRID, we use 255 random sentences from each speaker for evaluation, and the remainder for training, the same as previous state-of-the-art [4], [34]. The evaluation metrics for word-level and sentence-level tasks are classification accuracy and Word Error Rate (WER), respectively, where WER is defined as

$$\text{WER} = \frac{\text{# of substitutions, deletions, and insertions}}{\text{length of ground truth transcript}}$$

Hence lower WERs are preferred. Note that the two metrics are equivalent if one views word recognition as a one-word sentence recognition task, in the sense that $\text{ACC} = 1 - \text{WER}$.

C. Experiments on Manually Selected RoIs

Baseline results on the manually defined RoIs are shown in Table I, Table II, and III. We analyze the results step by step for the rest of this subsection.

Effectiveness of including extraoral regions. Experiment results clearly show that upper face and cheeks convey useful information, since recognition rates are far above chance. Counterintuitively, the upper face achieves nearly half the accuracy of face and mouth models. To ensure that the model has learned useful discriminative features instead of some unknown inherent bias within the dataset, we conduct an additional experiment, transferring from LRW to LRW-1000 while keeping the front-end fixed. Intuitively, if the front-end has learned spurious clues that are irrelevant to the task itself, it should behave poorly on a dataset it has not been trained on. However, despite significant differences between the two datasets in terms of quality and language, classification accuracy is still far above chance after we fixed the front-end, with only 2.26% absolute performance loss.

Seeing how the upper face and cheeks convey useful information for VSR, by feeding the model with the entire face, we would expect it to benefit from the additional spatial context, which is indeed the case. Using the entire face instead of only the mouth region yields 1.6% WER reduction on GRID, 3.07% improvement on LRW-1000, and 0.16% improvement on LRW (when directly cropped faces are used). The slight performance regression on LRW with aligned faces will be discussed next.

Making a case for face alignment. For LRW, there are two sets of face-based experiments, one with face alignment, and the other using faces directly cropped by fixed coordinates. We observe a small performance degradation on LRW when we align faces to our canonical template (0.2% to 0.3% drop relative to mouth and 7/8 original resolution direct crops). Since recognition performance using the upper face actually benefits from alignment, it can be argued that the performance drop is most likely due to slightly lower mouth resolution (about $70 \times 70$), and not removal of roll during the alignment process. This is not desirable, but acceptable, and there may be room for improvement if we adopt higher resolution inputs and improve landmarking quality. Therefore, we consider aligning the faces into stable face tracks to be beneficial, and use aligned faces for the remaining experiments. By any means, this is also necessary for structured facial region erasing with Cutout.

Failure modes. As an illustrative example, we further compare confusions made by the model under each crop setting in Table IV and Table V for the two English datasets. Overall, short words with less context and words that invoke weak extraoral motions perform worst. We observe that the words best predicted by the upper face are long words, extraoral motion. The face based model, which achieves comparable but slightly inferior performance compared to the mouth based model, fails to discriminate words with only subtle differences, such as “benefits” and “benefit”.

### Table I

**Evaluation on the LRW dataset with different RoI choices.** The second group uses directly cropped (upper) faces while the third applies face alignment.

| Region    | Resolution | Accuracy | Description                          |
|-----------|------------|----------|--------------------------------------|
| Mouth     | $88 \times 88$ | 83.30%   | Fixed bounding box [24], [26]        |
| Face      | $112 \times 112$ | 83.46%   | Nose-centered, 7/8 original size     |
| Upper face| $112 \times 112$ | 42.58%   | Resized upper half from above        |
| Face      | $112 \times 112$ | 83.19%   | Aligned with eye & nose landmarks [8]|
| Cheeks    | $112 \times 112$ | 62.49%   | Resized $40 \times 112$ crop from above |
| Upper face| $112 \times 112$ | 48.33%   | Resized upper half                   |
| Face (CBAM)| $112 \times 112$ | 83.14%   |                                    |
| Face (Cutout)| $112 \times 112$ | 85.02%   |                                    |

### Table II

**Evaluation on LRW-1000 with different RoI choices.**

| Region    | Resolution | Accuracy | Description                          |
|-----------|------------|----------|--------------------------------------|
| Mouth     | $88 \times 88$ | 38.64%   | Mouth-centered, no roll (as in [35]) |
| Face      | $112 \times 112$ | 41.71%   | Aligned with eye & nose landmarks     |
| Upper face| $112 \times 112$ | 15.84%   | Resized upper half                    |
| Cheeks    | $112 \times 112$ | 32.50%   | Resized $40 \times 112$ crop from above |
| Face (Cutout)| $112 \times 112$ | 45.24%   |                                    |
| Upper face| $112 \times 112$ | 13.58%   | Front-end loaded from LRW and fixed   |

### Table III

**Evaluation on the GRID corpus with different RoI choices.** A 5-gram, character-level language model is used during beam search. “CER” stands for Character Error Rate. The lower the error rates, the better.

| Region    | Resolution | WER | CER | Description                          |
|-----------|------------|-----|-----|--------------------------------------|
| Mouth     | $100 \times 50$ | 4.85% | 1.9% | Affine warped                        |
| Mouth     | $100 \times 50$ | 4.7% | 1.9% | Above (reproduced)                   |
| Face      | $100 \times 100$ | 3.1% | 1.3% | Aligned with eye & nose landmarks    |
| Upper face| $100 \times 50$ | 14.4% | 7.4% | Upper half of above                  |
| Cheeks    | $100 \times 50$ | 6.8% | 3.1% | Resized $36 \times 100$ crop from above |
| Face (Cutout)| $100 \times 100$ | 2.9% | 1.2% |                                    |
Recogising such words correctly requires analysis of tongue movement, which is hindered by the lowered overall resolution. However, on the GRID corpus, the face-based model seems to have identified idiosyncratic speaking styles, allowing it to make correct predictions for even short words and letters that are not easy to distinguish, and eventually obtain results better than cropped mouths.

**Superiority of using entire faces for cross-dataset transfer.** A byproduct of using the entire face rather than specifying sub-ROIs is that the visual front-end can be transferred across datasets easily with no explicit sub-ROI cropping after face detection and registration, which is already very routine in face recognition. Empirically, we found that transferring across face crops is much better than transferring across mouth crops from LRW to LRW-1000. When both fine-tuned for 9 epochs, mouth-based models suffer from source and target domain discrepancies, resulting in fluctuating validation loss and accuracy (best value 2.8482/38.40%), whereas the face-based model transferred stably, and converged to a better result (best value 2.7202/40.35%).

### D. Experiments on Enhancing Face Inputs with Cutout

Results of combining Cutout with aligned faces can also be found in previous tables. This strategy is extremely powerful, enabling state-of-the-art performance on all three benchmarks. In particular, Cutout significantly reduces overfitting on LRW to the point where the model starts to underfit, as can be seen from the training curves in Fig. 3. This is radically different from models trained on faces or mouths without Cutout, where the network eventually reaches near 100% accuracy on the training data, and proves that this harsh regularisation policy is indeed useful for VSR. Below we elaborate on a few interesting observations.

**Effect of the Cutout patch size.** The size of the masked out region is an important hyperparameter for Cutout. We experiment with four different sizes: 70 × 70, 56 × 56, 42 × 42, and 30 × 30 patches for each of the dataset. The derived maps are colored by intensity and overlaid onto the original frames. Each face thumbnail corresponds to a 224 × 224 feature map (28, 28, 1). We also identify regions that are important to the network’s predictions by spatiotemporal masking on the entire test set of LRW, and results are depicted in Fig. 4. It can be seen that for both models, the area that leads to the most significant performance degradation when masked is still the lip region, which agrees with common intuition. Further, the model trained with Cutout is also affected by occlusion in extraoral regions such as the cheeks and the upper face, showing that the model has learned strong visual features that encode these weak signals to complement lip motion. More importantly, the plain model suffers up to 40% performance drop even when only a 7 × 7 patch is.

**Visualising key facial regions.** Fig. 5 provides some saliency visualisations which show that models with Cutout can learn more discriminative features. Fig. 5(a) is generated by extracting feature maps from the third ResNet layer, and performing max-pooling across channels, and Fig. 5(b)(c) (GRID and LRW-1000) by computing back-prop saliency. The derived maps are colored by intensity and overlaid onto the original frames. Each face thumbnail corresponds to a time step, and the two rows can be compared side-by-side to see the effects of introducing Cutout, especially regions highlighted with a dotted red box. For example, for LRW the convolutional responses are no longer confined to the lips, and for GRID there is stronger saliency in the cheeks, the eyebrows, and other facial muscles. The third row shows that after transferring from LRW to LRW-1000, saliency in extraoral regions persist, which means that the learned facial features generalize well and are relatively robust.

### Table IV

| Target / Estimated | Smooth | Error | Network |
|--------------------|--------|-------|---------|
| Mouth - train      | 0      | 0     | 0       |
| Mouth - val        | 0      | 0     | 0       |
| Face - train       | 0      | 0     | 0       |
| Face - val         | 0      | 0     | 0       |

### Table V

| Examples of predictions on GRID under each crop setting. Prediction errors are highlighted in red. GT: Ground Truth; UF: Upper Face; C: Cheeks; M: Mouth; F: (Aligned) Face. |
|----------------------------------------------------------------------------------------------------------------------------------|
| GT | lay white in u four now | GT | lay white in q five again |
| UF | lay white in u four now | UF | set white at t five again |
| C  | lay white at o four now | C  | lay white at q five again |
| M  | lay white at o four now | M  | lay white at q five again |
| F  | lay white in u four now | F  | lay white in q five again |

**Fig. 3.** Accuracy curves on LRW as training progresses (only the first 75k steps of the temporal convolution backend stage is shown). Compared to vanilla mouth and face based models which easily achieve nearly 100% accuracy on the training set, Cutout (denoted by red curves) significantly reduces overfitting.

**Fig. 4.** Ablation results on LRW with different Cutout patch sizes. We achieve best validation accuracy with patches half the size of the input. 42, and 28 × 28, which are 5/8, 1/2, and 3/8 of the scale of the whole face, respectively. Experiments on LRW show that among those a 56 × 56 patch is most effective (see Fig. 4). This is probably because it is approximately the average size of the mouth, and the possibility of the entire mouth being blocked allows for more efficient utilisation of extraoral regions. Since we adopt the same canonical template, we use 1/2-size masks for all datasets (i.e. 50 × 50 for GRID).
occluded, the drop remains below 2% for the model trained with Cutout. This observation again strongly supports the usefulness of extraoral regions.

Comparison with attention-based RoI selection. The attention mechanism, inspired by human vision, has been widely used for implicit RoI selection and fine-grained image recognition. Here we compare Cutout with a Convolutional Block Attention Module (CBAM) [33] augmented baseline on LRW, where we plug CBAM into ResNet blocks. Here, we use $5 \times 5$ instead of $7 \times 7$ kernels for the spatial attention modules to accommodate to the smaller feature maps. Although CBAM is a powerful spatial-channel attention mechanism which achieved remarkable performance on ImageNet classification and robust remote heart rate estimation [20], results show that the attention-augmented model is only marginally better than the baseline on LRW. We believe this is because the subtle movements in extraoral regions are too weak to be captured with the attention mechanism, and the model is still biased towards lip motion.

Performance across pose. We are also interested in how well the Cutout-augmented model can handle pose variations. We turn to LRW-1000, where the data is divided into different difficulty levels according to yaw rotation. From Table VI, we can see that the model trained with Cutout outperforms the no-augmentation face-based baseline by about 2% on both the easy (yaw $\geq 20^\circ$) and the medium ($\geq 40^\circ$) subset, but degrades slightly on the hard subset ($\geq 60^\circ$). We believe this is mainly because effective areas are smaller in the hard setting, where there are more frames in profile view. The effective regions are more likely to be erased or occluded when there is significant yaw rotation.

V. CONCLUSION

In this paper, we have investigated a previously overlooked problem in VSR: the use of extraoral information. We demonstrate that extraoral regions in the face, such as the upper face and the cheeks, can also be included to boost performance. We show that using simple Cutout augmentation with aligned face inputs can yield stronger features, and vastly improve recognition performance by forcing the model to learn the less obvious extraoral cues from data. Beyond VSR, our findings also have clear implications for other speech-related vision tasks, such as realistic talking face generation, face spoofing detection and audio-visual speech enhancement. Next steps include extending the method to sentence-level VSR where more contextual clues are available, increasing input resolution, and eliminating the need for explicit face alignment.

VI. ACKNOWLEDGMENTS

We would like to thank Chenhao Wang and Mingshuang Luo’s extensive help with data processing. This work is supported in part by the National Key R&D Program of China (No. 2017YFA0700804), and Natural Science Foundation of China (No. 61702486, 61876171).
