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

Yuanhang Zhang[1,2], Shuang Yang[1], Jingyun Xiao[1,2], Shiguang Shan[1,2], Xilin Chen[1,2]

[1] Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing 100190, China
[2] University of Chinese Academy of Sciences, Beijing 100049, China

*Abstract*— **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 on the evaluation of 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
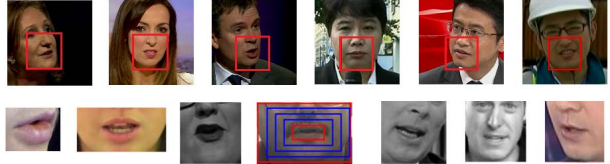
Fig. 1. **Common practices for RoI selection in VSR.** To date, there is no clear consensus on a best practice, resulting in very different RoIs in different works (see Sec. II-A). Top row: examples of frames from talking face videos. Bottom row: some examples of cropped RoIs in prior work, from left to right: [3], [4], [9], [17], [27], [1], [35].

in practical use, due to the diversity of the speaker's pose and other variations in the facial region that are unrelated to speech production. Accordingly, as part of the dataset creation pipeline, almost all researchers crop regions-of-interest (RoIs) around the mouth after obtaining face bounding boxes and landmarks. By observing only the cropped RoI, the model is expected to focus on fine-grained discrimination of "clean" motion signals within the most relevant lip region, and not be distracted by other parts of the face, whose utilities are less obvious. Some common practices for RoI selection in previous work are depicted in Fig. 1.

However, this convention of explicit mouth RoI cropping inevitably brings about many questions. Firstly, lip motion is not the only visual signal we can rely on to decode speech. Research on the advantage of performing speechreading with the whole face has a long history [28], [6]. In particular, movements of articulatory muscles such as the orbicularis oris (which is near the lips) lead to skin movements, often reflected by the cheeks being pushed and pulled, as well as changes in the visibility of the nasolabial folds. The now widely adopted term "speechreading", used in place of "lip reading" implies precisely the contribution of extraoral facial regions, such as the tongue, teeth, and cheeks, to the speech perception task. Evidence from psychology studies, and our experience in human communication also suggest that we in fact do not focus on the speaker's lip all the time throughout a conversation, even in very noisy environments. Instead, we scan different regions of the person's face periodically [31].

Secondly, if we do use a mouth RoI, there are many factors that need to be considered: how much of the face should the RoI cover? Will increased spatial context improve performance by supplying more information, or hinder performance by distracting the network with information unrelated to speech? Should we apply additional corrections

to handle pose variation? Answers to these questions are not evident or straight, 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 retraced 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 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



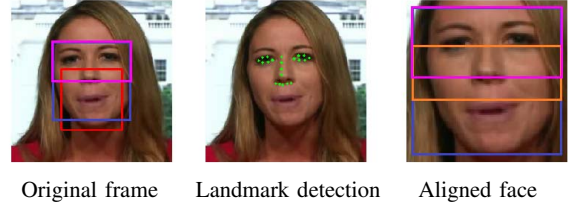| Original frame | Landmark detection | Aligned face |

Fig. 2. **Illustration of the sub-face RoIs defined in this paper.** We train baselines on the whole face (blue), the upper face (purple), the cheeks (orange), and the mouth (red).

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, $112 \times 112$ and $100 \times 100$. 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. $50 \times 50$ 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, since 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 $\mathbf{x} \in \mathbb{R}^{T \times H \times W \times C}$ 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, $p(i \mid \mathbf{x}) = \phi(\mathbf{x})_i$, where $1 \leq i \leq |V|$. We compute its gradient with respect to the input $\mathbf{x}$ using guided backpropagation. Likewise, for sentence-level VSR, we compute the likelihood of the greedily decoded output $\mathbf{y} \in (V \cup \{\_\})^*$, which is $\prod_t p(\mathbf{y}_t \mid \mathbf{x})$, and derive it against the input video sequence $\mathbf{x}$ 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.

| Region | Resolution | Accuracy | Description |
|---|---|---|---|
| Mouth | $88 \times 88$ | 83.30% | Fixed bounding box [21], [26] |
| Mouth | $88 \times 88$ | 83.30% | Mouth-centered [9] |
| Face | $112 \times 112$ | 83.46% | Nose-centered, 7/8 original size |
| Upper face | $112 \times 112$ | 42.28% | Resized upper half from above |
| Face | $112 \times 112$ | 83.10% | 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%** | |

| 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 |

| Region | Resolution | WER ↓ | CER ↓ | Description |
|---|---|---|---|---|
| Mouth | $100 \times 50$ | 4.8%[4] | 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%** | |

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, such as "Westminster" and "Temperatures", and cheeks are good at making predictions for words that invoke significant 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".

learning rate is initialised to 0.0003 for stage I / II and 0.001 for stage III, decreasing on log scale from the 5th or the 10th epoch onwards, depending on whether Cutout is applied. We apply identical spatial augmentation to every frame in the form of random cropping and random horizontal flipping during training, and take a central crop for testing. Also, for *LRW*-1000 training, we initialise front-end weights from the corresponding best model on LRW, following [35]. For sentence-level VSR, we use a fixed learning rate of 0.0001 and use no weight decay. We also apply random horizontal flipping, but no random cropping is used because of relatively accurate and stable landmarking results.

**Data partitioning, and evaluation metrics.** We use the train / validation / test partitioning provided 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 carry 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

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.

TABLE IV

TOP PREDICTED WORDS, WORST PREDICTED WORDS AND PAIRS EXHIBITING HIGHEST CONFUSION IN LRW UNDER EACH CROP SETTING.

| Accuracy / Network | | | | Accuracy / Network | | | |
|---|---|---|---|---|---|---|---|
| Mouth | Face | Cheeks | Upper Face | Mouth | Face | Cheeks | Upper Face |
| AGREEMENT (1) | ACCUSED (1) | AFTERNOON (0.98) | WESTMINSTER (0.96) | ASKED (0.58) | WORLD (0.6) | REALLY (0.32) | GREAT (0.18) |
| ALLEGATIONS (1) | AGREEMENT (1) | WEEKEND (0.98) | TEMPERATURES (0.92) | BRITAIN (0.58) | ANSWER (0.58) | COULD (0.3) | OTHER (0.18) |
| BEFORE (1) | BEFORE (1) | WELFARE (0.98) | AFTERNOON (0.88) | MATTER (0.58) | BECAUSE (0.58) | GETTING (0.3) | UNTIL (0.18) |
| PERHAPS (1) | CAMPAIGN (1) | WESTMINSTER (0.98) | SUNSHINE (0.88) | SPEND (0.58) | COURT (0.58) | MAKES (0.3) | WHICH (0.18) |
| PRIME (1) | FOLLOWING (1) | INFORMATION (0.96) | DESCRIBED (0.86) | TAKEN (0.58) | PERSON (0.58) | MATTER (0.3) | BRING (0.16) |

| Target / Estimated (# Errors) / Network | | | |
|---|---|---|---|
| Mouth | Face | Cheeks | Upper Face |
| SPEND / SPENT (11) | BENEFITS / BENEFIT (14) | CLAIMS / GAMES (11) | MILLION / BILLION (11) |
| PRESS / PRICE (11) | PRESS / PRICE (12) | CHALLENGE / CHANGE (10) | BENEFITS / BENEFIT (10) |
| WORST / WORDS (10) | LIVING / GIVING (12) | SYRIAN / SYRIA (10) | EVERYONE / EVERYBODY (10) |
| PRICE / PRESS (10) | WORST / WORDS (11) | GROUND / AROUND (9) | TAKEN / SECOND (9) |
| SERIES / SERIOUS (9) | THEIR / THERE (10) | INDUSTRY / HISTORY (9) | TERMS / TIMES (9) |

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 |

Recognising 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 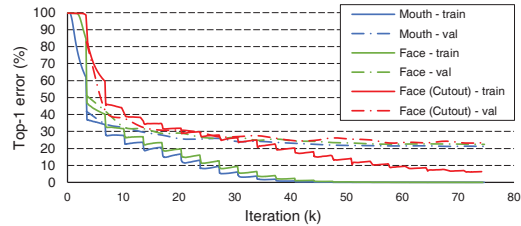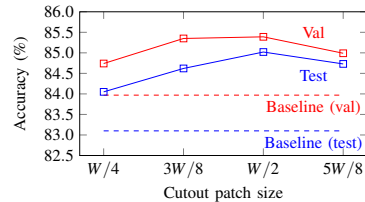different sizes: $70 \times 70$, $56 \times 56$, $42 \times$ 42, and $28 \times 28$, which are $5/8$, $1/2$, $3/8$, and $1/4$ the scale of the whole face, respectively. Experiments on LRW show that among those a $56 \times 56$ patch is most effective (see Fig. 4). This is probably because it is approximately 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. $50 \times 50$ for GRID).

**Visualising key facial regions.** Fig. 5 provide 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.

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. 6. 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. In addition, 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, while the plain model suffers up to 40% performance drop even when only a $7 \times 7$ patch is

Fig. 5. **Visualisations of model behaviour on LRW, *LRW*-1000, and GRID with Cutout applied.** In particular, note how models trained with Cutout preserve fine-grained features, and yields clearly visible saliency around the cheeks. (a) ResNet layer 3 feature maps for "hundred"; (b) Saliency maps for "bin blue"; (c) Saliency maps for "*er ling yi qi*" (twenty-seventeen).



Fig. 6. **Important facial regions determined by spatiotemporal masking.** Left: without Cutout. Middle: an aligned face for reference. Right: with Cutout. Regions that result in more accuracy drop when occluded are colored brighter, and more crucial for model prediction.

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 $\geq 20°$) and the medium ($\geq 40°$) subset, but degrades slightly on the hard subset

### TABLE VI
PERFORMANCE W.R.T POSE ON *LRW*-1000

| Methods | Easy | Medium | Hard | All |
|---|---|---|---|---|
| Mouth (ResNet34) [35] | 24.89% | 20.76% | 15.9% | 38.19% |
| Mouth | 25.95% | 21.55% | 18.36% | 38.64% |
| Face | 28.87% | 28.45% | **27.21**% | 41.71% |
| **Face (Cutout)** | **31.74**% | **30.04**% | 26.89% | **45.24**% |

($\geq 60°$). 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

## REFERENCES

[1] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman. Deep audio-visual speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2018.

[2] N. Alghamdi, S. Maddock, R. Marxer, J. Barker, and G. J. Brown. A corpus of audio-visual lombard speech with frontal and profile views. *The Journal of the Acoustical Society of America*, 143(6):EL523–EL529, 2018.

[3] I. Anina, Z. Zhou, G. Zhao, and M. Pietikäinen. OuluVS2: A multi-view audiovisual database for non-rigid mouth motion analysis. In *11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2015, Ljubljana, Slovenia, May 4-8, 2015*, pages 1–5, 2015.

[4] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas. LipNet: End-to-end sentence-level lipreading. *CoRR*, abs/1611.01599, 2016.

[5] H. L. Bear, R. W. Harvey, B. Theobald, and Y. Lan. Resolution limits on visual speech recognition. In *2014 IEEE International Conference on Image Processing, ICIP 2014, Paris, France, October 27-30, 2014*, pages 1371–1375, 2014.

[6] C. Benoit, T. Guiard-Marigny, B. Le Goff, and A. Adjoudani. Which components of the face do humans and machines best speechread? In *Speechreading by humans and machines*, pages 315–328. Springer, 1996.

[7] T. H. Chen and D. W. Massaro. Seeing pitch: Visual information for lexical tones of mandarin-chinese. *The Journal of the Acoustical Society of America*, 123(4):2356–2366, 2008.

[8] J. S. Chung, A. Jamaludin, and A. Zisserman. You said that? In *British Machine Vision Conference 2017, BMVC 2017, London, UK, September 4-7, 2017*, 2017.

[9] J. S. Chung and A. Zisserman. Lip reading in the wild. In *Computer Vision - ACCV 2016 - 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II*, pages 87–103, 2016.

[10] J. S. Chung and A. Zisserman. Lip reading in profile. In *British Machine Vision Conference 2017, BMVC 2017, London, UK, September 4-7, 2017*, 2017.

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

[12] E. Cvejic, J. Kim, and C. Davis. Prosody off the top of the head: Prosodic contrasts can be discriminated by head motion. *Speech Communication*, 52(6):555–564, 2010.

[13] C. Davis and J. Kim. Audio-visual speech perception off the top of the head. *Cognition*, 100(3):B21–B31, 2006.

[14] T. Devries and G. W. Taylor. Improved regularization of convolutional neural networks with cutout. *CoRR*, abs/1708.04552, 2017.

[15] L. Dungan, A. Karaali, and N. Harte. The impact of reduced video quality on visual speech recognition. In *2018 IEEE International Conference on Image Processing, ICIP 2018, Athens, Greece, October 7-10, 2018*, pages 2560–2564, 2018.

[16] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein. Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation. *ACM Trans. Graph.*, 37(4):112:1–112:11, 2018.

[17] A. Koumparoulis, G. Potamianos, Y. Mroueh, and S. J. Rennie. Exploring ROI size in deep learning based lipreading. In *Auditory-Visual Speech Processing, AVSP 2017, Stockholm, Sweden, 25-26 August 2017.*, pages 64–69, 2017.

[18] C. R. Lansing and G. W. McConkie. Word identification and eye fixation locations in visual and visual-plus-auditory presentations of spoken sentences. *Perception & psychophysics*, 65(4):536–552, 2003.

[19] Y. Li, J. Zeng, S. Shan, and X. Chen. Patch-gated CNN for occlusion-aware facial expression recognition. In *24th International Conference on Pattern Recognition, ICPR 2018, Beijing, China, August 20-24, 2018*, pages 2209–2214, 2018.

[20] X. Niu, X. Zhao, H. Han, A. Das, A. Dantcheva, S. Shan, and X. Chen. Robust remote heart rate estimation from face utilizing spatial-temporal attention. In *14th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2019, Lille, France, May 14-18, 2019*, pages 1–8, 2019.

[21] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic. End-to-end audiovisual speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018*, pages 6548–6552, 2018.

[22] T. N. Sainath, O. Vinyals, A. W. Senior, and H. Sak. Convolutional, long short-term memory, fully connected deep neural networks. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015*, pages 4580–4584, 2015.

[23] SeetaTech. SeetaFace2. https://github.com/seetafaceengine/SeetaFace2, 2019.

[24] J. Shiraishi and T. Saitoh. Optical flow based lip reading using non rectangular ROI and head motion reduction. In *11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2015, Ljubljana, Slovenia, May 4-8, 2015*, pages 1–6, 2015.

[25] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. A. Riedmiller. Striving for simplicity: The all convolutional net. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, 2015.

[26] T. Stafylakis, M. H. Khan, and G. Tzimiropoulos. Pushing the boundaries of audiovisual word recognition using residual networks and LSTMs. *Computer Vision and Image Understanding*, 176-177:22–32, 2018.

[27] T. Stafylakis and G. Tzimiropoulos. Combining residual networks with LSTMs for lipreading. In *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, pages 3652–3656, 2017.

[28] W. H. Sumby and I. Pollack. Visual contribution to speech intelligibility in noise. *The journal of the acoustical society of america*, 26(2):212–215, 1954.

[29] K. Sun, C. Yu, W. Shi, L. Liu, and Y. Shi. Lip-Interact: Improving mobile device interaction with silent speech commands. In *The 31st Annual ACM Symposium on User Interface Software and Technology, UIST 2018, Berlin, Germany, October 14-17, 2018*, pages 581–593, 2018.

[30] E. Vatikiotis-Bateson, I.-M. Eigsti, S. Yano, and K. G. Munhall. Eye movement of perceivers during audiovisualspeech perception. *Perception & psychophysics*, 60(6):926–940, 1998.

[31] M. L.-H. Võ, T. J. Smith, P. K. Mital, and J. M. Henderson. Do the eyes really have it? dynamic allocation of attention when viewing moving faces. *Journal of vision*, 12(13):3–3, 2012.

[32] Y. Wang, Y. Shen, Z. Liu, P. P. Liang, A. Zadeh, and L. Morency. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019.*, pages 7216–7223, 2019.

[33] S. Woo, J. Park, J. Lee, and I. S. Kweon. CBAM: convolutional block attention module. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII*, pages 3–19, 2018.

[34] K. Xu, D. Li, N. Cassimatis, and X. Wang. LCANet: End-to-end lipreading with cascaded attention-ctc. In *13th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2018, Xi'an, China, May 15-19, 2018*, pages 548–555. IEEE Computer Society, 2018.

[35] S. Yang, Y. Zhang, D. Feng, M. Yang, C. Wang, J. Xiao, K. Long, S. Shan, and X. Chen. 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, Lille, France, May 14-18, 2019*, pages 1–8, 2019.

[36] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, pages 818–833, 2014.

[37] G. Zhao, M. Barnard, and M. Pietikäinen. Lipreading with local spatiotemporal descriptors. *IEEE Trans. Multimedia*, 11(7):1254–1265, 2009.

[38] Y. Zhao, R. Xu, and M. Song. A cascade sequence-to-sequence model for chinese mandarin lip reading. In C. Xu, M. S. Kankanhalli, K. Aizawa, S. Jiang, R. Zimmermann, and W. Cheng, editors, *MMAsia '19: ACM Multimedia Asia, Beijing, China, December 16-18, 2019*, pages 32:1–32:6. ACM, 2019.

[39] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang. Random erasing data augmentation. *CoRR*, abs/1708.04896, 2017.