EFFICIENT MOUTH ALIGNMENT FOR VISUAL SPEECH RECOGNITION

Zhe Niu and Brian Mak

Department of Computer Science and Engineering
The Hong Kong University of Science and Technology
{zniu,bmak}@cse.ust.hk

ABSTRACT

Visual Speech Recognition (VSR), also known as lip-reading, requires accurate mouth tracking and alignment to achieve high performance. Conventional approaches rely on complex pipelines involving facial landmark detection, facial landmark smoothing, affine transformation estimation and warping. In this paper, we propose a streamlined solution: a fast Mouth Alignment Network (MAN) that directly estimates affine transformation parameters from a video of a talking face, significantly simplifying the process. Our approach also incorporates a complementary Mouth Scoring Network (MSN) that injects knowledge about adversarial perturbations into the training process, further improving alignment quality. Experiments on standard lip-reading benchmarks (LRS2 and LRS3) demonstrate that our method maintains the performance when integrated with state-of-the-art VSR systems, including Auto-AVSR and USR, while substantially improving the alignment quality and computational efficiency.

***Index Terms*—** lip-reading, visual speech recognition, mouth alignment.

1. INTRODUCTION

Visual Speech Recognition (VSR) is a challenging task aimed at recognizing spoken words from visual lip movements. Since the mouth region-of-interest (ROI) contains most of the information needed for VSR and processing entire high-resolution frames is computationally inefficient, VSR models typically focus on small (e.g., 88×88) aligned mouth ROIs. State-of-the-art VSR models, such as Auto-AVSR [1], USR [2], Llama-AVSR [3] and VSP-LLM [4], rely on a complex pipeline to obtain the mouth region, involving extracting facial landmarks per video frame, smoothing these landmarks, and estimating affine transformation parameters to warp the original frames into aligned mouth frames.

In this work, we aim to improve the efficiency and robustness of mouth alignment pipeline. We propose a fast Mouth Alignment Network (MAN) that directly predicts affine transformation parameters from raw video frames. This approach eliminates the conventional multi-step pipeline and avoids reliance on facial landmark detection. MAN is initially trained by distilling affine parameters estimated by the conventional approach. To further refine the training process, we introduce a Mouth Scoring Network (MSN) that injects knowledge of adversarial perturbations into the training process. MSN produces a differentiable score for an aligned mouth video, which is used to guide MAN in generating high-quality mouth alignment. MSN is trained in a self-supervised manner with contrastive loss between a preferred sample (*i.e.*, one estimated by the conventional pipeline) and a randomly perturbed non-preferred sample, at the same time leveraging the knowledge from pretrained lip encoder of the VSR model. This approach allows MAN to learn from negative examples, improving its ability to produce high-quality alignment parameters.

We demonstrate the effectiveness of our proposed method by conducting experiments on the LRS2 and LRS3 datasets. Our results show that the proposed MAN achieves a performance comparable to that of the conventional pipeline when integrated with state-of-the-art VSR models: Auto-AVSR [1] and USR [2], while improving efficiency and alignment quality estimated by MSN. Our contributions are: 1. A novel mouth alignment model that directly predicts affine transformation parameters, eliminating facial landmark detection and smoothing. 2. A fast mouth alignment network (MAN) guided by a pretrained mouth scoring network (MSN) that uses adversarial perturbations to improve alignment quality and detect poor alignment. 3. Extensive experiments on LRS2/LRS3 with Auto-AVSR and USR systems, showing comparable performance to conventional pipelines with improved efficiency.

2. RELATED WORK

2.1. Conventional mouth alignment methods

State-of-the-art VSR methods, including Auto-AVSR [1], USR [2], Llama-AVSR [3] and VSP-LLM [4], follow a multi-step approach for lip region alignment. These methods first detect facial landmarks using the FAN detector [5], then apply temporal smoothing with an averaging window, and finally compute affine transformations to align the lip region.

979-8-3503-2411-2/25/$31.00 ©2025 IEEE

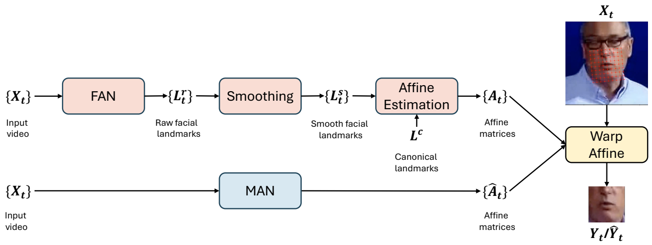


Fig. 1. Conventional (top) and proposed (bottom) mouth alignment pipelines for VSR. Both predict affine transformations to warp input frames into mouth frames.

As in fig. 1, this conventional alignment pipeline consists of several sequential stages: face detection, landmark detection, landmark smoothing, and affine transformation. This cascade approach relies on manually designed components that lack the robustness of end-to-end learned solutions. Moreover, the Hour-Glass layer [6] in the FAN [5] landmark extractor was specifically designed for precise landmark detection, which is computationally inefficient for the mouth alignment where only a few affine parameters are required to be predicted.

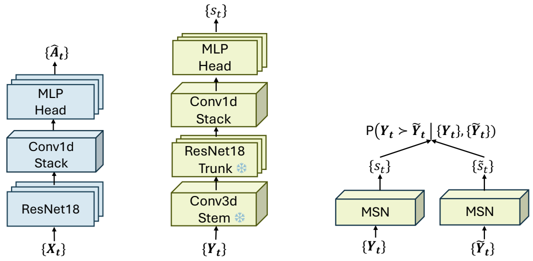
2.2. Spatial transformer networks

Our proposed mouth alignment network builds upon Spatial Transformer Networks (STNs) [7], which directly predict transformation parameters from input images and apply these transformations to the input. STNs are designed to achieve spatial invariance by learning optimal transformation parameters. This makes them particularly suited for tasks such as mouth alignment in VSR. By leveraging STNs, our approach benefits from end-to-end optimization, eliminating the need for landmark detection in conventional pipelines. This results in a more robust and efficient solution for mouth alignment in visual speech recognition.

3. METHODOLOGY

3.1. Preliminary

State-of-the-art visual speech recognition (VSR) models [1, 8, 2] typically rely on a mouth alignment pipeline that estimates affine transformations based on facial landmarks. As illustrated in fig. 1, given an input video with T RGB frames $\{\boldsymbol{X}_t\}_{t=1}^T \in \mathbb{R}^{T\times H\times W\times 3}$, the conventional pipeline first employs a face alignment network (FAN) [5] to detect 68 2D facial landmarks for each frame independently: $\boldsymbol{L}_t^r = \text{FAN}(\boldsymbol{X}_t), \boldsymbol{L}_t^r \in \mathbb{R}^{68\times 2}$. These raw landmarks are then temporally smoothed using a sliding window, producing $\boldsymbol{L}_t^s \in \mathbb{R}^{68\times 2}$. Subsequently, the smoothed landmarks are used to compute an affine transformation that aligns the face with a canonical face template $\boldsymbol{L}_t^c \in \mathbb{R}^{68\times 2}$. This estimation is typically performed using the Least Median of Squares (LMedS) method [9], producing a sequence of 3×3 affine transformation matrices $\{\boldsymbol{A}_t\} \in \mathbb{R}^{T\times 3\times 3}$. Once the affine matrices are obtained, each mouth frame $\boldsymbol{Y}_t$ can be extracted by warping the affine matrices:



(a) MAN. (b) MSN. (c) Pref. model.

Fig. 2. Mouth alignment, scoring networks, and preference modeling. Snowflakes indicate frozen weights.

$$\boldsymbol{Y}_t = \mathcal{W}(\boldsymbol{X}_t, \boldsymbol{A}_t), \tag{1}$$

where $\mathcal{W}$ is the affine warp function. The warp function first generates a sampling grid that maps the pixel coordinates of the output image from the input image using the affine transformation matrix $\boldsymbol{A}_t$. Then, the sampling grid is used to interpolate pixels from the input frame $\boldsymbol{X}_t$ to produce the aligned mouth frame $\boldsymbol{Y}_t$.

3.2. MAN distillation

To eliminate the need for explicit facial landmark detection, we propose a novel approach to directly produce the affine transformation parameters using a Mouth Alignment Network (MAN), as shown in the bottom of fig. 1. MAN is a light-weight mouth alignment network which has the architecture as shown in fig. 2a. It consists of a visual frontend and a temporal modeling module. The visual frontend is a ResNet-18 [10] that extracts per-frame features independently, while the temporal modeling module is a stack of two full pre-activation residual blocks [11], with group normalization [12] as the normalization layer and GELU [13] as the activation layer, which captures temporal dependencies across frames. The output of MAN is a sequence of affine transformation parameters $\{\hat{\boldsymbol{A}}_t\}_{t=1}^T$.

MAN is initially trained by distilling from the conventional mouth alignment pipeline by minimizing the L1 loss between the predicted parameters and the affine transformation extracted from the conventional pipeline:

$$\mathcal{L}_1 = \sum_{t=1}^{T} \|\hat{\boldsymbol{A}}_t - \boldsymbol{A}_t\|_1. \tag{2}$$

When distilling the conventional FAN-based pipeline, MAN inevitably inherits any of alignment errors from it. To push

MAN beyond these inherited limitations, we add a quality-aware training signal in the form of a differentiable score produced by a Mouth Scoring Network (MSN).

3.3. Mouth scoring network

MSN takes mouth video $\{\boldsymbol{Y}_t\}_{t=1}^{T}$ as input and outputs alignment quality scores, higher the better. As shown in fig. 2b, MSN uses a visual frontend (Conv3d stem with ResNet-18 trunk from Auto-AVSR [1]) followed by a temporal module (two stacked residual blocks as in MAN). Formally, MSN outputs quality scores $\{s_t\} = \mathcal{S}(\{\boldsymbol{Y}\})_t$ for each frame based on the full video sequence. Since ground-truth alignment quality labels are unavailable, we use preference modeling as described below.

3.4. Preference pairs construction

To apply preference modeling, pairs of preferred (high-quality) and non-preferred (degraded) samples are required. Since generating superior examples is difficult, we create degraded versions by applying controlled perturbations to affine transformation parameters, pairing original mouth frames with their perturbed counterparts.

The perturbation process simulates alignment errors through two methods: *global perturbation* applies a single random matrix across all frames to simulate systematic errors, while *local perturbation* modifies random frame subsets with independent matrices to create temporal inconsistencies. One pattern is randomly selected per training sample. We corrupt FAN's affine matrix using four transforms: translation $\boldsymbol{P}_t$ (shifting by δ_x, δ_y), scaling $\boldsymbol{P}_s$ (scaling by s_x, s_y), rotation $\boldsymbol{P}_r$ (rotating by θ), or composite $\boldsymbol{P}_c = \boldsymbol{P}_t\,\boldsymbol{P}_r\,\boldsymbol{P}_s$. Parameters are sampled from uniform distributions: $\delta_x, \delta_y \sim \mathcal{U}[-0.5, 0.5]$, $s_x, s_y \sim \mathcal{U}[0.5, 1.5]$, $\theta \sim \mathcal{U}[-\pi/2, \pi/2]$.

During training, two levels of perturbations are constructed. Starting with the original matrix $\boldsymbol{A}_t$, a random perturbation $\boldsymbol{P}_t^1$ is first applied to obtain $\tilde{\boldsymbol{A}}_t^1 = \boldsymbol{P}_t^1 \boldsymbol{A}_t$, and then a second random perturbation of the same type $\boldsymbol{P}_t^2$ is drawn to generate $\tilde{\boldsymbol{A}}_t^2 = \boldsymbol{P}_t^2 \tilde{\boldsymbol{A}}_t^1$. Warping the input frames $\boldsymbol{X}_t$ with these two matrices yields two mouth video variants: the once-perturbed frame $\tilde{\boldsymbol{Y}}_t^1 = \mathcal{W}(\boldsymbol{X}_t, \tilde{\boldsymbol{A}}_t^1)$, and the twice-perturbed sequence $\tilde{\boldsymbol{Y}}_t^2 = \mathcal{W}(\boldsymbol{X}_t, \tilde{\boldsymbol{A}}_t^2)$. We then can construct three pairs of mouth videos $(\{\boldsymbol{Y}_t\}, \{\tilde{\boldsymbol{Y}}_t^1\})$, $(\{\boldsymbol{Y}_t\}, \{\tilde{\boldsymbol{Y}}_t^2\})$ and $(\{\tilde{\boldsymbol{Y}}_t^1\}, \{\tilde{\boldsymbol{Y}}_t^2\})$. The preference pairs $(\{\boldsymbol{Y}_t\}, \{\tilde{\boldsymbol{Y}}_t^1\})$ and $(\{\boldsymbol{Y}_t\}, \{\tilde{\boldsymbol{Y}}_t^2\})$ enable the preference model to establish a clear decision boundary between clean and perturbed sequences. Meanwhile, the pair $(\{\tilde{\boldsymbol{Y}}_t^1\}, \{\tilde{\boldsymbol{Y}}_t^2\})$ trains the model to quantify the magnitude of perturbation between sequences that have undergone different levels of degradation.

3.5. Preference modeling

With the constructed preference pairs, we train MSN using a preference loss. The goal is to ensure MSN assigns higher quality scores to preferred mouth videos compared to their less preferred (perturbed) counterparts. As shown in fig. 2c, we utilize the Bradley-Terry model [14] to formalize this preference relationship. For any pair of mouth videos, $\{\boldsymbol{Y}_t^a\}$ and $\{\boldsymbol{Y}_t^b\}$, the probability that frame t from video a is preferred over frame t from video b, given the context of both videos, is modeled as:

$$p(\boldsymbol{Y}_t^a \succ \boldsymbol{Y}_t^b \mid \{\boldsymbol{Y}_t^a\}, \{\boldsymbol{Y}_t^b\}) = \sigma(s_t^a - s_t^b), \tag{3}$$

where $\{s_t^a\}_{t=1}^{T} = \mathcal{S}(\{\boldsymbol{Y}_t^a\})$ and $\{s_t^b\}_{t=1}^{T} = \mathcal{S}(\{\boldsymbol{Y}_t^b\})$ are the sequences of frame-wise quality scores produced by MSN $\mathcal{S}$, and σ denotes the sigmoid function. MSN is trained by maximizing the likelihood of observing these preferences across all relevant frames. This corresponds to minimizing the total negative log-likelihood loss $\mathcal{L}_s$, which aggregates the losses from the three types of pairs:

$$\mathcal{L}_s = f(\{\boldsymbol{Y}_t\}, \{\tilde{\boldsymbol{Y}}_t^1\}) + f(\{\boldsymbol{Y}_t\}, \{\tilde{\boldsymbol{Y}}_t^2\}) + f(\{\tilde{\boldsymbol{Y}}_t^1\}, \{\tilde{\boldsymbol{Y}}_t^2\}), \tag{4}$$

where the preference loss for a generic pair $(\{\boldsymbol{Y}_t^a\}, \{\boldsymbol{Y}_t^b\})$ is defined as:

$$f(\{\boldsymbol{Y}_t^a\}, \{\boldsymbol{Y}_t^b\}) = -\frac{\sum_t m_t \log \sigma(s_t^a - s_t^b)}{\sum_t m_t}. \tag{5}$$

Here, $m_t \in \{0, 1\}$ is a binary indicator that equals 1 if the t-th frame in $\boldsymbol{Y}_t^b$ has been perturbed relative to $\boldsymbol{Y}_t^a$, and 0 otherwise. Under local perturbation the loss targets only those perturbed frames, whereas under global perturbation it considers all frames.

3.6. MSN guidance

After pre-training the Mouth Scoring Network (MSN), we freeze its parameters and utilize it to guide the training of the Mouth Alignment Network (MAN). Let $\{\hat{\boldsymbol{A}}_t\}_{t=1}^{T}$ be the sequence of affine transformations predicted by MAN for an input video $\{\boldsymbol{X}_t\}_{t=1}^{T}$. The corresponding mouth video sequence is generated by warping: $\{\hat{\boldsymbol{Y}}_t\}_{t=1}^{T}$, where $\hat{\boldsymbol{Y}}_t = \mathcal{W}(\boldsymbol{X}_t, \hat{\boldsymbol{A}}_t)$. We then use the frozen MSN to compute frame-wise quality scores for this sequence and the original mouth video sequence $\{\boldsymbol{Y}_t\}_{t=1}^{T}$:

$$\{\hat{s}_t\}_{t=1}^{T} = \mathcal{S}(\{\hat{\boldsymbol{Y}}_t\}), \quad \{s_t\}_{t=1}^{T} = \mathcal{S}(\{\boldsymbol{Y}_t\}). \tag{6}$$

To maximize the average quality score over all frames, we define the guidance loss $\mathcal{L}_g$ as:

$$\mathcal{L}_g = \frac{1}{T} \sum_{t=1}^{T} \max(s_t - \hat{s}_t, 0). \tag{7}$$

Minimizing $\mathcal{L}_g$ encourages MAN to generate mouth sequences that MSN deems to be of high quality. We stopped optimizing $\hat{s}_t$ once it gets higher than the score of the original mouth video sequence s_t to avoid over-optimization.

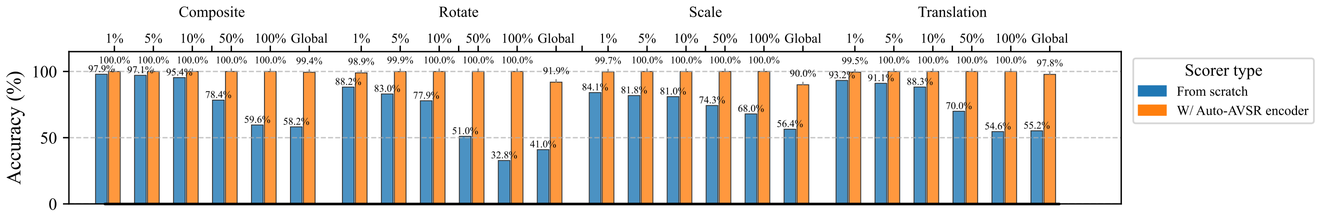


Fig. 3. Accuracy of different scorers on the LRS3 test set.

4. EXPERIMENTS

4.1. Experimental setup

4.1.1. Datasets

Following the conventional VSR protocol as in [1], our evaluation employed two benchmark datasets: LRS2 [15] and LRS3 [16]. The LRS2 dataset consists of 160 × 160 resolution video samples from BBC television programs, and the LRS3 dataset contains 224 × 224 resolution videos from TED talks. For our training regime, we used both LRS2 and LRS3 datasets. We randomly split LRS3's *trainval* set into 30,982 training samples and 1,000 validation samples since LRS3 has no official validation set. For LRS2, we used the *train* split containing 45,839 samples for training and the official validation set of 1,082 samples. We report results on both the validation sets and official test splits, comprising 1,243 samples from LRS2 and 1,321 samples from LRS3. Both MAN and MSN models were trained on the combined training set and validated on the merged validation sets. The evaluation was performed at 5,000 step intervals and we report the best performance achieved in the combined validation set.

4.1.2. Training details

MSN was trained for 10k steps using the Adam [17] optimizer with a learning rate of 2×10^{-4} and a batch size of 8. For MAN, we employed the Adam optimizer with an initial learning rate of 2×10^{-4}, applying a step-wise decay that reduced the rate by a factor of 0.1 every 10k steps. Training continued for 30k steps with a batch size of 4. All experiments were conducted on NVIDIA RTX 4090 GPU with 24GB memory. We utilized the following two state-of-the-art VSR models to test the trained MAN models: *Auto-AVSR*[1]: We employed the checkpoint trained on 3,448 hours of data, which achieved 19.1% WER on the LRS3 test set and 14.6% on the LRS2 test set. *USR*[2]: We used the high-resource model trained on both LRS3 and VoxCeleb2 [18] datasets, which achieved 22.3% WER on the LRS3 test set. Both models were originally built with the conventional FAN-based mouth alignment pipeline. In our evaluation, we replace the FAN-based mouth alignment module with our proposed MAN-based method, while keeping the rest of the VSR model unchanged.

4.2. MSN evaluation

We trained MSN using samples from the combined training set. During training, input mouth videos were randomly perturbed either globally (affecting all frames) or locally (affecting a randomly selected percentage of frames), each with a 50% chance by a random perturbation type. To evaluate the trained MSN's performance, we designed an artificial perturbation test. This test assesses the ability of MSN to assign a higher score to the preferred sample (the video processed by the conventional pipeline) compared to a randomly perturbed version of the same video. Accuracy is computed as the percentage of preference frame pairs where the preferred sample receives a higher score than its perturbed counterpart.

The evaluation used a test set derived from LRS3. For each sample, we applied four perturbation types (composite, rotation, scaling, and translation) at six variants: five local perturbation rates (1 – 100% of frames) plus one global application. Results in fig. 3 show the scorer with pretrained Auto-AVSR encoder consistently outperforms the from-scratch version[3], indicating the pretrained encoder better captures mouth alignment differences.

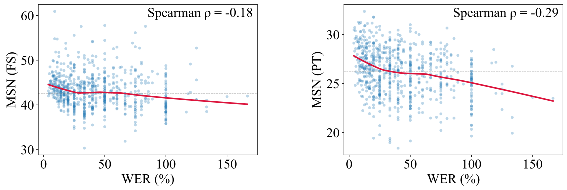


Fig. 4. MSN scores and WERs of Auto-AVSR on LRS3 test samples with WER $>$ 0% (673/1,321 samples). FS: from scratch; PT: frozen pretrained encoder. Dashed lines show medians.

In fig. 4, we plot MSN scores and the WERs of the Auto-AVSR model on the LRS3 test set with mouth aligned by the conventional FAN-based pipeline. MSN with the pretrained encoder exhibits a stronger Spearman correlation with WER on samples having WER $>$ 0% (-0.29) compared to the one

[1] Auto-AVSR: `https://github.com/mpc001/Visual_Speech_Recognition_for_Multiple_Languages`.

[2] USR: `https://github.com/ahaliassos/usr`.

[3] The *from-scratch* model randomly initialized encoder parameters and trained them with MSN, rather than using frozen Auto-AVSR weights.

trained from scratch (-0.18). This indicates that the pretrained encoder is more sensitive to the quality of mouth alignment and can better capture the relationship between mouth alignment quality and VSR performance. Moreover, the Lowess curve in fig. 4 shows that the negative correlation between MSN score and WER is more pronounced when the WER is high ($> 50\%$), suggesting that a certain amount of samples with poor WER is associated with a low MSN score.

4.3. MAN evaluation

After MSN training, we trained MAN using two strategies: (1) with only $\mathcal{L}_1$; and (2) with $\mathcal{L}_1$ and MSN guidance loss $\mathcal{L}_g$. We evaluated alignment performance by integrating the STNs with different VSR networks and compared against conventional FAN-based alignment. Results in table 1 show MAN trained with both $\mathcal{L}_1$ and $\mathcal{L}_g$ consistently outperforms the $\mathcal{L}_1$-only version in MSN scores, achieving better WERs in five of eight cases with only one worse result. This demonstrates that the proposed guidance loss enhances alignment performance, achieving comparable results to FAN-based methods, while $\mathcal{L}_1$-only training shows slightly degraded performance.

Table 1. MSN score and WER (%) on LRS2 and LRS3. MSN scores are produced by model with the pretrained Auto-AVSR encoder, *i.e.*, MSN (PT). Highlights show change w.r.t. the reproduced FAN baseline (bold: better, gray: worse).

Data	Method	MSN (PT) ↑		Auto-AVSR WER (%) ↓		USR WER (%) ↓	
		Val	Test	Val	Test	Val	Test
LRS2	FAN [5] (reported)	–	–	–	14.6	–	-
	FAN [5] (reproduced)	34.1	33.5	23.3	14.2	37.0	30.4
	MAN ($\mathcal{L}_1$)	33.5	32.9	23.8	14.5	37.1	30.7
	MAN ($\mathcal{L}_1 + \mathcal{L}_g$)	34.1	33.5	23.7	14.5	37.1	30.4
LRS3	FAN [5] (reported)	–	–	–	19.1	–	22.3
	FAN [5] (reproduced)	35.2	34.0	10.0	19.2	4.2	22.0
	MAN ($\mathcal{L}_1$)	34.9	33.8	10.1	19.5	4.2	22.4
	MAN ($\mathcal{L}_1 + \mathcal{L}_g$)	**35.5**	**34.4**	10.2	19.2	**4.1**	22.3

4.4. Efficiency analysis

As shown in table 2, while both models have comparable parameter counts, MAN demonstrates dramatically improved computational efficiency with only 1.8 GMac per frame compared to FAN's 14.0 GMac, which is a 7.7× reduction in computational complexity. This translates to significantly faster inference, with MAN processing frames in 5.9ms compared to FAN's 49.4ms, representing an 8.4× speed improvement.

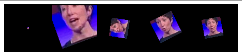
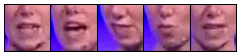
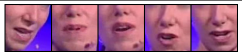
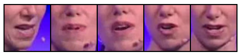
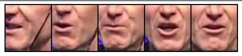
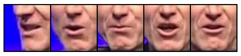
4.5. Qualitative study

In table 3, we present a qualitative comparison of samples with low MSN scores from the LRS3 test set. The FAN landmarks used in the conventional mouth alignment pipeline were obtained from the Auto-AVSR[4]. The samples show that the MSN score positively correlates with the mouth alignment quality, with a more stable mouth alignment leading to a higher score. The proposed MAN addresses severe misalignment issues, improving the performance of both VSR models. This is particularly evident in the cases of *ROgFmb3oTLo/00007* and *ROgFmb3oTLo/00005*, where WER improvements are substantial. Besides, MAN also effectively resolves subtle alignment issues, as seen in *CgNx9Bgac1I/00002*, even when these misalignments are not severe enough to negatively impact WER in the baseline system.

Table 2. Efficiency metrics on NVIDIA RTX 4090 GPU. Inference times are averaged over 1,000 runs of a 1-second clip at 25 FPS.

Metric	FAN [5]	MAN (Ours)
Parameters (M)	12.1	12.0
Flops per frame (GMac)	14.0	1.8
Inference time (ms)	49.4 ± 0.8	5.9 ± 0.4

Table 3. Qualitative comparison of representative utterances from the LRS3 test set. Five frames, evenly sampled across time, are shown from each video sequence.

ID	Method	Frames	MSN (PT) ↑	WER (%) ↓	
				Auto-AVSR	USR
ROgFmb3oTLo/00007	FAN		-72.9	100.0	87.5
↪	MAN		**34.7**	**37.5**	**50**
ROgFmb3oTLo/00005	FAN		19.7	37.5	37.5
↪	MAN		**35.3**	**0**	**25**
CgNx9Bgac1I/00002	FAN		21.5	0	0
↪	MAN		**34.7**	0	0

5. CONCLUSION

In summary, we replace the multistage landmark pipeline used in visual speech recognition with a compact Mouth Alignment Network (MAN) that directly regresses affine parameters from raw frames and is refined by a self-supervised Mouth Scoring Network (MSN). When plugged into state-of-the-art VSR models (Auto-AVSR, USR) on LRS2 and LRS3, MAN matches their word-error rates while cutting per-frame computation by roughly an order of magnitude, and it reliably corrects failure cases that defeat landmark-based alignment,

[4] FAN landmarks: `https://github.com/mpc001/Visual_Speech_Recognition_for_Multiple_Languages?tab=readme-ov-file#autoavsr-models`

improving the alignment quality measured by MSN. These results demonstrate that end-to-end, quality-guided mouth alignment can deliver both efficiency and robustness, paving the way for fast lip-reading systems and offering a template for learned alignment in other vision tasks.

6. ACKNOWLEDGEMENTS

The work described in this paper was partially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. T45-407/19N-1).

7. REFERENCES

[1] Pingchuan Ma, Alexandros Haliassos, Adriana Fernandez-Lopez, Honglie Chen, Stavros Petridis, and Maja Pantic, “Auto-AVSR: Audio-visual speech recognition with automatic labels,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[2] Alexandros Haliassos, Rodrigo Mira, Honglie Chen, Zoe Landgraf, Stavros Petridis, and Maja Pantic, “Unified speech recognition: A single model for auditory, visual, and audiovisual inputs,” *arXiv preprint arXiv:2411.02256*, 2024.

[3] Umberto Cappellazzo, Minsu Kim, Honglie Chen, Pingchuan Ma, Stavros Petridis, Daniele Falavigna, Alessio Brutti, and Maja Pantic, “Large language models are strong audio-visual speech recognition learners,” in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[4] Jeonghun Yeo, Seunghee Han, Minsu Kim, and Yong Man Ro, “Where visual speech meets language: VSP-LLM framework for efficient and context-aware visual speech processing,” in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, Eds., Miami, Florida, USA, Nov. 2024, pp. 11391–11406, Association for Computational Linguistics.

[5] Adrian Bulat and Georgios Tzimiropoulos, “How far are we from solving the 2D & 3D face alignment problem?(and a dataset of 230,000 3D facial landmarks),” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1021–1030.

[6] Alejandro Newell, Kaiyu Yang, and Jia Deng, “Stacked hourglass networks for human pose estimation,” in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*. Springer, 2016, pp. 483–499.

[7] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu, “Spatial transformer networks,” *arXiv preprint arXiv:1506.02025*, 2015.

[8] Young Jin Ahn, Jungwoo Park, Sangha Park, Jonghyun Choi, and Kee-Eung Kim, “SyncVSR: Data-efficient visual speech recognition with end-to-end cross-modal audio token synchronization,” *arXiv preprint arXiv:2406.12233*, 2024.

[9] Desire L Massart, Leonard Kaufman, Peter J Rousseeuw, and Annick Leroy, “Least median of squares: a robust method for outlier and model error detection in regression and calibration,” *Analytica Chimica Acta*, vol. 187, pp. 171–179, 1986.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Identity mappings in deep residual networks,” in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*. Springer, 2016, pp. 630–645.

[12] Yuxin Wu and Kaiming He, “Group normalization,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[13] Dan Hendrycks and Kevin Gimpel, “Gaussian error linear units (GELUS),” *arXiv preprint arXiv:1606.08415*, 2016.

[14] Ralph Allan Bradley and Milton E Terry, “Rank analysis of incomplete block designs: I. the method of paired comparisons,” *Biometrika*, vol. 39, no. 3/4, pp. 324–345, 1952.

[15] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman, “Deep audio-visual speech recognition,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 12, pp. 8717–8727, 2018.

[16] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman, “LRS3-TED: a large-scale dataset for visual speech recognition,” *arXiv preprint arXiv:1809.00496*, 2018.

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

[18] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, “Voxceleb2: Deep speaker recognition,” *arXiv preprint arXiv:1806.05622*, 2018.