Lip2Vec: Efficient and Robust Visual Speech Recognition via Latent-to-Latent Visual to Audio Representation Mapping (Supplementary Material)

Yasser Abdelaziz Dahou Djilali[1,2] Sanath Narayan[1] Haithem Boussaid[1]
Ebtessam Almazrouei[1] Merouane Debbah[1]
[1]Technology Innovation Institute, UAE [2]Dublin City University, Ireland

Here, we present additional quantitative and qualitative results of the Lip2Vec approach addressing the problem of visual speech recognition.

A. Varying the ASR Model and Video Encoder

The prior network $f_{\theta}(\cdot)$ in our Lip2Vec framework can be potentially trained with different off-the-shelf (pretrained) ASR models and video encoders. Here, we evaluate the performance of our Lip2Vec approach when utilizing VQ-Wav2Vec [1] as ASR model and VATLM [8] as the video encoder.

ASR model: The choice of utilizing VQ-Wav2Vec as an alternate ASR model is motivated by the fact that it is semantically different from Wav2Vec2.0, as it relies on a discrete latent space. Particularly, the model first encodes an input audio signal as vector quantized (VQ) representations through a codebook learned on top of the feature extractor. Then, the resulting discrete representations of the audio are input to BERT [2], which outputs enhanced representations based on their respective surrounding context. Finally, an acoustic model is utilized to predict text from the BERT output representations. While pretrained VQ-Wav2Vec and BERT models are readily avialable[1], the associated acoustic model is not. Therefore, we train a 6-layer transformer decoder (CE auto-regressive decoding) along with a linear layer (for CTC decoding) on the BERT representations using the audio-text pairs in LRS3 training set. This acoustic model obtains 11.2 WER on the LRS3 test set when using CE+CTC decoding.

Utilizing this VQ-Wav2Vec in our Lip2Vec indeed requires changing the prior network training objective to deal with codebook indices instead of continuous audio representations. Thus, we plug a classification head on the prior output to predict the codebook indices. Hence, we replace the cosine similarity loss with a standard cross entropy loss.

Table A.1: **Supervised finetuning *vs*. latent-to-latent training.** Comparison in terms of WER on LRS3 test set is shown. The same pretrained video encoder from AV-HuBERT [7] is finetuned (supervised w/ CE) or utilized for training the prior network in our Lip2Vec with two different ASR models: VQ-Wav2Vec and Wav2Vec2.0.

Encoder	Pretrain	Finetune	Supervised	Ours: Lip2Vec	
			S2S w/ CE	VQ-Wav2Vec	Wav2Vec2.0
Base	433h	30h	51.8	54.0	49.5
	1759h	30h	46.1	42.2	40.6
Large	433h	30h	44.8	57.5	55.4
	1759h	30h	32.5	33.5	31.2

Table A.2: **AV-HuBERT *vs*. VATLM as video encoder.** Comparison in terms of WER on LRS3 test set is shown. The pretrained video encoders from AV-HuBERT [7] and VATLM [8] are utilized for training the prior network in our Lip2Vec framework. The same ASR model (Wav2Vec2.0) is utilized for both experiments.

Encoder	Pretrain	Finetune	Video Encoder	
			VATLM	AV-HuBERT
Base	1759h	30h	42.5	40.6
Large	1759h	30h	33.0	31.2

Table A.1 shows the performance of our Lip2vec when using VQ-Wav2Vec as the ASR model in the low-resource setting (30h of finetuning data). We observe that it performs comparably with supervised finetuning of [7] across different settings, while requiring similar complexity due to CE+CTC decoding. The performance of our Lip2Vec when using Wav2Vec2.0 ASR model with CTC decoding alone is also shown for ease of comparison.

Video encoder: Here, we evaluate the performance of Lip2Vec when utilizing a different self-supervised video en-

[1] https://github.com/facebookresearch/fairseq/blob/main/examples/wav2vec/README.md#vq-wav2vec

Table A.3: **Impact of varying video length.** Comparison is shown in terms of WER on the LRS3 test set (denoted by All) along with four subsets of the same test set partitioned based on the length of the videos. LR and HR denote the low- and high-resource training with 30h and 433h of LRS3, respectively. Typically, text prediction is degraded for short sequences (less than 2 seconds) due to lack of contextual information during visual feature encoding.

Model	All	Video Length (in seconds)			
		0-2	2-4	4-6	> 6
VTP [6]	40.6	46.2	41.5	36.8	29.4
VTP [6] (2676h)	30.7	38.0	31.1	24.5	21.3
Ma *et al.* [5]	32.3	41.1	31.6	22.5	17.1
Ours: Lip2Vec (LR)	31.2	38.8	31.7	22.7	17.2
Ours: Lip2Vec (HR)	26.0	34.2	24.5	15.9	17.2

coder from VATLM [8]. It is worth mentioning that VATLM follows the same architecture and training procedure as AV-HuBERT. However, VATLM additionally utilizes the text modality during pretraining to enhance the features and promote for a unified latent space. Table A.2 shows the performance comparison when utilizing AV-HuBERT and VATLM encoders for training our prior network in the low-resource setting. Both encoders are pretrained on 1759h of LRS3+VoxCeleb2-en data.

Since VATLM utilizes text modality during pretraining, the resulting encoder representations are likely to be better aligned to the task of text prediction than for representing the lip sequences. Despite this, the VATLM encoder-based Lip2Vec achieves WER scores of 42.5 and 33.0 WER when using the Base and Large encoder architectures, respectively and performs comparably with the AV-HuBERT encoder-based Lip2Vec.

In summary, the aforementioned results and discussion demonstrate the capability of our Lip2Vec approach to successfully adapt to different ASR models and video encoders for learning the prior network using unlabelled video-audio pairs. Consequently, the Lip2Vec forms a viable alternative to video-text supervised finetuning.

B. Additional Results

In this section, we analyse the robustness of our Lip2Vec approach when varying the video sequence lengths and head poses of the speaker at test time. This is followed by a discussion on common failure cases and model consistency.

Varying the Video Length: Table A.3 shows the performance comparison on different folds obtained by partitioning the LRS3 test set based on the video sequence length. We observe that shorter videos (less than 2 seconds, *i.e.*, 50 frames) present a bottleneck, which results in performance degradation of the approaches from their corresponding average WER on the whole LRS3 test set (denoted as All in Table A.3). This is likely due to the lack of rich contextual features in shorter video sequences, which leads to sub-optimal temporal modeling in the video encoder. Consequently, the resulting representations output by the video encoder are not sufficiently discriminative for decoding the text correctly. Furthermore, we observe that the SoTA approaches and our Lip2Vec generally perform better with longer videos as input, indicating the importance of temporal modeling of visual features for accurate text decoding. However, targeting this issue is an important line of research to follow.

Table A.4: **Impact of head pose.** Comparison is shown in terms of WER on the LRS3 test set (denoted by All) along with two subsets: Frontal and Extreme, partitioned based on the head pose of the speaker in the video. LR and HR denote the low- and high-resource training with 30h and 433h of LRS3, respectively. Decoding text from partial/occluded lip motion at extreme head poses is challenging compared to frontal videos, where the lips are fully visible. See text for more details.

Model	All	Frontal	Extreme
VTP [6]	40.6	38.5	37.7
VTP [6] (2676h)	30.7	29.4	28.4
Ma *et al.* [5]	32.3	28.8	33.4
Ours: Lip2Vec (LR)	31.2	25.9	33.4
Ours: Lip2Vec (HR)	26.0	19.4	29.4

Varying Head Poses: Figure A.1 shows example frames from videos with frontal and extreme head poses in the LRS3 dataset. For this experiment, we select random 132 videos from LRS3 test for each of the subsets: frontal and extreme. We recover the 3D head pose by using a recently introduced method [3] targeting monocular 3D face reconstruction from talking face videos. Given a parametric 3D model [4] built from large datasets of 3D scans of human faces, this approach regresses the 3D model parameters that best fit to each image frame. We consider frontal and extreme based on predefined face angles.

Table A.4 shows the performance comparison between different approaches on both theses subsets, in terms of WER. We observe that decoding text from videos with extreme head poses is challenging since the lip sequences in such videos are only partially visible, resulting in less discriminative representations output by the video encoder. Among the approaches, only VTP achieves comparable results for both subsets. This is likely due to VTP utilizing the sequence of full images as input instead of the cropped lip sequences.

In summary, the presented Lip2Vec framework that learns a prior network using video-audio pair data performs favorably in comparison to other approaches across different settings with varying video lengths and head poses.

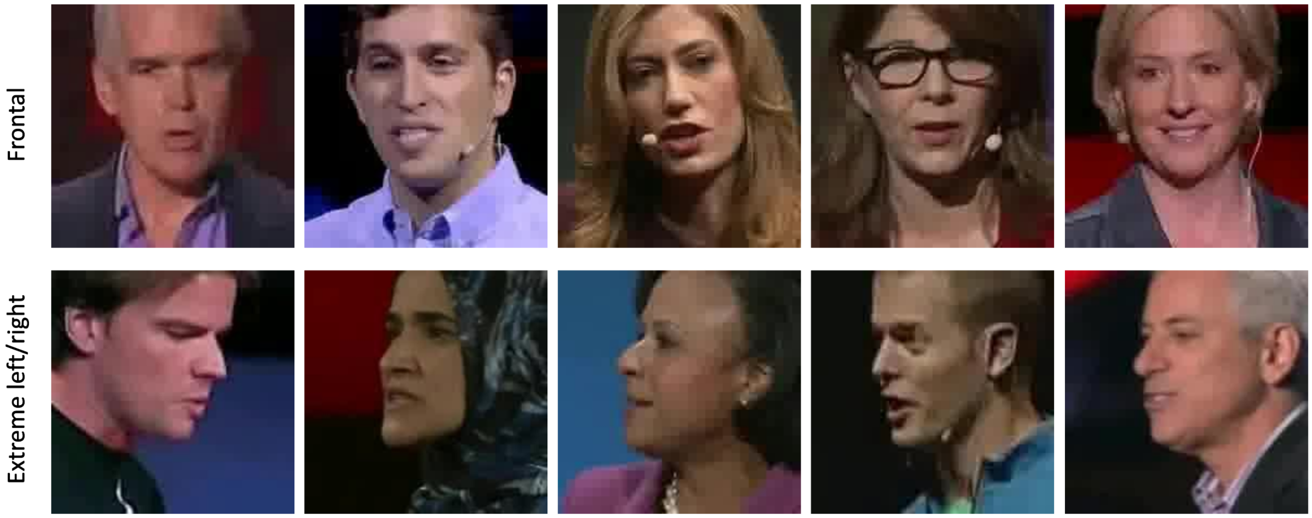

Figure A.1: **Frontal *vs*. extreme head poses in videos.** Top and bottom rows show example frames from videos having speakers with frontal and extreme (right/left) head poses, respectively. The lips sequences in extreme head poses are not completely visible and are likely to result in less discriminative representations output by the video encoders.

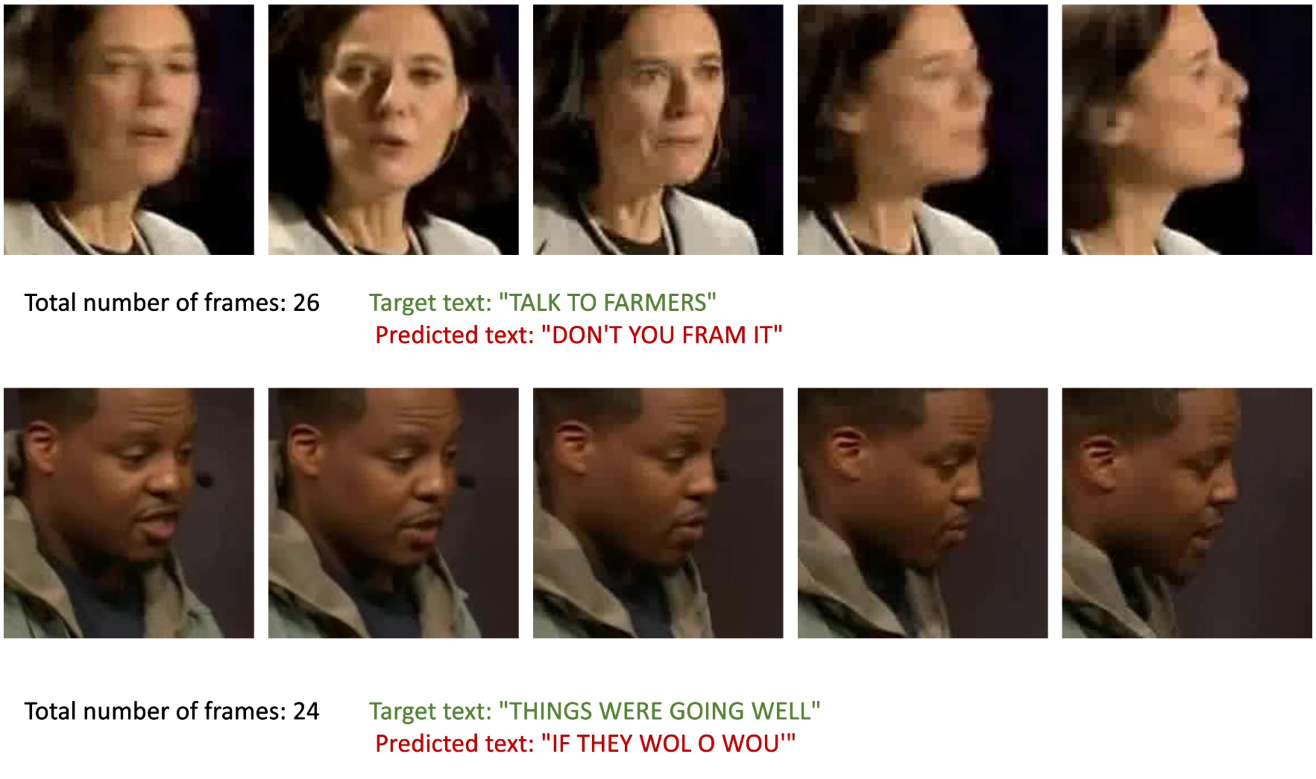
Total number of frames: 26
Target text: "TALK TO FARMERS"
Predicted text: "DON'T YOU FRAM IT"
Total number of frames: 24
Target text: "THINGS WERE GOING WELL"
Predicted text: "IF THEY WOL O WOU'"

Figure A.2: **Illustration of failure cases.** We observe the text decoding to be less accurate in case of short videos (around 1 second), where contextual representation is difficult. Furthermore, rapid variation of poses with blurry frames (top row) and extreme poses (bottom row) present a challenge for accurate text decoding. It is worth mentioning that although the predicted sentence for the top row video is not accurate, it has the same lip motion as the target sentence (*i.e*., they are homophemes).

Failure cases: Figure A.2 illustrates example failure cases of the Lip2Vec framework. In the top row, the model fails to adapt to rapid head motion (the speaker turns the head suddenly from left to right while talking) in a short sequence.

Additionally, the frames appear blurred due to the rapid motion, which likely affects the visual representations as well. The predicted sentence in this case, although incorrect, is still a homopheme and has the same lip motion as the target text. The bottom row example appears to be more challenging, since the subject has an extreme head pose all along the short sequence, leading to a set of poor visual representations and hence, failed decoding. A potential future direction, beyond the scope of the current work, could be to employ head pose normalization techniques as a preprocessing step to frontalize the videos and use them as input.

References

[1] Alexei Baevski, Steffen Schneider, and Michael Auli. vq-wav2vec: Self-supervised learning of discrete speech representations. *arXiv preprint arXiv:1910.05453*, 2019. 1

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[3] Panagiotis P. Filntisis, George Retsinas, Foivos Paraperas-Papantoniou, Athanasios Katsamanis, Anastasios Roussos, and Petros Maragos. Visual Speech-Aware Perceptual 3D Facial Expression Reconstruction from Videos. 7 2022. 2

[4] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics (TOG)*, 36(6), 11 2017. 2

[5] Pingchuan Ma, Stavros Petridis, and Maja Pantic. Visual speech recognition for multiple languages in the wild. *Nature Machine Intelligence*, pages 1–10, 2022. 2

[6] KR Prajwal, Triantafyllos Afouras, and Andrew Zisserman. Sub-word level lip reading with visual attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5162–5172, 2022. 2

[7] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. *arXiv preprint arXiv:2201.02184*, 2022. 1

[8] Qiushi Zhu, Long Zhou, Ziqiang Zhang, Shujie Liu, Binxing Jiao, Jie Zhang, Lirong Dai, Daxin Jiang, Jinyu Li, and Furu Wei. Vatlm: Visual-audio-text pre-training with unified masked prediction for speech representation learning. *arXiv preprint arXiv:2211.11275*, 2022. 1, 2