Visual Speech Recognition for Multiple Languages - Extended Abstract

Pingchuan Ma[1] Stavros Petridis[1,2] Maja Pantic[1,2]

[1]Imperial College London [2]Meta AI

1. Introduction

Visual speech recognition (VSR) aims to recognise the content of speech based on the lip movements without relying on the audio stream. Advances in deep learning and the availability of large audio-visual datasets have led to the development of much more accurate and robust VSR models than ever before. However, these advances are usually due to larger training sets rather than the model design. In this work, we demonstrate that designing better models is equally important to using larger training sets. We propose the addition of prediction-based auxiliary tasks to a VSR model and highlight the importance of appropriate data augmentations. We show that such model works for different languages and outperforms all previous methods trained on publicly available datasets by a large margin. We show furthermore that using additional training data, even in other languages or with automatically generated transcriptions, results in further improvement.

2. Methodology

2.1. Prediction-based Auxiliary Tasks

The proposed model is shown in Fig. 1. It is based on the hybrid/CTC architecture proposed in [8] which is augmented with the addition of auxiliary tasks. We propose as an auxiliary task the prediction from intermediate layers of audio and visual representations learned by pre-trained ASR and VSR models (pre-trained as explained in [8]). This is inspired by the recent success of prediction tasks in self-supervised learning. In particular, good audio representations can be learned by predicting handcrafted audio features [12] or by using joint audio and visual supervision [16]. Similarly, visual speech representations can be learned by predicting audio features [7]. Hence, the proposed auxiliary task provides additional supervision to the intermediate layers of the model which in turns results in better visual representations and improved performance. This results in the following loss term added to loss function:

$$\begin{aligned}\mathcal{L}_{AUX} = \beta_a \left\| h_a(f^l(\mathbf{x}_v)) - g_a^l(\mathbf{x}_a)\right\|_1 \\ + \beta_v \left\| h_v(f^l(\mathbf{x}_v)) - g_v^l(\mathbf{x}_v)\right\|_1\end{aligned} \tag{1}$$

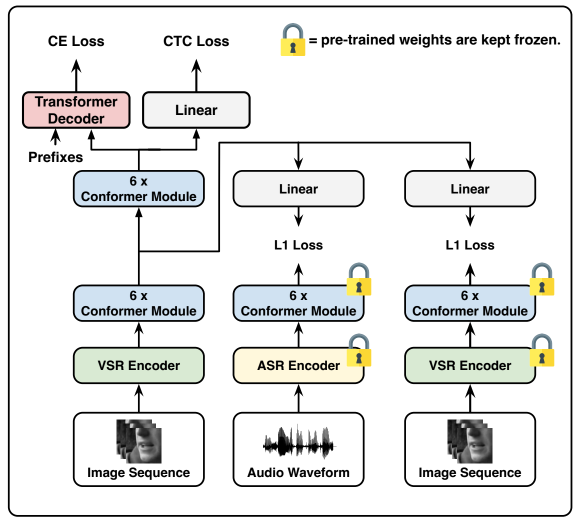

Figure 1. Summary of the proposed model with prediction-based auxiliary tasks. In this figure, the pre-trained ASR/VSR encoders and some conformer layers are frozen and their internal representations are used as targets for the audio and visual predictors.

where $\mathbf{x}_v$ and $\mathbf{x}_a$ are the visual and audio input sequences, respectively, g_v and g_a are the pre-trained visual and audio encoders, respectively. f is the subnetwork up to layer l whose intermediate representation is used as input to the audio and visual predictor h_a and h_v, respectively. β_a and β_v are the coefficients for each loss term and $\|\cdot\|_1$ is the ℓ_1-norm.

The model performs VSR and at the same time attempts to predict audio and visual representations from intermediate layers. Hence, the final loss is the following:

$$\mathcal{L} = \mathcal{L}_{VSR} + \mathcal{L}_{AUX} \tag{2}$$

$$\mathcal{L}_{VSR} = \alpha\mathcal{L}_{CTC} + (1-\alpha)\mathcal{L}_{att} \tag{3}$$

where $\mathcal{L}_{VSR}$ is the loss of the hybrid CTC/attention architecture used. $\mathcal{L}_{CTC}$ is the CTC loss, $\mathcal{L}_{att}$ the loss of the attention mechanism and α controls the relative weight of each loss term.

2.2. Time Masking

In this work we propose the use of time masking which is commonly used in training ASR models [11]. It works by

Method	Pre-training Set	Training Set	Training Sets Total Size (hours)	WER	CER
Results on the LRS3 dataset					
Using Publicly Available Datasets					
KD+CTC [3]	VoxCeleb2^{clean}	LRS3	772	59.8	-
CM-seq2seq [8]	LRW	LRS3	595	43.3	-
Ours	-	LRS3	438	**37.9**	-
Ours	LRW	LRS2+LRS3+AVSpeech+VoxCeleb2	3 388	**26.1**	-
Using Non-Publicly Available Datasets					
TM-seq2seq [1]	MVLRS+LRS2	LRS3	1 391	58.9	-
V2P [15]	-	LSVSR	3 886	55.1	-
RNN-T [10]	-	YT-31k	31 000	33.6	-
ViT3D-TM [13]	-	YT-90k	90 000	25.9	-
ViT3D-CM [14]	-	YT-90k	90 000	19.3	-
Results on the CMLR dataset					
LIBS [19]	-	CMLR	61	-	31.3
CTCH [9]	-	CMLR	61	-	22.0
Ours	-	CMLR	61	-	**9.1**
Results on the CMU-MOSEAS-Spanish (CM_{es}) dataset					
CM-seq2seq [8]	LRW	CM_{es}+MT_{es}	244	58.1	-
Ours	LRW	CM_{es}+MT_{es}	244	**50.4**	-

Table 1. Summary of our results. WER: Word Error Rate. CER: Character Error Rate.

randomly masking n consecutive frames by replacing them with the mean sequence frame. This allows the model to more effectively use contextual information and can better disambiguate similar lip movements which correspond to different phonemes. It also makes the model more robust to short missing segments.

3. Experiments

3.1. Datasets

For the purposes of this study we use the **LRS3** [2] dataset, which is the largest publicly audio-visual English dataset collected from TED talks, **CMLR** [18], which is the largest audio-visual Mandarin dataset collected from Chinese national news program, and **CMU-MOSEAS-Spanish (CM_{es})** [17], which is an audio-visual Spanish dataset. Furthermore, we also use the English-only version of VoxCeleb2 [4], and AVSpeech [6]. The transcriptions for these datasets are automatically generated using the ASR model from Wav2Vec2-Base-960h[1].

[1] https://huggingface.co/facebook/wav2vec2-base-960h

3.2. Results

Results on LRS3, which is an English audio-visual dataset, are presented in Table 1. Our proposed approach significantly outperforms all existing works which are trained using publicly available datasets. In particular, our method leads to better performance than the state-of-the-art [8] even though it is trained only on the LRS3 training set and no external datasets are used for pre-training. In case of additional training data being available, our method leads to an 17.2 % absolute improvement in word error rate (WER) over the state-of-the-art [8]. It is worth pointing that such a significant improvement is observed although automatically generated transcriptions are used for AVSpeech and VoxCeleb2. This confirms the recent trend observed in the literature where using larger training sets results in better performannce. We should also emphasize that we achieve a very similar WER to [13] despite using 26.5 times less training data.

Results on the CMLR dataset, which is a Mandarin audio-visual dataset, are also shown in Table 1. We report performance in terms of character error rate (CER) instead of WER because Chinese characters are not separated by spaces. Our approach results in a significant reduction in the CER over all existing works. We achieve an absolute improvement of 12.9 % in CER over the state-of-the-art [9].

Results on the CMU-MOSEAS-Spanish dataset, which is an audio-visual Spanish dataset, are shown in Table 1. Given that this is a small dataset it is not possible to train an accurate model without using additional data. For this purpose, we first pre-train the model on the LRW dataset [5] and then fine-tune it on the training sets of CMU-MOSEAS using the Spanish videos only. Since this is a new dataset and there are no results from prior works, we have trained the end-to-end model presented in [8] to serve as the baseline. Our proposed approach results in a 7.7 % absolute reduction in the WER.

4. Conclusion

In this work, we presented our approach for visual speech recognition and demonstrated that state-of-the-art performance can be achieved not only by using larger datasets, which is the current trend in the literature, but also by carefully designing a model. We proposed a new architecture based on auxiliary tasks where the VSR model also predicts audio visual representations learned by pre-trained ASR and VSR models. Our approach outperforms all existing VSR works trained on publicly available datasets in English, Spanish and Mandarin by a large margin.

Acknowledgements

All training, testing, and ablation studies have been conducted at Imperial College London.

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*, 2018. 2

[2] T. Afouras, J. S. Chung, and A. Zisserman. LRS3-TED: a large-scale dataset for visual speech recognition. Preprint at https://arxiv.org/abs/1809.00496, 2018. 2

[3] T. Afouras, J. S. Chung, and A. Zisserman. ASR is all you need: Cross-modal distillation for lip reading. In *Proceedings of the 45th IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2143–2147, 2020. 2

[4] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. In *Proceedings of the 19th Annual Conference of International Speech Communication Association*, pages 1086–1090, 2018. 2

[5] J. S. Chung and A. Zisserman. Lip reading in the wild. In *Proceedings of the 13th Asian Conference on Computer Vision*, volume 10112, pages 87–103, 2016. 3

[6] 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 Transactions on Graphics*, 37(4):112:1–112:11, 2018. 2

[7] P. Ma, R. Mira, S. Petridis, B. W. Schuller, and M. Pantic. LiRA: Learning visual speech representations from audio through self-supervision. In *Proceedings of the 22nd Annual Conference of International Speech Communication Association*, pages 3011–3015, 2021. 1

[8] P. Ma, S. Petridis, and M. Pantic. End-to-end audio-visual speech recognition with conformers. In *Proceedings of the 46th IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7613–7617, 2021. 1, 2, 3

[9] S. Ma, S. Wang, and X. Lin. A transformer-based model for sentence-level chinese mandarin lipreading. In *Proceedings of the 5th IEEE International Conference on Data Science in Cyberspace*, pages 78–81, 2020. 2, 3

[10] T. Makino, H. Liao, Y. Assael, B. Shillingford, B. Garcia, O. Braga, and O. Siohan. Recurrent neural network transducer for audio-visual speech recognition. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, pages 905–912, 2019. 2

[11] D. S. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le. Specaugment: A simple data augmentation method for automatic speech recognition. In *Proceedings of the 20th Annual Conference of International Speech Communication Association*, pages 2613–2617, 2019. 1

[12] S. Pascual, M. Ravanelli, J. Serrà, A. Bonafonte, and Y. Bengio. Learning problem-agnostic speech representations from multiple self-supervised tasks. In *Proceedings of the 20th Annual Conference of International Speech Communication Association*, pages 161–165, 2019. 1

[13] D. Serdyuk, O. Braga, and O. Siohan. Audio-visual speech recognition is worth 32×32×8 voxels. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, pages 796–802, 2021. 2

[14] D. Serdyuk, O. Braga, and O. Siohan. Transformer-based video front-ends for audio-visual speech recognition. Preprint at https://arxiv.org/abs/2201.10439, 2022. 2

[15] B. Shillingford, Y. Assael, M. W. Hoffman, T. Paine, C. Hughes, U. Prabhu, H. Liao, H. Sak, K. Rao, L. Bennett, et al. Large-scale visual speech recognition. In *Proceedings of the 20th Annual Conference of International Speech Communication Association*, pages 4135–4139, 2019. 2

[16] A. Shukla, S. Petridis, and M. Pantic. Learning speech representations from raw audio by joint audiovisual self-supervision. In *Proceedings of the 37th International Conference on Machine Learning Workshop*, 2020. 1

[17] A. B. Zadeh, Y. Cao, S. Hessner, P. P. Liang, S. Poria, and L. Morency. CMU-MOSEAS: A multimodal language dataset for spanish, portuguese, german and french. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 1801–1812, 2020. 2

[18] Y. Zhao, R. Xu, and M. Song. A cascade sequence-to-sequence model for chinese mandarin lip reading. In *Proceedings of the 1st ACM International Conference on Multimedia in Asia*, pages 1–6, 2019. 2

[19] Y. Zhao, R. Xu, X. Wang, P. Hou, H. Tang, and M. Song. Hearing lips: Improving lip reading by distilling speech recognizers. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, pages 6917–6924, 2020. 2