Speech Recognition Models are Strong Lip-readers

K R Prajwal, Triantafyllos Afouras, Andrew Zisserman

VGG, University of Oxford

{prajwal,afourast,az}@robots.ox.ac.uk

Abstract

In this work, we show that a large pre-trained ASR model can be adapted to perform lip-reading. Our method enables an ASR model like Whisper to interpret lip movements in a video and output text transcriptions. We achieve this by learning a cross-modal mapping from a lip sequence to a speech sequence, allowing a pre-trained ASR model to directly perform lip-reading. The mapping can be learnt simply by backpropagating the cross-entropy loss on the text labels through the pre-trained, frozen ASR model. We achieve an impressive gain of 5.7 WER in the low data regime on the LRS3 benchmark over previous lip-reading methods. Finally, we demonstrate that the same strategy can be extended to other visual speech tasks, such as identifying the spoken language in silent videos.

Index Terms: speech recognition, lip reading

1. Introduction

The advent of foundational models has brought to light previously unexplored dimensions of deep learning's ability to generalize. Large models pre-trained on massive datasets achieve state-of-the-art performance on a wide range of downstream tasks. Even more remarkably, these models also demonstrate strong cross-modal learning capabilities [1, 2, 3].

In this paper, we study the cross-modal generalization capabilities of a pre-trained automatic speech recognition (ASR) model like Whisper [4] for the highly challenging task of visual speech recognition (lip-reading). This allows us to leverage the speech modeling capabilities of state-of-the-art ASR models. ASR models can be trained on massive audio datasets and learn to form compact latent speech representations that are then decoded into text. Therefore, in order to learn a strong lip-reading model, we can rely on a pre-trained ASR model for the speech and language modeling and need only focus on learning two things specific to the visual modality: a good visual representation, and a module for mapping the visual features into a speech-like representation. This approach paves the road to much more sample-efficient learning for lip-reading, which is important because publicly available transcribed video datasets are orders of magnitude smaller than their audio counterparts. For instance, among the public datasets, VoxPopuli [5] consists of 400K hours of multi-lingual speech data, whereas LRS3 [6], the most commonly used public lip-reading benchmark contains only about 400 hours. In addition to sample-efficiency, using a pre-trained ASR model provides an additional benefit of saving computational resources for training, as a large part of the model is reused.

The essence of our approach to lip-reading is to replace the audio input sequence of a pre-trained ASR model with the visual input sequence of a talking face, while maintaining the speech transcription abilities of the pre-trained ASR model. We study the recently released Whisper ASR model [4], which is based on a vanilla Transformer encoder-decoder architecture using audio spectrograms as input, and has been trained on 680k hours of weakly-labelled audio speech data.

Our contributions are: (1) re-purposing a large-scale, frozen, pre-trained ASR model to perform lip-reading; (2) achieving state-of-the-art results on the standard lip-reading benchmarks, and (3) demonstrating that our approach can be successfully extended to the visual language identification task, obtaining a significant improvement over previous methods. Please follow the website for demo samples and code release: https://www.robots.ox.ac.uk/~vgg/research/whisperer

2. Related Work

2.1. Lip reading

The advent of deep learning, along with the creation of larger datasets such as LRW [7], LRS2 [8] and LRS3 [6] has led to rapid development of strong machine lip reading models in the past few years. Sentence-level lip reading methods followed two main paradigms, borrowed from the machine translation and ASR literature, namely sequence-to-sequence [9] and CTC [10, 11], as well as hybrid approaches [12]. The sequence learning architecture has also evolved from LSTMs to Transformers [13], to recent works using Conformers [14, 15, 16] that can model both the local information and long-term temporal information.

Self-supervised learning for lip reading. Following the trend in speech recognition [17, 18, 19, 20], there have been many works [21, 22, 23] that learn audio-visual speech representations from unlabeled videos. These models are data-efficient – they can be fine-tuned with far fewer video-text pairs to achieve a much better performance. The most prominent one among them is AV-HuBERT [23] which learns representations by predicting masked cluster assignments. In this work, we adapt these AV-HuBERT features to condition a pre-trained ASR model like Whisper.

Utilizing ASR for lip-reading. Over the years, efforts have been made to leverage ASR models to enable lip-reading training. For example [24] obtain noisy text labels for training by running an ASR on the accompanying audio. A similar data-efficient and scalable approach was shown by Auto-AVSR [25], which obtained impressive performance gains by pre-training on large amounts of videos labeled with noisy transcripts from a pre-trained ASR. Although our method also relies on a pre-trained ASR model, we follow a different approach: we do not merely use ASR as a labeling tool – the ASR is the key component of the lip-reading model itself. This idea is by no means a

new one, previous methods [26, 27] have shown that ASR models can be adapted for lip-reading. However, they achieve a far worse performance and employ additional pre-training strategies and loss functions to match the audio and video features. Our approach is straightforward and much faster to train – we directly train for lip-reading using limited video-text pairs and achieve a superior performance as shown in Table 1. Our approach is easily extensible to solve additional tasks (Section 6).

2.2. Visual language Identification

Language identification is a well-studied problem [28, 29] as it is the first step for multi-lingual ASR. With the recent efforts being made towards multi-lingual lip-reading [16, 30], identifying the spoken language from lips becomes an important task. Few modern works exist in this space with more recent ones [31] aggregating lip-reading representations with shallow temporal processing networks, followed by simple classifier heads.

3. Method

In this section, we describe our approach. Our goal is to show that we can adapt a pre-trained ASR model to lip-read, by mapping visual features representing lip movements into the same space as the ASR model's audio features. In this paper, we realize this goal using the Whisper ASR model [4]. An overview of the approach is given in Fig. 1. It consists of three major parts: (i) a visual feature encoder, (ii) a cross-modal mapping network, and (iii) a pre-trained ASR network, which is the Whisper model in our case. We explain each of the three parts below.

3.1. Visual backbone

Our lip reading network's input is a silent video clip of T frames, $\boldsymbol{x} \in \mathbb{R}^{T\times H\times W\times 3}$. We use a pre-trained AV-HuBERT-Large model [23] to extract the visual features for this input. It consists of two modules: (i) A ResNet18-like CNN, and a (ii) Transformer encoder, both of which are trained in a self-supervised manner through masked cluster prediction. The video input sequence is divided into sub-clips of 5 frames (i.e. 0.2s) with a unit frame stride. The sub-clips are then processed by a ResNet18-like CNN architecture, where only the first CNN layer is spatio-temporal. The CNN outputs T feature vectors, one per input frame: $\boldsymbol{f} \in \mathbb{R}^{T\times C}$, where $C = 1024$ in our implementation. The feature sequence $\boldsymbol{f}$ is processed by a series of Transformer encoder layers after a linear projection to give the visual features $\boldsymbol{g_{enc}} \in \mathbb{R}^{T\times C}$. Note that we obtain our visual features without any text supervision.

3.2. Cross-modal mapping network

Given the visual features $\boldsymbol{g_{enc}}$, we would like to now map them to the space of the ASR model's audio features. We start by re-sampling the video features to match the number of time-steps in the audio stream. In the case of Whisper, the audio input is at twice the frame rate (50 FPS) of the video input stream, so visual inputs are simply repeated along the time dimension in order to match the audio time-steps. We denote the time-steps as T instead of $2T$ for simplicity. The re-sampled visual features ($\boldsymbol{g_{enc}}$) are passed through a cross-modal mapping network $\mathbb{M}$, which consists of an an MLP and a stack of N Transformer encoder layers to map these features to Whisper's audio space. The output of the mapping network $\boldsymbol{h_{enc}}$ can be directly fed into the pre-trained ASR model.

$$\boldsymbol{h_{enc}} = \mathbb{M}(\boldsymbol{g_{enc}} + PE_{1:T}) \in \mathbb{R}^{T\times C}.$$

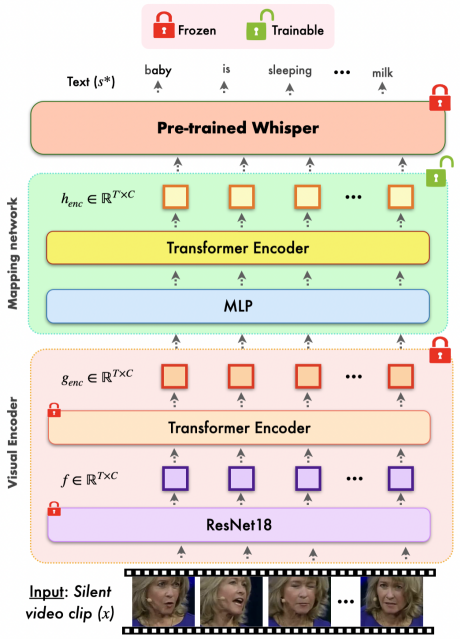

Figure 1: ***Proposed lip reading network***. *Given a silent video clip, we first encode it using a pre-trained AV-HuBERT encoder consisting of a ResNet18 followed by a Transformer encoder. The obtained visual features are temporally up-sampled (not shown in the figure for simplicity) and passed through a cross-modal mapping network to obtain inputs that can be fed into a pre-trained Whisper model. The frozen Whisper model outputs text transcriptions. Only the mapping network (specified by open lock) is trained.*

where PE is the temporal position embedding.

3.3. Conditioning the ASR with visual inputs

The next step is to condition the ASR with the mapped features $\boldsymbol{h_{enc}}$. We briefly describe the ASR we use below.

Whisper architecture. Whisper inputs audio mel-spectrogram features and embeds these using a stack of convolutional layers to get the initial audio feature vectors $\boldsymbol{A} \in \mathbb{R}^{T\times D}$. These are then processed by a Transformer encoder network to get $\boldsymbol{A_{enc}} = \mathbb{W}_{enc}(\boldsymbol{A}) \in \mathbb{R}^{T\times D}$. A Transformer decoder is used to auto-regressively decode the text tokens s while cross-attending to $\boldsymbol{A_{enc}}$. We use Whisper-medium in all our experiments and keep the Whisper model weights frozen.

Feeding the mapped visual features to Whisper. There are several possibilities for feeding in the mapped visual features $\boldsymbol{h_{enc}}$ to Whisper. We found that feeding $\boldsymbol{h_{enc}}$ in place of $\boldsymbol{A}$, i.e., directly to the start of the Transformer encoder of Whisper ($\mathbb{W}_{enc}$) gives the best results. Thus, the conditioning of Whisper on the visual features $\boldsymbol{h_{enc}}$ to transcribe the text sequence s can be expressed as follows:

$$\log p(s|\boldsymbol{h_{enc}}) = \sum_{t=1}^{T_{dec}} \log p(s_t|\, \mathbb{W}_{enc}(\boldsymbol{h_{enc}}), s_{1:t-1}) \quad (1)$$

The output sequence s is represented in Byte-Pair-Encoding (BPE) tokens, exactly the same as the ones used by the Whisper model [4]. At inference, we perform a left-to-right beam search of width $B = 40$. We do not perform any test-time augmentations and do not use any external language models.

3.4. Training objective

We maximise the log likelihoods of the text transcriptions by optimising the following objective:

$$L = -\mathbb{E}_{(x,s^*)\in\mathcal{D}} \log p(s^*|\boldsymbol{x}) \quad (2)$$

4. Experiments

4.1. Data

The visual feature extractor from AV-HuBERT is pre-trained on the English split of VoxCeleb2 [32] and LRS3 [6], a total of 1759 hours of unlabeled video data. Our model is trained and evaluated on the largest publicly available sentence-level lip-reading dataset, LRS3 [6]. LRS3 contains clips from over $5,000$ TED and TEDx talks in English, available on YouTube, totalling 475 hours. The manual text transcriptions were automatically aligned to the audio using force-alignment to yield word boundaries; these word alignments enable training at any granularity. The dataset is split into pretrain (403h), trainval (30h) and test sets. For the low-resource setting, we use only the LRS3-trainval set of 30h, as also done in [23]. The models are evaluated on the LRS3 test set containing 1,321 samples. Both LRS3 and VoxCeleb have been created using a detection and tracking pipeline that produces clips loosely cropped around the speaker's talking head. All videos are available at 25 fps, accompanied with audio at 16 kHz. We follow AV-HuBERT's official code to extract tight mouth ROIs from the face tracks. The transcripts are tokenized with Byte-Pair Encoding (BPE) in the exact same manner as done in Whisper's official code.

4.2. Implementation details

We now provide the implementation details of our best performing model. We use AV-HuBERT Large as our visual feature extractor. The pre-trained model weights are obtained by optimizing for the self-supervised clustering objective on 1759 hrs of unlabeled video data from LRS3 [6] and the English split of VoxCeleb2 [32]. The feature encoder consists of a ResNet18 followed by 24 transformer encoder layers with a hidden dimension of 1024 and 16 heads. Our mapping network $\mathbb{M}$ consists of (i) an MLP with two FC layers with ReLU activation and Layer Normalization, and (ii) a $N = 3$ layer Transformer encoder. The feature dimensions and the number of heads match that of the Whisper's encoder, which is $C = 1024, heads = 16$. For the Whisper model, we use Whisper-medium-English, whose weights are also publicly available. We use the Adam optimizer [33] with an initial learning rate of $5e^{-5}$. The learning rate is reduced by 5 every time the WER on the LRS3 validation set does not improve for 2 consecutive epochs, with a minimum learning rate of $1e^{-6}$. We finetune the AV-HuBERT encoder at this minimum learning rate for two epochs and choose the checkpoint with the best validation word error rate. We clip all gradients to have a maximum L2 norm of 0.1. The training takes only a few hours on 4 Tesla V100 GPUs.

5. Results

5.1. Quantitative scores

In Table 1, we compare our method with the previous works on the LRS3 test set. We categorize the previous works in Table 1 into three groups. The first two groups use only publicly available text transcripts for training and evaluation such as LRS2 [8], LRS3 [6]. The first group consists of models that are trained only with text supervision. The second group involves a mix of self-supervised representation learning and/or ASR-aided training, followed by finetuning with manually annotated clean transcripts. The third and final group are methods that use large, proprietary datasets for training.

Method	**Unlabeled (hrs)**	**Labeled (hrs)**	**WER (%)**
Fully Supervised models with publicly available data			
ASR distillation [24]	-	590	68.8
Conv-Seq2Seq [34]	-	855	60.1
Discriminative AVSR [35]	-	590	57.8
Hyb. + Conformer [14]	-	590	43.3
VTP [36]	-	698	40.6
Self-supervised pre-training + Supervised finetuning			
ASR distillation [24]	334	590	59.8
LiRA [37]	433	30	71.9
	1,759	433	49.6
LiteVSR [27]	639	59	45.7
AV-HuBERT Large [23]	1,759	30	32.5
		433	28.6
SynthVSR [38]	3,652	438	27.9
Auto-AVSR [25]	1,759	433	25.0
Lip2Vec [26]	1,759	30	31.2
		433	26.0
Ours	1,759	30	**25.5**
		433	**24.3**
Trained on large-scale non-publicly available datasets			
Deep-AVSR [9]	-	1,519	58.9
Large-scale AVSR [11]	-	3,886	55.1
RNN-T [39]	-	31,000	33.6
VTP [36]	-	2,676	30.7
ViT-3D [40]	-	90,000	17.0
Synth-VSR [38]	3,652	3,068	16.9
LP [15]	-	1,00,000	12.5

Table 1: *WER (%) of our models and the comparison with prior works on the LRS3 test set. Among the different approaches trained on similar number of hours of labeled data, our method gives clear improvements.*

Our method falls into the second category. We consider two settings, depending on the amount of labeled data used for training. In the 30hrs setting, we achieve a WER of **25.5**, which is **5.7 WER** points better than the next best model (31.2) at this data scale. Our best model that uses 433hrs of labeled data achieves a WER of **24.3**, surpassing all other approaches that train using the same datasets. Thus, our approach is particularly useful when there is limited video-text paired training data.

5.2. Whisper decoder's cross-attention heatmaps

We now investigate how the pre-trained Whisper model behaves when we feed in (i) audio input to perform ASR, (ii) the mapped visual features (h_{enc}) to perform lip-reading. We specifically visualize the cross-attention heatmaps between the text tokens and the input source (audio or video). Figure 2, plots the cross-attention heatmaps for a sample from the LRS3 test set for audio and video inputs. It can be seen that, except for a few minor differences, the attention plots are strikingly similar, indicating that the Whisper's text decoder is able to interpret the input from different modalities in a similar manner.

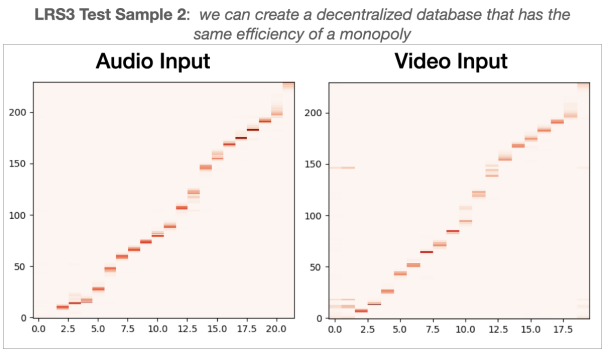

Figure 2: ***Whisper's cross attention map*** *for a specific decoder layer's attention head. X-axis denotes the text tokens. Y-axis denotes the input source time-steps which can be either audio (ASR) or video (lip reading). The cross-attention plot are very similar across modalities, indicating that the decoder is not distinguishing between the modalities.*

5.3. Ablation studies

In the arXiv version of the paper, we provide ablations of our approach, analysing the various design choices such as: (i) the right place to feed the visual features to Whisper, (ii) which mapping network is the best, (iii) how the size of the Whisper model affects the lip-reading performance, (iv) does fine-tuning the ASR model help? We also analyze the performance of our model against various sequence lengths.

6. Extending to Visual Language Identification

We demonstrate that our approach can be easily extended to another task that Whisper is proficient at: language identification. Given a speech input, Whisper has been trained to identify the language of the speech in addition to transcribing what is being spoken. Similar to our re-purposing of Whisper for lip-reading, we show that we can identify the language spoken purely from the lip movements, i.e. visual language identification [31].

Dataset. Due to the lack of public benchmarks for this task, we show results by creating train-val-test splits on the LRS3-lang+ dataset [31]. This dataset contains 14 languages sourced from 19k TEDx talks on YouTube. We split the dataset into 85-5-10 for train-val-test splits. We ensure that the same TEDx talk is not repeated across the splits to ensure there is no overlap of content or speakers. Details on the distribution of the languages in the dataset is provided in the original paper [31] and also in the arXiv version of the paper.

Approach. We follow the same approach as we did for lip-reading, with only a few minor changes. We now switch from a monolingual (English) Whisper model to the multi-lingual variant. For example, when given a French speech, the multilingual Whisper model would output a transcripts like <|fr|><|transcribe|> *Chacun voit midi à sa porte.* <|endoftranscript|>. For this task, we would like to output only the first token, which gives us the language code. Thus, the task of language identification is cast as a sequence-to-sequence task, where the input can be speech (visual or audio) and output is the language code.

Implementation details and Evaluation. We follow the same hyper-parameter settings that we use for the lip-reading model. We train on 1–5 second clips. At test-time, we evaluate on clips of lengths up to 5 seconds. We use the average class accuracy (chance: 7%) to measure the model's performance.

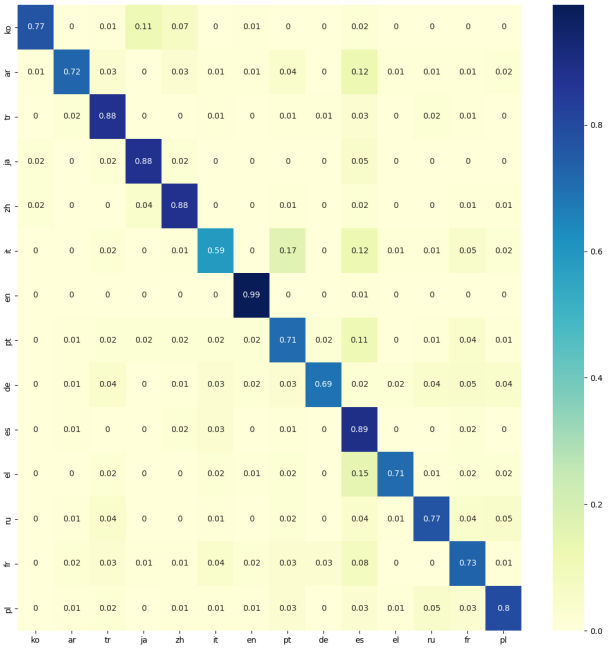

Figure 3: ***Row-normalized Confusion matrix*** *of the language identification predictions on the test set. The languages are as follows: ('ko' : Korean), ('ar' : Arabic, ('tr' : Turkish), ('ja' : Japanese), ('zh' : Mandarin), ('it' : Italian), ('en' : English), ('pt' : Portugese), ('de' : German), ('es' : Spanish), ('el' : Greek), ('ru' : Russian), ('fr' : French), and ('pl' : Polish).*

The current best model [31] evaluates on 5s clips and reports a mean class accuracy of **67.2%**. Unfortunately, the authors did not release the train-test splits to directly compare with this score. On our test split, which can contain clips of less than 5s (which makes it harder due to shorter segments), we achieve a mean class accuracy of **78.7%**, which is already far better than the previously reported performance. We also show the confusion matrix across the different languages in Fig 3. English and Spanish are by far the most accurate of the languages, presumably because there are more training data for these two languages. Italian seems to be a very difficult language, often being confused with Portuguese, Spanish or French. The visual language identification results serve as a strong evidence that our approach can be extended to solve several other lip-related tasks that Whisper performs with audio inputs.

7. Conclusion and extensions

We present an approach that employs an ASR model to achieve state-of-the-art performance in lip-reading. Our method is also well-positioned to make use of the rapid ongoing progress in large-scale speech recognition models [23, 41, 22]. We believe that this work is only the tip of an iceberg – it is yet to be explored if we can adapt the visual features to the full range of Whisper's capabilities other than English speech recognition and language identification: (i) multi-lingual lip-reading, (ii) visual speech detection (detecting the speech time-steps), (iii) visual speech-to-text translation (e.g. English lip movements to German text), and (iv) conditioning on a larger context. We believe that our approach can have an impact across all these tasks, especially when there is a scarcity of labeled data.

Acknowledgements. Funding for this research is provided by the EPSRC Programme Grant VisualAI EP/T028572/1, and a DeepMind Graduate Scholarship.

8. References

[1] M. Tsimpoukelli, J. L. Menick, S. Cabi, S. Eslami, O. Vinyals, and F. Hill, "Multimodal few-shot learning with frozen language models," *NeurIPS*, vol. 34, pp. 200–212, 2021.

[2] J.-B. Alayrac *et al.*, "Flamingo: a visual language model for few-shot learning," *arXiv preprint arXiv:2204.14198*, 2022.

[3] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," 2023.

[4] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *arXiv preprint arXiv:2212.04356*, 2022.

[5] C. Wang, M. Riviere, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, "VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," in *ACL*, Aug. 2021, pp. 993–1003.

[6] T. Afouras, J. S. Chung, and A. Zisserman, "Lrs3-ted: a large-scale dataset for visual speech recognition," in *arXiv preprint arXiv:1809.00496*, 2018.

[7] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *ACCV*, 2016.

[8] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *CVPR*, 2017.

[9] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE TPAMI*, 2019.

[10] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas, "Lipnet: Sentence-level lipreading," *arXiv:1611.01599*, 2016.

[11] B. Shillingford, Y. Assael, M. Hoffman, T. Paine, C. Hughes, U. Prabhu, H. Liao, H. Sak, K. Rao, L. Bennett, M. Mulville, B. Coppin, B. Laurie, A. Senior, and N. Freitas, "Large-scale visual speech recognition," in *Interspeech*, 2019.

[12] S. Petridis and M. Pantic, "Deep complementary bottleneck features for visual speech recognition," *ICASSP*, pp. 2304–2308, 2016.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[14] P. Ma, S. Petridis, and M. Pantic, "End-to-end audio-visual speech recognition with conformers," in *ICASSP*, 2021.

[15] O. Chang, H. Liao, D. Serdyuk, A. Shah, and O. Siohan, "Conformers are all you need for visual speech recogntion," *arXiv preprint arXiv:2302.10915*, 2023.

[16] P. Ma, S. Petridis, and M. Pantic, "Visual speech recognition for multiple languages in the wild," *NMI*, pp. 1–10, 2022.

[17] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *arXiv preprint arXiv:1904.05862*, 2019.

[18] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[19] W.-N. Hsu *et al.*, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[20] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "Data2vec: A general framework for self-supervised learning in speech, vision and language," in *ICML*, 2022, pp. 1298–1312.

[21] W.-N. Hsu and B. Shi, "u-hubert: Unified mixed-modal speech pretraining and zero-shot transfer to unlabeled modality," in *Advances in Neural Information Processing Systems*, 2022.

[22] J. Lian, A. Baevski, W.-N. Hsu, and M. Auli, "Av-data2vec: Self-supervised learning of audio-visual speech representations with contextualized target representations," *arXiv preprint arXiv:2302.06419*, 2023.

[23] B. Shi, W.-N. Hsu, K. Lakhotia, and A. Mohamed, "Learning audio-visual speech representation by masked multimodal cluster prediction," *arXiv preprint arXiv:2201.02184*, 2022.

[24] T. Afouras, J. S. Chung, and A. Zisserman, "ASR is all you need: Cross-modal distillation for lip reading," in *International Conference on Acoustics, Speech, and Signal Processing*, 2020.

[25] P. Ma, A. Haliassos, A. Fernandez-Lopez, H. Chen, S. Petridis, and M. Pantic, "Auto-avsr: Audio-visual speech recognition with automatic labels," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[26] Y. A. D. Djilali, S. Narayan, H. Boussaid, E. Almazrouei, and M. Debbah, "Lip2vec: Efficient and robust visual speech recognition via latent-to-latent visual to audio representation mapping," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 13 790–13 801.

[27] H. Laux, E. Mededovic, A. Hallawa, L. Martin, A. Peine, and A. Schmeink, "Litevsr: Efficient visual speech recognition by learning from speech representations of unlabeled data," *arXiv preprint arXiv:2312.09727*, 2023.

[28] F. Richardson, D. Reynolds, and N. Dehak, "Deep neural network approaches to speaker and language recognition," *IEEE signal processing letters*, vol. 22, no. 10, pp. 1671–1675, 2015.

[29] J. Gonzalez-Dominguez, D. Eustis, I. Lopez-Moreno, A. Senior, F. Beaufays, and P. J. Moreno, "A real-time end-to-end multilingual speech recognition architecture," *IEEE Journal of selected topics in signal processing*, vol. 9, no. 4, pp. 749–759, 2014.

[30] M. Anwar, B. Shi, V. Goswami, W.-N. Hsu, J. Pino, and C. Wang, "Muavic: A multilingual audio-visual corpus for robust speech recognition and robust speech-to-text translation," *arXiv preprint arXiv:2303.00628*, 2023.

[31] T. Afouras, J. S. Chung, and A. Zisserman, "Now you're speaking my language: Visual language identification," *Proceedings of ISCA 2020*, no. 2020, 2020.

[32] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *INTERSPEECH*, 2018.

[33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[34] X. Zhang, F. Cheng, and S. Wang, "Spatio-temporal fusion based convolutional sequence learning for lip reading," in *ICCV*, 2019.

[35] B. Xu, C. Lu, Y. Guo, and J. Wang, "Discriminative multi-modality speech recognition," in *CVPR*, 2020.

[36] K. Prajwal, T. Afouras, and A. Zisserman, "Sub-word level lip reading with visual attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5162–5172.

[37] P. Ma, R. Mira, S. Petridis, B. Schuller, and M. Pantic, "LiRA: Learning visual speech representations from audio through self-supervision," in *Interspeech*, 2021.

[38] X. Liu, E. Lakomkin, K. Vougioukas, P. Ma, H. Chen, R. Xie, M. Doulaty, N. Moritz, J. Kolar, S. Petridis, M. Pantic, and C. Fuegen, "Synthvsr: Scaling up visual speech recognition with synthetic supervision," in *CVPR*, June 2023, pp. 18 806–18 815.

[39] 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 *Interspeech*, 2019.

[40] D. Serdyuk, O. Braga, and O. Siohan, "Transformer-based video front-ends for audio-visual speech recognition," *arXiv preprint arXiv:2201.10439*, 2022.

[41] A. Haliassos, P. Ma, R. Mira, S. Petridis, and M. Pantic, "Jointly learning visual and auditory speech representations from raw data," *arXiv preprint arXiv:2212.06246*, 2022.