Scaling Multilingual Visual Speech Recognition

K R Prajwal*
VGG, University of Oxford
prajwal@robots.ox.ac.uk

Sindhu Hegde*
VGG, University of Oxford
sindhu@robots.ox.ac.uk

Andrew Zisserman
VGG, University of Oxford
az@robots.ox.ac.uk

***Abstract*—Visual Speech Recognition (lip-reading) has witnessed tremendous improvements, reaching word error rates as low as 12.8 WER in English. However, the performance in other languages is lagging far behind, due to the lack of labeled multilingual video data. In this work, we reduce the performance gap with the help of three key advances: (i) introducing the largest multilingual lip-reading dataset to date, (ii) proposing a single multi-task architecture that can perform two tasks simultaneously: identify the language and transcribe the utterance, and (iii) jointly training this architecture on all the languages together, resulting in large WER improvements as opposed to training monolingual models separately. We achieve state-of-the-art performance in both visual language identification and multilingual lip-reading tasks. Moreover, our pipeline uses zero manual annotations, as all the training transcriptions are obtained using a pre-trained ASR model. We also show that our multilingual model can be readily fine-tuned for new low-resource languages on which models trained from scratch do not converge. Our data, code, and models are available at: www.robots.ox.ac.uk/~vgg/research/multivsr.**

***Index Terms*—lip-reading, multilingual dataset, lip-to-text**

I. Introduction

Automatic Speech Recognition (ASR) models have taken giant strides in recent years, achieving human-like accuracy [1]–[5] across a large number of languages [6] while also being fast and real-time. Visual speech recognition (VSR), a.k.a. lip-reading, is a far more challenging task compared to ASR, because lip movements are inherently ambiguous to transcribe from. In spite of this, the word error rates (WER) have plummeted from 50-60 [7], [8] to just 12.8 [9] on the LRS3 [10] test set in a matter of five years. However, even today, in stark contrast to ASR models, lip-reading models have been fairly accurate only for English. One possible reason for this could be the lack of resources for non-English data. The state-of-the-art multilingual lip-reading model today [11], in spite of making significant improvements in WER, still only uses 331 hours of Spanish data, which is 25% smaller than that used for English even in 2018 [12], and 95% smaller than what is used for English lip-reading today [13].

In this work, we scale multilingual lip-reading with the help of three key advances. We start by collecting *MultiVSR*, a large-scale multilingual dataset from YouTube videos. With $\approx 12,000$ hours of video data across various languages, this is the largest multilingual lip-reading dataset to date, as far as we know. Our data collection process requires little to no manual effort, because the transcriptions are automatically obtained using the pre-trained models in the WhisperX [14], [15] ASR library. Using this data, we train a single model that can perform two tasks in a single framework: identifying language from lip movements and transcribing what was uttered. We train this multi-task model jointly across all languages. This is in contrast to previous works [11], [16] where each language has a separate model. We clearly show that the joint multilingual training leads to large gains in each individual language. We also demonstrate that our pre-trained multilingual model can be fine-tuned for a new low-resource language on which training from scratch diverges (> 100 WER).

In summary, our contributions are: (i) A new large-scale multilingual dataset, *MultiVSR*, comprising nearly $12,000$ hours of video data, (ii) A multi-task model capable of both language identification and transcription from silent lip

*equal contribution

This research was funded by EPSRC Programme Grant VisualAI EP/T028572/1, and a Royal Society Research Professorship.

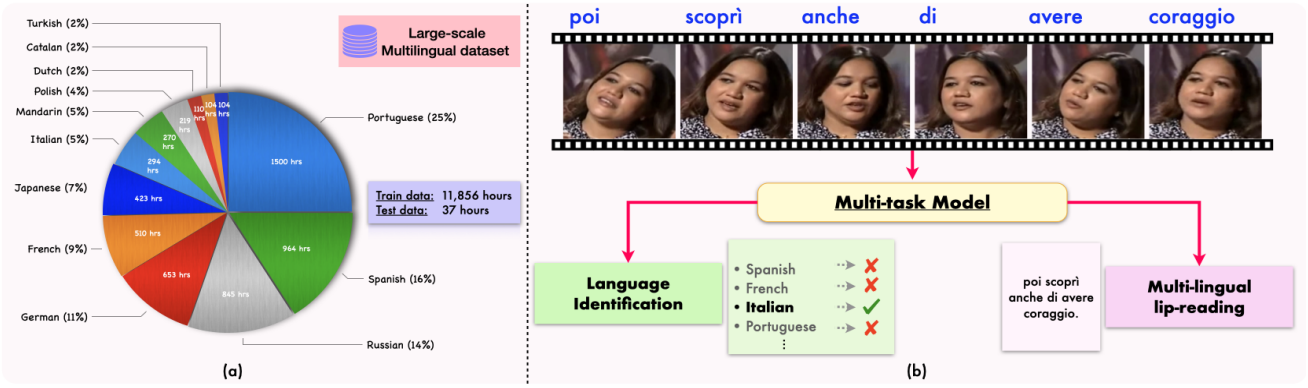


Fig. 1. (a) The *MultiVSR* dataset: we curate the largest multilingual lipreading corpus to date, consisting of $\approx 12,000$ hours of video paired with word-aligned transcripts (5859 hours English & 5996 hours non-English). (b) We develop a unified multi-task framework capable of performing both lip-reading and visual language identification tasks. The model, trained on our newly collected dataset, achieves state-of-the-art results in both tasks.

movements, and (iii) A unified multilingual lip-reading model that is far superior to previous monolingual models across all the languages, and can generalize to new low-resources languages with limited training data.

II. Related Work

English Lip-reading: Lip-reading has gained significant attention in recent years, with advancements in both the scale of training data and model architectures leading to substantial performance improvements. Early works primarily focused on English-speaking datasets, such as LRW [17], LRS2 [7], and LRS3 [10], which laid the foundation for visual speech recognition. With the availability of vast amounts of video data, numerous works [8], [12], [13], [18]–[30] have emerged, resulting in WER dropping from 50-60 [7], [8] to just 12.8 [9] on the LRS3 [10] test set in a matter of five years.

Non-English Lip-reading: Non-English Lip-reading remains under-explored primarily due to the lack of large-scale multilingual datasets. Existing multilingual video datasets such as VoxCeleb2 [31] and AVSpeech [32] are not accompanied by transcripts, and the individual language subsets are relatively small in scale. The multilingual TEDx (mTEDx) [33] dataset provides speech transcriptions in several languages, but it lacks video data and even if we take the corresponding (unofficial) video clips from YouTube, it is too small for training accurate lip-reading models. MuAViC [34] extends mTEDx by including videos, but has been mainly used for audio-visual ASR and translation. Recently, a few works [11], [16], [35] have focused on multilingual lip-reading, proposing various methods to either combine or extend existing datasets to cover multiple languages. The largest multilingual lip-reading dataset to date was curated in [11], where the authors merged data from mTEDx, VoxCeleb2, and AVSpeech to create a dataset spanning four languages: Spanish (384 hours), French (331 hours), Portuguese (420 hours), and Italian (152 hours). Despite a big drop in WER, we note that the LRS3 dataset [10] used for training English lip-reading in 2018 [12] was bigger than each of the above language splits. Moreover, most of the recent works [11], [16], [35] have only trained monolingual lip-reading models, which we show is sub-optimal.

Visual Language Identification: A critical pre-processing step for multilingual lip-reading is to identify what language is being uttered. Only a handful of works [25], [36], [37] have tackled this problem, and none of them perform multilingual transcription. We outperform the previous best model [25], while also doing this jointly with multilingual lip-reading.

III. A New Large-scale Multilingual Dataset

We present *MultiVSR*, a new large-scale multilingual lip-reading dataset that includes video clips paired with aligned text transcripts. The dataset covers English along with 12 other languages: Portuguese, Spanish, Russian, German, French, Japanese, Italian, Mandarin, Polish, Dutch, Catalan, and Turkish, with the last three being the low-resource languages. Each clip features a visible speaker who is the sole person speaking in the video. In total, the dataset encompasses $11,856$ hours of video segments, representing a diverse range of individuals, languages, and facial poses. Figure 1 (left) gives an overview of the data and Table I contains the detailed dataset statistics.



Fig. 2. ***MultiVSR* Dataset:** (a) We begin by collecting a large set of $200K$ YouTube videos across 13 languages. (b) These videos are pre-processed by detecting and tracking person crops. (c) In the final step, we utilize SyncNet [38] to discard clips that do not contain the face of the speaker and the clips with synchronization offsets of $\geq \pm 10$ frames.

A. Data collection and pre-processing

The dataset is automatically gathered from publicly available YouTube videos. We obtained the list of YouTube IDs from the AVSpeech dataset [32] which consists of $\approx 200K$ YouTube videos across a large number of languages. While AVSpeech [32] uses only a few seconds from each video to yield 4700 hrs of data, we download and process the entire video, leading to a vastly different dataset in terms of number of identities and the number of hours (approx. $12K$ hours). This *scale* is essential for most of the non-English languages.

Pre-processing: Our pre-processing pipeline (Fig 2) involves two stages: (i) person tracking and cropping, and (ii) synchronization detection. We begin by resampling all videos to 25 fps and then detecting faces in each video frame using the S3FD face detector [39]. We then track segments where the face is clearly visible and speaking (see Figure 2 (b)). Clips shorter than 1 second are discarded, and we also filter out faces with a resolution below 96×96 pixels. Next, in the synchronization stage, we verify whether the crop is synchronized with the corresponding audio segment. Specifically, we use SyncNet [38] to calculate synchronization offsets and retain only the crops where the offset is within ± 10 frames by sync correction (see Figure 2 (c)). The clips containing a different person other than the speaker, and clips where the synchronization offset is above ± 10 frames are discarded. This step is critical since we rely on a pre-trained automatic speech recognition (ASR) model to generate transcripts from the audio segments (described below), rather than manual labeling.

B. Automatic labelling

As demonstrated in [11], automatic annotations are scalable and produce results comparable to manual labeling for lip-reading; we thus opt for automatic label generation for our dataset. We first identify the language using a model trained on the VoxLingua dataset [40]. We generate the word-level time-aligned transcripts for the video segments using the pre-trained Whisper Large-v3 ASR model in WhisperX [15]. We chose the 13 major languages that had at least 100 hours of training data. This process results in approximately $1.66M$ video clips across $123K$ YouTube videos, comprising a total of $11,893$ hours of footage.

TABLE I

DATA STATISTICS: THE *MultiVSR* DATASET COMPRISES VIDEO WITH ALIGNED TRANSCRIPTS ACROSS 13 LANGUAGES, TOTALING ≈ 12,000 HOURS.

Language	**Split**	English	Portuguese	Spanish	Russian	German	French	Japanese	Italian	Mandarin	Polish	Dutch	Catalan	Turkish	**Total**
# clips	trainval	843377	176978	128241	90455	93000	84986	69125	41165	41789	30515	33016	18343	11396	**1,662,386**
	test	n/a	2056	1961	1191	1053	1058	933	728	337	263	309	212	240	**10,341**
# unique videos	trainval	62750	13570	10285	7553	7161	5320	4545	3161	2829	2384	1484	1225	1102	**1,23,369**
	test	n/a	380	298	201	191	147	126	87	62	58	47	36	38	**1,671**
avg. per clip (secs)	trainval	25.01	30.52	27.06	33.62	25.28	21.61	22.03	25.74	23.24	25.85	12.01	20.34	32.94	**25.67**
	test	n/a	5.06	3.89	7.59	6.39	5.77	3.66	4.45	5.22	6.93	5.58	6.76	4.46	**5.13**
duration (hours)	trainval	5859.09	1500.38	963.79	844.85	652.95	510.16	423.06	294.35	269.73	219.11	110.17	103.66	104.27	**11,855.57**
	test	n/a	7.03	6.48	5.90	4.08	4.17	2.54	2.84	1.02	0.96	0.70	0.94	0.62	**37.28**

C. Data Splits

In Table I, we show the data splits: trainval and test, with no speaker common across the two splits. We also exclude the speakers present in the test sets of LRS3 [10] and mTEDx [33], ensuring that models trained on our data can be benchmarked on these two datasets, if necessary. We also manually verify our test set and correct transcription errors.

IV. MULTI-TASK NETWORK

We design a multi-task encoder-decoder model as illustrated in Figure 3. Given a silent video sequence with N_f frames, $V = \{V_1, V_2, \ldots, V_{N_f}\}$, the objective is to predict both the language token L and the corresponding text sequence, $T = \{T_1, T_2, \ldots, T_W\}$, where W is the number of words. The model is jointly trained in a sequence-to-sequence framework [41] to perform the tasks of language identification and lip-reading simultaneously.

The visual features are extracted using the pre-trained Visual Transformer Pooling (VTP) model [23], which has a temporal footprint of 5 frames and serves as a robust visual backbone. It outputs per-frame lip movement features, $f_{vtp} \in \mathbb{R}^{N_f \times d}$. We follow a simple Transformer encoder-decoder architecture from [23]. The N-layer Transformer encoder temporally encodes the VTP features to produce $E_{enc} \in \mathbb{R}^{N_f \times d}$. This is followed by a N-layer Transformer Decoder, which cross-attends to the E_{enc} and auto-regressively predicts both the language token and the text transcription.

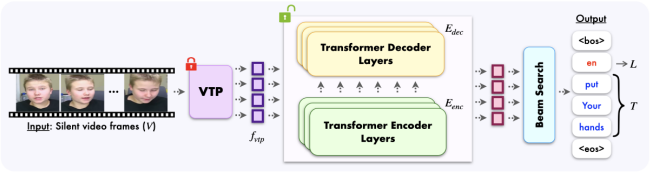


Fig. 3. **The multi-task architecture:** The input video frames are first passed through the Visual Transformer Pooling (VTP) model, which extracts the visual features, f_{vtp}. These features are then processed by a Transformer encoder-decoder model, generating the embeddings E_{dec}. Finally, the network outputs both the language token L for visual language identification and the text sequence T for lip-reading, using beam search.

A. Implementation details

Input/output data: We use Whisper's tokenizer [14] with a shared vocabulary of 51865 tokens for all the languages. This also includes the special tokens such as the start and end tokens, and language tokens. At each training iteration, using the start-end time of the word boundaries from WhisperX [15], we randomly sample segments of up to 14 seconds in length from the full video clips. For the mTEDx dataset, we use the full clips specified by [11].

Hyper-parameters and Training: We set the number of layers N to 12, with a hidden dimension of 768 and number of attention heads as 12. We train the lip-reading model with the Adam optimizer with β_1 = 0.9, β_2 = 0.98 and a learning rate of $1e^{-4}$. We reduce the learning rate to $2e^{-5}$ when validation loss plateaus and train until convergence (4 days).

Inference: The first predicted token is always the language token L. Thus, the model directly performs language identification as part of the transcription. While decoding the text T, we use beam search with a beam width of 20. We do not use any external language model to improve the performance.

V. EXPERIMENTS

A. Datasets

We leverage our newly collected large-scale dataset of approximately 12,000 hours, as outlined in Table I, for both training and testing. A single model is jointly trained on all of the major 10 languages (11,537 hrs), for both lip-reading and VLID tasks. The remaining 3 languages (shaded in blue in Table I) contain 100 hours each and are used to show low-resource fine-tuning performance (Sec V-D). To compare with the existing multilingual works, we also report our results on mTEDx dataset [33] as explained below.

Additional dataset – mTEDx: Multilingual TEDx (mTEDx) [33] is a publicly available multilingual speech corpus, but lacks video data. Thus, we prepare the video data following [11], [35] and use it to evaluate against the current lip-reading models. As done in prior state-of-the-art lip-reading works [11], [16], [35], we consider four languages for evaluation: Spanish, French, Portuguese, and Italian.

B. Lip-reading Results

Results on the *MultiVSR* dataset: We evaluate our model on the test set of our newly collected dataset across nine non-English high resource languages and report the word error rate (WER). For comparison, we use the pre-trained model released by Yeo et al. [11] and evaluate it on our test set. [11] trains separate models for each language, whereas ours is a single model trained on all languages collectively. As shown

TABLE II
PERFORMANCE COMPARISON OF LIP-READING AND VISUAL LANGUAGE IDENTIFICATION (VLID) ACROSS NINE LANGUAGES ON *MultiVSR* DATASET.

Task	Method	Portuguese	Spanish	Russian	German	French	Japanese	Italian	Mandarin	Polish
Lip-reading (WER) ↓	Yeo et al. [11]	40.91	40.99	n/a	n/a	45.31	n/a	41.26	n/a	n/a
	Ours	**33.76**	**37.04**	**39.49**	**40.35**	**34.63**	**37.21**	**38.66**	**55.54**	**40.44**
VLID (accuracy) ↑	Prajwal et al. [25]	77.22	77.55	80.03	63.17	78.32	90.82	35.56	73.38	46.80
	Ours	**98.34**	**95.18**	**97.64**	**96.52**	**96.61**	**98.92**	**91.21**	**82.28**	**95.80**

in Table II, our model clearly outperforms the current state-of-the-art across all languages resulting in a $\approx 3-7$ WER average improvement over [11] across four languages. The gains can largely be attributed to the (i) larger number of hours in each language and the quality of the transcriptions, and (ii) joint multilingual training to further increase the data scale. Note that [11] also trains the visual backbone along with Conformer layers [22], which is far superior to our method that trains a simple Transformer on pre-trained frozen VTP features [23].

Results on mTEDx dataset: To validate the effectiveness of our proposed model and dataset, and to maintain consistency with previous lip-reading studies, we compare with the prior state-of-the-art [11], [16], [35] on the standard mTEDx dataset. We first discard all clips that have a face size of $\leq 64 \times 64$ pixels. We fine-tune on the train split specified in [11] and report results on the unseen speakers of test clips [11] in Table III. In-line with our earlier findings, our model outperforms existing methods across all languages.

TABLE III
LIP-READING WERS ON MTEDX DATASET [33]. LOWER IS BETTER.

Method	Portuguese	Spanish	French	Italian
Ma et al. [16]	61.50	56.30	66.20	57.40
Kim et al. [35]	58.57	56.96	64.92	59.74
Yeo et al. [11]	47.89	45.71	58.30	51.79
Ours	**45.25**	**42.25**	**52.39**	**44.57**

C. Visual Language Identification (VLID) Results

Our network is also capable of identifying the language based solely on lip movements. Notably, we do not train a separate model for this task; instead, it is learned jointly during the lip-reading training process. Table II shows the results of VLID across all the 9 languages. We compare with the VLID model from Prajwal et al. [25] after fine-tuning it on the MultiVSR dataset. Note that [25] trains a VLID network independently from the lip-reading task. From Table II, it is evident that our joint model achieves an unprecedented accuracy compared to the previous best model [25], owing to the benefits of training with a joint lip-reading objective.

D. Extending to new low-resource languages

To address the challenge of training models for languages with limited data, we investigate the feasibility of using our pre-trained multilingual model for enabling lip reading for low-resource languages. Specifically, we consider three languages for this study, each with about 100 hours of training data, as opposed to at least 200 hours with the other main 10 languages in the dataset. We curate the training and testing data for these languages (Dutch, Catalan, Turkish) in the exact same manner as described in Section III.

Firstly, we note that training models only on these languages from scratch does not work – the models overfit without learning meaningful representations and obtain a word error rate of $97 - 105$. Fine-tuning from our joint pre-trained multilingual model (trained on $11,537$ hours) produces good results. We fine-tune for each of these languages separately, at four different data scales: 25h, 50h, 75h and 100h. As shown in Figure 4, fine-tuning with as little as 25h already yields a WER of ≈ 60, with consistent improvements with more language-specific fine-tuning data. This demonstrates the power of leveraging large-scale multilingual pre-training to enhance lip-reading performance, particularly for low-resource languages, setting the stage for broader and more inclusive applications of visual speech recognition.

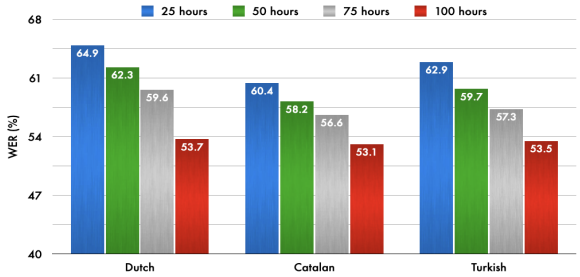


Fig. 4. Fine-tuning the joint model on as little as 25 hours of a new language yields a reasonable performance, which further improves with more data.

E. Monolingual v/s Multilingual

In Table IV, we assess the benefits of multilingual training. The first row contains results from training a separate model for top five non-English languages. As each of these separate models learn from only a small amount of data, it hurts the performance significantly. On the other hand, training a joint model on all languages yields a single model that can identify and transcribe all languages with far lower WERs.

TABLE IV
MONOLINGUAL V/S MULTILINGUAL TRAINING ON *MultiVSR* DATASET.

Method	Portuguese	Spanish	Russian	German	French
Monolingual	39.96	46.59	52.94	57.18	51.78
Multilingual	**33.76**	**37.04**	**39.49**	**40.35**	**34.63**

VI. LIMITATIONS AND CONCLUSION

In this work, we built the largest multilingual lip-reading dataset to date. We train a state-of-the-art lip-reading model that can jointly identify the language and transcribe the utterance. A clear limitation of our dataset is that it has a large composition of English clips. While our multilingual training strategy makes good use of this, a promising future direction would be to increase the number of languages and generalize to new languages in a few-shot manner.

REFERENCES

[1] Y. He, T. N. Sainath, R. Prabhavalkar, I. McGraw, R. Alvarez, D. Zhao, D. Rybach, A. Kannan, Y. Wu, R. Pang *et al.*, "Streaming end-to-end speech recognition for mobile devices," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6381–6385.

[2] J. Li *et al.*, "Recent advances in end-to-end automatic speech recognition," *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, 2022.

[3] R. Prabhavalkar, T. Hori, T. N. Sainath, R. Schlüter, and S. Watanabe, "End-to-end speech recognition: A survey," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[4] N. Manh Tien Anh and T. Ho Sy, "Improving speech recognition with prompt-based contextualized asr and llm-based re-predictor," in *Interspeech*, 2024, pp. 737–741.

[5] Y. Hono, K. Mitsuda, T. Zhao, K. Mitsui, T. Wakatsuki, and K. Sawada, "Integrating pre-trained speech and language models for end-to-end speech recognition," in *Findings of the Association for Computational Linguistics ACL 2024*, 2024, pp. 13 289–13 305.

[6] A. Tjandra, N. Singhal, D. Zhang, O. Kalinli, A. Mohamed, D. Le, and M. L. Seltzer, "Massively multilingual asr on 70 languages: Tokenization, architecture, and generalization capabilities," in *ICASSP*, 2023, pp. 1–5.

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

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

[9] O. Chang, H. Liao, D. Serdyuk, A. Shahy, and O. Siohan, "Conformer is all you need for visual speech recognition," in *ICASSP*. IEEE, 2024, pp. 10 136–10 140.

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

[11] J. H. Yeo, M. Kim, S. Watanabe, and Y. M. Ro, "Visual speech recognition for languages with limited labeled data using automatic labels from whisper," in *ICASSP*, 2024, pp. 10 471–10 475.

[12] T. Afouras, J. S. Chung, and A. Zisserman, "Deep lip reading: a comparison of models and an online application," in *INTERSPEECH*, 2018.

[13] P. Ma, A. Haliassos, A. Fernandez-Lopez, H. Chen, S. Petridis, and M. 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.

[14] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *ICML*, 2023, pp. 28 492–28 518.

[15] M. Bain, J. Huh, T. Han, and A. Zisserman, "Whisperx: Time-accurate speech transcription of long-form audio," in *INTERSPEECH*, 2023.

[16] P. Ma, S. Petridis, and M. Pantic, "Visual speech recognition for multiple languages in the wild," *Nature Machine Intelligence*, vol. 4, no. 11, pp. 930–939, 2022.

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

[18] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic, "End-to-end audiovisual speech recognition," in *ICASSP*. IEEE, 2018, pp. 6548–6552.

[19] X. Zhang, F. Cheng, and S. Wang, "Spatio-temporal fusion based convolutional sequence learning for lip reading," in *Proceedings of the IEEE/CVF International conference on Computer Vision (ICCV)*, 2019, pp. 713–722.

[20] 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 AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 6917–6924.

[21] 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.

[22] P. Ma, S. Petridis, and M. Pantic, "End-to-end audio-visual speech recognition with conformers," in *ICASSP*. IEEE, 2021, pp. 7613–7617.

[23] K. R. Prajwal, T. Afouras, and A. Zisserman, "Sub-word level lip reading with visual attention," in *Proc. CVPR*, 2022.

[24] J. H. Yeo, M. Kim, and Y. M. Ro, "Multi-temporal lip-audio memory for visual speech recognition," in *ICASSP*. IEEE, 2023, pp. 1–5.

[25] K. R. Prajwal, T. Afouras, and A. Zisserman, "Speech recognition models are strong lip-readers," in *INTERSPEECH*, 2024.

[26] 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.

[27] 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.

[28] X. Liu, E. Lakomkin, K. Vougioukas, P. Ma, H. Chen, R. Xie, M. Doulaty, N. Moritz, J. Kolar, S. Petridis *et al.*, "Synthvsr: Scaling up visual speech recognition with synthetic supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 806–18 815.

[29] 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," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 10 391–10 395.

[30] 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.

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

[32] 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.*, vol. 37, 2018.

[33] E. Salesky, M. Wiesner, J. Bremerman, R. Cattoni, M. Negri, M. Turchi, D. W. Oard, and M. Post, "Multilingual tedx corpus for speech recognition and translation," in *INTERSPEECH*, 2021.

[34] 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.

[35] M. Kim, J. H. Yeo, J. Choi, and Y. M. Ro, "Lip reading for low-resource languages by learning and combining general speech knowledge and language-specific knowledge," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 15 359–15 371.

[36] J. L. Newman and S. J. Cox, "Language identification using visual features," *IEEE Transactions on audio, speech, and language processing*, vol. 20, no. 7, pp. 1936–1947, 2012.

[37] T. Afouras, J. S. Chung, and A. Zisserman, "Now you're speaking my language: Visual language identification," in *INTERSPEECH*, 2020.

[38] J. S. Chung and A. Zisserman, "Out of time: automated lip sync in the wild," in *Workshop on Multi-view Lip-reading, ACCV*, 2016.

[39] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, "S3fd: Single shot scale-invariant face detector," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 192–201.

[40] J. Valk and T. Alumäe, "VoxLingua107: a dataset for spoken language recognition," in *Proc. IEEE SLT Workshop*, 2021.

[41] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*. MIT Press, 2014, p. 3104–3112.