ViSpeR: Multilingual Audio-Visual Speech Recognition

Sanath Narayan[1*] Yasser Abdelaziz Dahou Djilali[1*] Ankit Singh[1]
Eustache Le Bihan[2†] Hakim Hacid[1]

[1]Technology Innovation Institute, UAE [2]ENS Paris-Saclay, France

Abstract

This work presents an extensive and detailed study on Audio-Visual Speech Recognition (AVSR) for five widely spoken languages: Chinese, Spanish, English, Arabic, and French. We have collected large-scale datasets for each language except for English, and have engaged in the training of supervised learning models. Our model, ViSpeR, is trained in a multi-lingual setting, resulting in competitive performance on newly established benchmarks for each language. The datasets and models are released to the community with an aim to serve as a foundation for triggering and feeding further research work and exploration on Audio-Visual Speech Recognition, an increasingly important area of research. Code available at https://github.com/YasserdahouML/visper.

1. Introduction

Visual Speech Recognition (VSR), also known as sentence-level VSR [17], poses significant challenges in training deep learning models due to the ambiguous nature of the input data, and the lack of large scale datasets compared to, e.g., the ones in Audio Speech Recognition (ASR) [3, 4]. Indeed, acquiring VSR data involves recording and annotating both the audio and video streams simultaneously, which is a more complex and resource-intensive process compared to acquiring only audio data for ASR. Additionally, factors like lighting conditions, camera angles, and speaker variations can introduce noise and variability in the visual data, making it harder to capture high-quality and consistent VSR data at scale. Furthermore, before extracting the visual features from a video stream, it is necessary to first detect the active speaker in the video and then locate and track the mouth region of the person of interest, which adds further complexity to the data acquisition process.

To this end, not only are the current VSR datasets (e.g., LRS3 [1] and VoxCeleb [12]) smaller in size compared to ASR datasets, they are mostly focused on the English language. This limits the applicability of VSR models to other languages and accents. Indeed, existing non-English VSR datasets [9, 20] are significantly shorter in duration, and often recorded in a controlled environment. The recent work of MuAVIC [2] introduced a multilingual audio-visual corpus, providing $1,200$ hours of audio-visual speech in nine languages extracted from TED talks. Authors of [19] leveraged the existing large-scale, unlabeled multilingual audio-visual speech datasets, such as VoxCeleb2 [12] (2442 hours) and AV-Speech [7] (around $4,700$ hours), and used Whisper [16] to transcribe the segments. This led to the creation of a multi-lingual dataset for four languages, namely: French, Italian, Spanish and Portuguese.

Given the scarcity of publicly available VSR data for non-English languages, the first natural step is to collect such data for the most four spoken languages at scale. To this end, we develop a data pipeline for efficiently collecting and processing videos from the wild. In this work, we collect and process data for Arabic (*ar*), Spanish (*es*), French (*fr*), and Chinese (*zh*).

Contributions: Based on the above discussion, the main contributions of this work are as follows:

- *Data pipeline and dataset*: We develop an efficient data collection pipeline for VSR to obtain 787h, 1200h, 794h, and 872h of data for Chinese, Arabic, Spanish, and French languages, respectively.
- *Benchmarks*: We also carefully create benchmarks for each language so they can be used to further measure progress in this field.
- *ViSpeR*: We engage in the training of supervised VSR and AVSR models, and establish them as baselines on the introduced benchmarks.

2. Building the ViSpeR dataset

Exploiting publicly available online content as a data source for creating audio/visual speech recognition datasets has become a popular approach [1, 7, 11, 12]. Clearly however,

[*]Joint first author; order determined by flipping a coin.
[†]Work done during an internship at TII.
Please send correspondence to: visper@tii.ae

producing a VSR dataset of satisfactory quality necessitates meticulous processing of raw videos to create pairs of visual sequences, specifically visual lip movements matched with text labels. Due to the computational complexity, it is important to carefully filter and extract relevant content from the extensive online repository before proceeding. Therefore, the procedure should involve two primary stages: 1) identifying pertinent videos likely to contain VSR content, and 2) processing the selected videos to form the required data pairs.

In terms of the initial phase, previous studies have employed two distinct methodologies. First, the YTD18 [11] dataset of 30k hours of utterances (i.e. not publicly accessible), there is no explicit mention of their approach to video selection. Another approach utilized, as seen in the creation of the publicly accessible and extensively utilized LRS3 [1], involves constraining content sources to high-quality videos, specifically focusing on TED and TEDx talks. These talks are chosen due to their reliable transcripts and visuals that closely align with the targeted objectives, thereby maximizing the efficiency of a processing pipeline aimed at constructing a VSR dataset. However, this approach significantly reduces the pool of available content from YouTube and may restrict the dataset's capacity to represent real-world performances, as TED talks typically share similar visual contexts, featuring underrepresented individuals and employing a formal language vocabulary. We opt for an approach that combines the strengths of both strategies by focusing on keyword-based searches within a curated set of high-quality video sources. This hybrid approach allows us to expand the dataset's diversity while still ensuring the content's relevance and quality. By targeting videos that are keyword-tagged with specific themes and subjects, we enhance our ability to include a broader array of visual contexts and linguistic styles. This method not only diversifies the visual and linguistic input but also broadens the demographic representation within the dataset. Additionally, this strategy mitigates the limitations imposed by exclusively using TED and TEDx talks, thereby providing a more comprehensive foundation for developing robust and effective audio visual speech recognition models.

In this collection pipeline, and to avoid the complexity of large-scale processing, we aim to ensure that the initially targeted videos are likely to contain VSR content. For this, we train a simple binary classifier that streams the first 100 frames and assigns a score whether the input should be further considered for processing. Thus, we reduce the search to approximately only 20 % of the initial pool. The training set used to create the classifier was built by doing a first pass and then recursively tracking the videos that did not output any clips, followed by labeling the data accordingly.

2.1. Data Gathering

We leverage the capabilities of the YouTube search API, which provides various filtering options: (*i*) Search using keywords, (*ii*) target most relevant videos in regard of a specified language (Here French, Spanish, Arabic and Chinese). Indeed, we initiate the search using a set of 200 keywords, such as ”interview” or ”discussion,” to acquire the most pertinent content, extending the content-oriented filtering methodology explored in [1]. Additionally, we ensure that these videos do not duplicate content from existing multi-lingual datasets (i.e. VoxCeleb2 [12] and AV-Speech [7]) by cross-referencing YouTube IDs. The subsequent videos are then processed as detailed next.

2.2. Data Processing

Each video is divided into multiple shots using a scene change detection [8], which relies on alterations in three-dimensional histograms. Faces within each frame of the video are identified using YOLOv5n0.5-Face [14], chosen for its high accuracy-to-compute cost ratio. Then, the detected faces are matched and tracked across frames to generate multiple face tracks. These tracks are then filtered using SyncNet [5], leveraging active speaker detection to isolate face track segments featuring speakers corresponding to the audio content. Finally, we utilize the ASR model Whisper [15] to detect the language, also obtaining automated-transcripts. Utilizing word timestamps from Whisper outputs, tracks are segmented into clips ranging from 2.0 to 16 seconds in duration.

2.3. ViSpeR Statistics

Training: As shown in Table 1, our proposed dataset surpasses others in scale and coverage. ViSpeR exhibits substantial increases in both the number of clips and the total duration across all languages, making it a comprehensive resource for non-English VSR research.

Test: To ensure fair and robust evaluations, we take additional measures. Firstly, we obtain a second transcription using the Seamless-M4T model [13] for a pool of considered samples to build the test sets. Then, we retain only the clips that match the transcripts generated by Whisper, thus ensuring the creation of high-quality and reliable test sets. Additionally, we curate a subset from both the TedX and Wild splits. This ensures that our evaluation is both thorough and consistent with the LRS3 English TedX benchmark. For English, we use our previously introduced benchmark WildVSR [6] that is challenging and gives an accurate estimate of how existing English VSR models perform in the wild.

Table 1. **Comparison of VSR datasets**. Our proposed ViSpeR dataset is larger in size compared to other datasets that cover non-English languages for the VSR task. For our dataset, the numbers in parenthesis denote the number of clips. We also give the clip coverage under TedX and Wild subsets of our ViSpeR dataset.

Dataset	**French** (*fr*)	**Spanish** (*es*)	**Arabic** (*ar*)	**Chinese** (*zh*)
MuAVIC	176	178	16	–
VoxCeleb2	124	42	–	–
AVSpeech	122	270	–	–
ViSpeR (TedX)	192 (160k)	207 (151k)	49 (48k)	129 (143k)
ViSpeR (Wild)	680 (481k)	587 (383k)	1152 (1.01M)	658 (593k)
ViSpeR (full)	872 (641k)	794 (534k)	1200 (1.06M)	787 (736k)

Table 2. **Test set size per language.** For both TedX and Wild splits, the duration is given in hours. The numbers in parenthesis denote the number of clips.

	TedX	**WildVSR**
French (*fr*)	0.31 (221)	2.01 (1442)
Spanish (*es*)	0.65 (429)	1.21 (828)
Arabic (*ar*)	0.26 (208)	1.19 (745)
Chinese (*zh*)	0.37 (387)	3.30 (2989)

3. Experiments

3.1. Experimental Setup

The processed multilingual VSR video-text pairs are utilized to train a encoder-decoder model in a fully-supervised manner. The encoder-decoder model closely follows the structure of the state-of-the-art AutoAVSR [10]. The models are trained under a multi-lingual setting. While the encoder size is 12 layers, the decoder size is 6 layers. The hidden size, MLP and number of heads are set to 768, 3072 and 12, respectively. The unigram tokenizers are learned for all languages combined and have a vocabulary size of 21k. The models are trained for 150 epochs on 64 Nvidia A100 GPUs (40GB) using AdamW optimizer with max LR of 1e-3 and a weight decay of 0.1. A cosine scheduler with a warm-up of 5 epochs is used for training. The maximum batch size per GPU is set to 1800 video frames.

3.2. Results

Table 3 shows the performance of multilingual VSR and AVSR models for five languages (*fr*, *es*, *ar*, *zh*, and *en*) on our proposed benchmarks. In general, we observe that the model performance on Latin languages is better (lower WER) compared to the non-latin languages (*ar* and *zh*) on both VSR and AVSR tasks. For Arabic, a likely explanation is the diversity of accents and the dynamic spellings for the same words. In addition, since we used Whisper to transcribe the segments, we expect a higher level of label-noise in non-Latin languages given the fact that Whisper performance on these languages isn't on par with the Latin ones. Furthermore, the models perform better on the wild test split, compared to the TedX test split for all languages except English. This can likely be attributed to the following reasons: (*i*) the number of clips in train set belonging to the wild are far higher (4-10x) than those from TedX, and (*ii*) while the wild set generally covers high-quality clips with faces near to the camera, the quality of the clips in TedX (non-English TedX isn't the official organisation and the setting isn't as professional) is lower due to farther camera angles, thereby resulting in sub-optimal decoding of the text from noisy video features.

Table 3. **Performance comparison on different languages.** Here, our multilingual ViSpeR models (VSR and AVSR) are evaluated on the TedX and Wild test splits combined. For *en*, we combine the LRS3 [1] and WildVSR [6] test sets. Both VSR and AVSR models are trained on the full set (TedX+Wild). Performance (lower is better) is reported in terms of WER for *en*, *fr*, *es* and *ar*, while CER is used for *zh*.

	VSR	AVSR
French (*fr*)	29.8	5.7
Spanish (*es*)	39.4	4.4
Arabic (*ar*)	47.8	8.4
Chinese (*zh*)	51.3	15.4
English (*en*)	49.1	8.1

Moreover, these baseline model performances are still on the higher side ($\geq$30 WER) when compared to the state-of-the-art models on English (<20 WER) on LRS3 (*i.e.*, less than 1h of duration). This can be attributed to (*i*) smaller training sets in the proposed ViSpeR (800 to 1200 hours per language), compared to English (1700 to 3450 hours in training) and (*ii*) harder test sets in our ViSpeR dataset ($\geq$1.5 hours per language) compared to English LRS3 test set (0.9 hours). Additionally, the recent work of [6] showed that VSR models drop in performance when tested on the newly introduced English benchmark (WildVSR), which is consistent with our findings.

Furthermore, the results presented in Table 3 illustrates a significant performance disparity between Audio-Visual Speech Recognition (AVSR) and Visual Speech Recognition (VSR) across all evaluated languages (French, Spanish, Arabic, Chinese, and English). This difference can be largely attributed to the integration of audio cues in AVSR, which significantly enhances the model's ability to predict spoken words, even in challenging conditions such as noisy environments or videos with suboptimal visual clarity.

4. Discussion

Ethical Considerations: Given that the data collection for building the ViSpeR dataset utilizes publicly available videos from YouTube, biases inherent on the platform are likely to be present in the collected training and test sets. While steps have been taken to enhance the diversity of the content, the data is likely to be non-uniform and skewed towards certain content types during the filtering. Although the content used to derive the dataset is publicly available, we will provide a mechanism for content creators to opt out of the dataset. If anyone wishes to have their data removed, they can contact us and we will promptly exclude the associated clips and update the dataset.

Future works and why ViSpeR dataset is important:
Self-Supervised Learning methods for VSR: A likely future direction includes training multilingual self-supervised models (similar to AV-HuBERT [18]) on the proposed dataset to create a foundational model for VSR. This will include finding suitable clustering methods for creating the pseudo-labels to account for the multi-lingual aspect.
Multi-lingual supervised models: When training a 'single-model-for-multiple-languages', a few important questions that arise include: What is the optimal vocabulary size of the tokenizer? How to avoid tokens switching when the model mixes languages in the prediction? How to predict the spoken language if not know *a priori*?
VSR translation: Another interesting question involves leveraging the dataset to train models for translating visual speech from one language to another, such as from French to English. This capability holds great potential for facilitating seamless communication across linguistic barriers.
Other applications: Beyond VSR, our dataset could also be used for lip syncing, speaker identification, etc.

Conclusion We proposed a large-scale multilingual dataset called ViSpeR for the task of Audio Visual Speech Recognition. The ViSpeR dataset contains nearly 3.2 million clips with more than 3600 hours duration in total, covering four languages: Chinese, Arabic, Spanish and, French. The clips were filtered from different settings like interviews, talks, *etc.* to ensure sufficient diversity in the dataset. Furthermore, the test set contains two splits (TedX and WildVSR) per language to aid effective evaluation of the trained models. Moreover, we trained multi-lingual baseline models in a fully-supervised manner on the ViSpeR dataset for VSR and AVSR. We observed a reasonable performance on our proposed benchmarks, with clear gap between VSR and AVSR.

References

[1] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Lrs3-ted: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*, 2018. 1, 2, 3

[2] Mohamed Anwar, Bowen Shi, Vedanuj Goswami, Wei-Ning Hsu, Juan Pino, and Changhan Wang. Muavic: A multilingual audio-visual corpus for robust speech recognition and robust speech-to-text translation. *arXiv preprint arXiv:2303.00628*, 2023. 1

[3] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*, 2019. 1

[4] Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, et al. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. *arXiv preprint arXiv:2106.06909*, 2021. 1

[5] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13*, pages 251–263. Springer, 2017. 2

[6] Yasser Abdelaziz Dahou Djilali, Sanath Narayan, Eustache LeBihan, Haithem Boussaid, Ebtesam Almazrouei, and Merouane Debbah. Do vsr models generalize beyond lrs3? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6635–6644, 2024. 2, 3

[7] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619*, 2018. 1, 2

[8] Igor S. Gruzman and Anna S. Kostenkova. Algorithm of scene change detection in a video sequence based on the threedimensional histogram of color images. In *2014 12th International Conference on Actual Problems of Electronics Instrument Engineering (APEIE)*, pages 1–1, 2014. 2

[9] Denis Ivanko, Alexandr Axyonov, Dmitry Ryumin, Alexey Kashevnik, and Alexey Karpov. Rusavic corpus: Russian audio-visual speech in cars. In *Proceedings of the thirteenth language resources and evaluation conference*, pages 1555–1559, 2022. 1

[10] 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)*, pages 1–5. IEEE, 2023. 3

[11] Takaki Makino, Hank Liao, Yannis Assael, Brendan Shillingford, Basilio Garcia, Otavio Braga, and Olivier Siohan. Recurrent neural network transducer for audio-visual speech recognition. In *2019 IEEE automatic speech recognition and understanding workshop (ASRU)*, pages 905–912. IEEE, 2019. 1, 2

[12] Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman. Voxceleb: Large-scale speaker verification in the wild. *Computer Speech & Language*, 60:101027, 2020. 1, 2

[13] Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. Scaling speech technology to 1,000+ languages. *arXiv preprint arXiv:2305.13516*, 2023. 2

[14] Delong Qi, Weijun Tan, Qi Yao, and Jingfeng Liu. Yolo5face: Why reinventing a face detector. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*, pages 228–244. Springer, 2023. 2

[15] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*, 2022. 2

[16] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023. 1

[17] Changchong Sheng, Gangyao Kuang, Liang Bai, Chenping Hou, Yulan Guo, Xin Xu, Matti Pietikäinen, and Li Liu. Deep Learning for Visual Speech Analysis: A Survey. 2022. 1

[18] 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. 4

[19] Jeong Hun Yeo, Minsu Kim, Shinji Watanabe, and Yong Man Ro. Visual speech recognition for languages with limited labeled data using automatic labels from whisper. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10471–10475. IEEE, 2024. 1

[20] Amir Zadeh, Yan Sheng Cao, Simon Hessner, Paul Pu Liang, Soujanya Poria, and Louis-Philippe Morency. Cmu-moseas: A multimodal language dataset for spanish, portuguese, german and french. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, page 1801. NIH Public Access, 2020. 1