AudioVSR: Enhancing Video Speech Recognition with Audio Data

Xiaoda Yang[1]*, Xize Cheng[1]*, Jiaqi Duan[3], Hongshun Qiu[2],
Minjie Hong[1], Minghui Fang[1], Shengpeng Ji[1], Jialung Zuo [1],
Zhiqing Hong[1], Zhimeng Zhang[1], Tao Jin[1]†,

[1]Zhejiang University, [2]Beijing University Of Technology, [3]Qingdao University
Correspondence: xiaodayang@zju.edu.cn

Abstract

Visual Speech Recognition (VSR) aims to predict spoken content by analyzing lip movements in videos. Recently reported state-of-the-art results in VSR often rely on increasingly large amounts of video data, while the publicly available transcribed video datasets are insufficient compared to the audio data. To further enhance the VSR model using the audio data, we employed a generative model for data inflation, integrating the synthetic data with the authentic visual data. Essentially, the generative model incorporates another insight, which enhances the capabilities of the recognition model. For the cross-language issue, previous work has shown poor performance with non-Indo-European languages. We trained a multi-language-family modal fusion model, AudioVSR. Leveraging the concept of modal transfer, we achieved significant results in downstream VSR tasks under conditions of data scarcity. To the best of our knowledge, AudioVSR represents the first work on cross-language-family audio-lip alignment, achieving a new SOTA in the cross-language scenario.

1 Introduction

Visual speech recognition (VSR), also known as lip reading, aims to recognize the content of speech based on lip movements, without relying on the audio stream. Previous work primarily utilizes video information to build VSR models. However, for any language, the amount of recorded audio information far exceeds that of video information. This raises the question: *Can these audio resources be leveraged to enhance the capabilities of VSR models?* There are two methods to leverage audio data to build a strong VSR model: combining synthetic visual data with authentic visual data and leveraging the alignment of the audio and video.

For the method of combining synthetic visual data with the authentic visual data, previous research (Azizi et al., 2023; Wang et al., 2024) has successfully used generated data to enhance image classification and contrastive learning capabilities. However, the effectiveness of this method varies from problem to problem and lacks discussion regarding the lip-reading task. Does the chicken-and-egg problem apply to the task of VSR? SyncVSR (Liu et al., 2023) applies synthetic data to VSR tasks but does not provide an in-depth discussion of mixing ratios and cross-language scenarios. Our work settles these problems and validates the data inflation method in different scenarios. The effectiveness and robustness of lip-reading can be further enhanced with data inflation. This is essentially because the addition of new data increases the coupling of the dataset, which leads to better-learned knowledge. Based on this, we enrich the dataset using talking head generation (TFG) models. We compare multiple TFG models and find that using the TFG model for data inflation needs to be combined with the original data. Furthermore, since using generated data for data inflation does not always work (Wang et al., 2024; Ji et al., 2024e), we give a relatively stable mixing ratio. Our method achieves new SOTA results on zero-shot and full-shot VSR tasks.

For the method of leveraging the alignment of the audio and video, previous work (Han et al., 2024) has conducted extensive research on Indo-European languages, proposing a cross-language model within the language family. However, there is a weakness when crossing language families, as there are significant linguistic differences between the language families. Based on this, we propose a language-independent pre-training model for the two major language families, and use this model to accomplish the downstream VSR task. Specifically, for any language, its lips and sounds always correspond to each other, so we use multilingual audio and video data to pre-train the audio-lip alignment model. Considering that the audio data for any language is much larger than its video data (Kim

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 15352–15361
November 12-16, 2024 ©2024 Association for Computational Linguistics

et al., 2021; Ren et al., 2021a; Zhao et al., 2020a; Ji et al., 2024b,a), we transfer the knowledge of audio-lip alignment gained in the pre-training phase to the downstream task, using audio to augment the capabilities of the VSR model.

Our contributions are highlighted as follows.

- We investigate the significance of synthetic data for enhancing the capability of VSR modeling and found that the ratio of synthetic data, the generation model, and the quality of audio data have an impact on the VSR model. We give the optimal mixing ratio and analyze the mechanisms behind data inflation.

- We are the first to consider the impact of language families on Visual Speech Recognition (VSR) tasks. We utilized self-supervised learning to train a cross-language-family modal fusion model, AudioVSR.

- We provide solutions for challenges such as insufficient video data and cross-language-family problem, which enhance the robustness of the VSR model. Our method achieves state-of-the-art (SOTA) results in zero-shot, full-shot, and cross-language scenarios.

2 Related work

2.1 Visual Speech Recognition

Visual Speech Recognition (VSR) aims to predict the spoken words through the analysis of lip movements in video. Early works (Chung and Zisserman, 2017; Petridis et al., 2017; Stafylakis and Tzimiropoulos, 2017) focused on word-level VSR by using CNN and the RNN. Large-scale lip-reading sentence datasets (Afouras et al., 2018) have boosted the development of sentence-level VSR. By employing Transformer (Vaswani et al., 2017) architecture, (Afouras et al., 2022a) proposed a powerful sentence-level end-to-end VSR model. Recent VSR technologies (Chang et al., 2023) also employed transformer-variant architectures and improved the VSR performances. In pursuit of advanced training strategies, many researchers focus on narrowing the gap between visual and audio modalities. They (Ma et al., 2021; Ren et al., 2021b) studied how to effectively transfer audio knowledge into the VSR model by using knowledge distillation (Hinton et al., 2015) and memory network (Becattini and Uricchio, 2022).

2.2 Data Inflation

Traditional data augmentation involves simple processing based on the original data, for image data augmentation includes rotation, translation, scaling, etc., for audio data augmentation includes adding noise, variable speed, etc.. Traditional data augmentation can increase model generalization, but for the downstream task of VSR, traditional data augmentation does not introduce new valid information, so we take a generative model to expand the dataset, which is known as data inflation (Wang et al., 2024). This approach not only increases the amount of data but also introduces new variants that are consistent with real-world distribution.

Enriching datasets with generative models (Fang et al., 2024; Ji et al., 2024c,d) has been widely explored in the field of visual problems. (Azizi et al., 2023) explored this approach on a visual categorization task, and (Wang et al., 2024) combined generated data with traditional data inflation to explore its influence on contrastive learning (Yang et al., 2024). Building on previous studies (Wang et al., 2024), which indicate that data inflation does not always work, we conducted further research on its influencing factors and the appropriate ratio.

The main TFG generative models include (Prajwal et al., 2020; Wang et al., 2023; Zhou et al., 2021, 2020). However, (Zhou et al., 2021, 2020) focus on specific characteristics like audio and 3D, which do not meet our requirements for the downstream VSR task. (Wang et al., 2023) uses a lip supervisor, but since the model only focuses on region of interest (ROI), too many loss functions result in insufficient attention to non-lip regions, which leads to abrupt edges in ROI regions. So we focus on Wav2lip (Prajwal et al., 2020), and its derived model Wav2lip-gan.

2.3 Audio-Video Information Interaction

AV-HuBERT (Shi et al., 2022) extends HuBERT (Hsu et al., 2021) to the audio-visual setting by taking the masked audio-visual stream as input and predicting the hidden units initialized with MFCC clusters, iteratively refining them with layerwise features. The framework has proven effective for multiple downstream tasks, including ASR, VSR and AVSR. Our work adopts their modal random dropout technique. AV-Former (Seo et al., 2023) integrates lightweight modules into an audio-only speech recognizer, utilizing visual information to improve speech recognition capabilities. (Hsu and

Shi, 2022) utilizes the audio information to enhance the English VSR model. Recently, (Han et al., 2024) proposed a multilingual AVSR model based on the MuAViC dataset (Anwar et al., 2023), which includes a total of 9 languages. Inspired by these works, we align the audio and video and train a cross-language-family VSR model.

3 Method

To achieve an effective VSR model in both zero-shot and full-shot scenarios, we rethink the essence of deep learning in Sec. 3.1. Based on these insights, we employ a talking head generation (TFG) model for data inflation in Sec. 3.2. For the cross-language challenge, aiming for language independence, we focus on universal applicability across language families in Sec. 3.3. Consequently, we utilize self-supervised learning and introduce an audio-lip-alignment model in Sec. 3.4.

3.1 Rethinking Deep Learning

The goal of deep learning is to accurately perceive the real world, and its essence is to learn the distribution of the real world. Utilizing a dataset for training is based on the assumption that the dataset can simulate the distribution of the real world. But in practice, datasets often cannot perfectly simulate the objective world, and the gap between them represents our space for optimization. Typically, a better model leverages the current data distribution more effectively. Our work, however, aims to alter this data distribution to better reflect the distribution of the objective world.

In the full-shot scenario, the greater the data coupling, the more likely it is to achieve good results on the test set. However, does achieving good results on the test set truly indicate a good model? This depends on a key assumption: the chosen dataset must accurately simulate the distribution of the target language domain. In contrast, a good outcome achieved in the zero-shot scenario is more competitive compared to the full-shot scenario, as it faces the dual challenges of architecture and dataset. For the architecture, we choose the AudioVSR model, which has achieved good performance on VSR tasks; for the dataset, we use an audio-to-lip model for data inflation.

3.2 Data Inflation through TFG model

Based on a sequence model, AV-HuBERT highlights the importance of temporal correlations in video data for VSR tasks. Audio inherently carries temporal information, so by transferring audio knowledge to the corresponding video, the audio-to-lip model generates a video stream that contains the same temporal information. The model significantly enhances the semantic richness of video content, because it utilizes the temporal information in audio to strengthen the continuity among video frames. Furthermore, the higher accuracy rates of both AVSR and ASR compared to VSR further demonstrate that audio contains a wealth of knowledge. The audio-to-lip model serves as a bridge for transferring the audio knowledge to the VSR model.

Denotes M as the tokens of the real videos, M' as the tokens of the generated videos, the rank of the M is less than the rank of the $concate(M, M')$, where the rank of a matrix is the size of the largest linearly uncorrelated set of matrix vectors. It demonstrates that the TFG models can enrich the semantic information of the input videos.

Denote the distribution of real data as P_d, the distribution of generative data as P_g, the real dataset as D_d, the generative dataset as D_g, then the distribution of the synthetic data can be denoted as $P_t = \beta \cdot P_d + (1-\beta) \cdot P_g$, where $\beta = |D_d|/(|D_d| + |D_g|)$. Initially, $D_d : D_g = \beta : (1-\beta)$, we repeat D_d N times to modify the value of β, i.e., $\beta = N * |D_d|/(N * |D_d| + |D_g|)$. Repeating real data D_d does not alter the distribution P_d. Factors affecting P_t include the proportion of generated data β and the distribution of generated data P_g. During this process, as the generated data remains unchanged, the distribution P_g remains constant. Only its influence on the overall data distribution, β, changes, thus achieving the goal of controlling variables.

3.3 Rethinking Cross-language VSR

The differences between language families are significant, and their phonetic and semantic capabilities vary. The Indo-European and Sino-Tibetan language families are two of the world's major language families. Most current research focuses on the Indo-European. However, research on other language families is lacking. Recently, (Han et al., 2024) introduced a multilingual model based on the MuAViC dataset (Anwar et al., 2023), which encompasses nine languages: English, Arabic, German, Greek, Spanish, French, Italian, Portuguese and Russian. However, it has been noted that Arabic performs noticeably worse compared to other languages. This discrepancy is due to the fact that the other eight languages in the dataset are from the Indo-European family, whereas Arabic belongs to the Semitic family. During training, Arabic data only constituted 1% of the dataset and did not receive any additional reweighting. As a result, the model is essentially an Indo-European model, offering linguistic universality but not universality across language families. Based on this, we propose AudioVSR, a self-supervised model designed to be universally applicable to the two primary language families.

3.4 AudioVSR

Regardless of the language, there is always a correspondence between sound and lip movements. Based on this, different languages will benefit from each other rather than hinder one another, so we conduct training across multiple languages at the same time. In the pre-training phase, modality dropout involves randomly masking f_t^a and f_t^v with the goal of obtaining the same z_t, fostering a closer relationship and a stronger bond between audio and video. Based on this, we propose AudioVSR, a unified model that maps the audio and video of all languages into the same space. With the AudioVSR model, we input audio from entirely new datasets and fine-tune on downstream tasks, developing a model suitable for VSR tasks. The pipeline is shown in Fig. 1. During the pre-training stage, all modules are tunable. After the audio and video are aligned, in the fine-tuning stage, the encoder is frozen while the decoder remains tunable. Notably, during the fine-tuning stage, only audio is used as input, whereas in the inference stage, video can be used as input.

The AudioVSR model randomly masks parts of the audio and video data, forcing the model to learn deeper semantic connections. It binds audio and video semantically, ensuring functionality across diverse languages, which implies that the model captures a universal semantic framework beyond mere linguistic elements. As a result, whether audio and video are input separately or together, the embeddings produced are similar.

Pre-train The encoder is pre-trained using a self-supervised method, which is the same as (Hsu et al., 2021). Two steps are alternated during pre-training: feature clustering and mask prediction. The clustering phase uses a discrete latent variable model to form $\{z_t\}_{t=1}^T$, and the model then performs mask prediction through the Transformer (Vaswani et al., 2017) architecture to learn a better representation of the audio and video in the semantic space $f^m = \{f_t^m\}_{t=1}^T \in \mathbb{R}^{T\times D}$, where T is the sequence length, D is the embedding dimension, and m means modality. Given the output probability $\{p_t\}_{t=1}^T$, The pre-training loss is:

$$M = M_a \cup M_v \tag{1}$$

$$L = -\sum_{t\in M} \log p_t(z_t) - \alpha \sum_{t\notin M} \log p_t(z_t) \tag{2}$$

where M^a and M^v denote the frames that are masked for the audio and video. α controls the contribution of the unmasked regions in the overall objective.

Fine-tune Denote the features processed by the encoder as $f^v = \{f_t^v\}_{t=1}^T \in \mathbb{R}^{T\times D}$. A tunable transformer decoder is appended to autoregressively decode f^v into probabilities $p(\omega_t|\{\omega_i\}_{i=1}^{t-1}, f^v)$, where ω_i is the ground-truth transcription.

The lip-reading loss is a sequence-to-sequence loss, which is calculated after the decoder module using cross-entropy loss. Define S as the length of the target text, and the lip-reading loss can be expressed as:

$$L_{lip} = -\sum_{t=1}^{S} \log p(w_t|\{w_i\}_{i=1}^{t-1}, f^v). \tag{3}$$

4 Exprienment

4.1 Dataset

The LRS3 dataset (Afouras et al., 2018) consists of obtained from TED talks covering a large number of speakers and background noise environments. These videos contain sentences in the English language, which is beneficial for lip-reading and visual speech recognition research. The LRS3 dataset is widely used in research on lip-reading techniques due to its diversity and size. The LRS2 dataset (Afouras et al., 2022b) is from BBC, which consists of video, audio, and text for each sample, where the sample rates are 16 kHz for audio and 25 fps for video. The dataset contains more than 1,000 speakers, nearly 150,000 utterance instances, and nearly 63,000 different words, which makes the dataset extremely rich in data. The LRS2 dataset is particularly suitable for studying how to perform effective lip-reading in long video sequences, as it covers

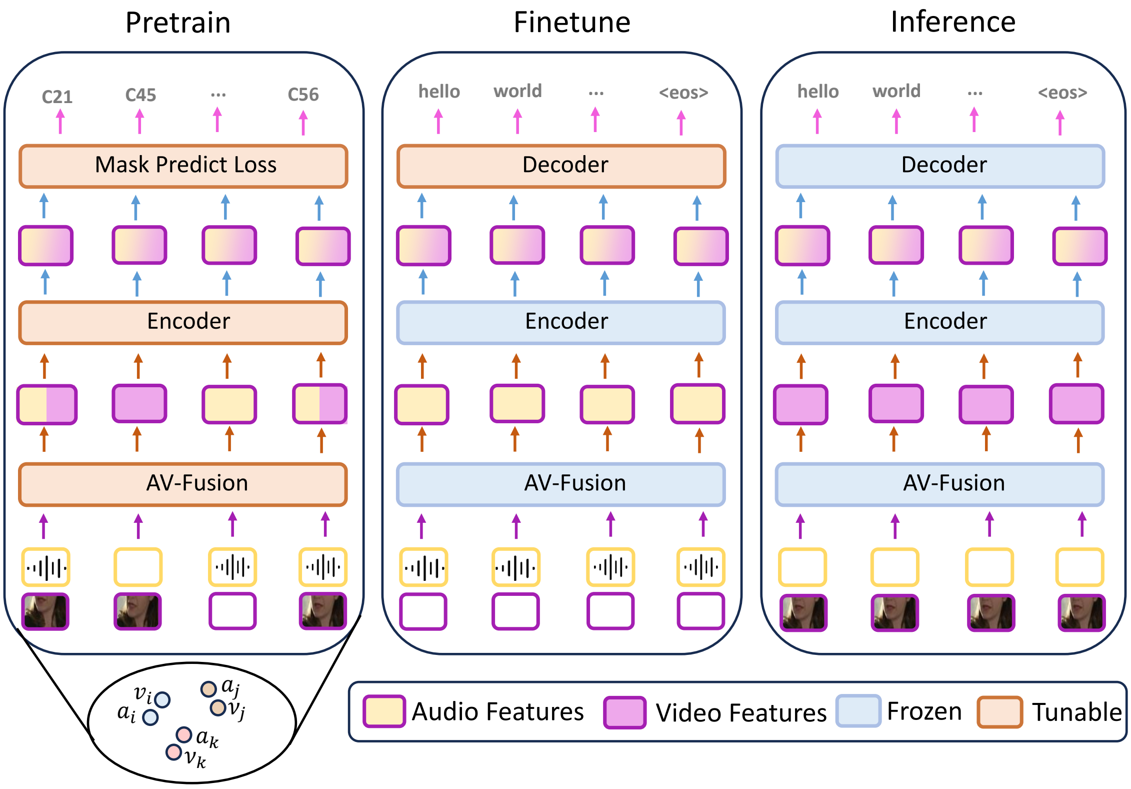

Figure 1: The pipeline of the AudioVSR. During the pre-training phase, random dropout is employed to align audio and video. In the fine-tuning stage, the encoder is frozen to maintain this alignment and prevent corruption, and it will be unfrozen before the end of the process. Throughout the fine-tuning process, audio is used as the primary input. During inference, the model uses video input to generate labels.

continuous natural conversations. The VoxCeleb2 dataset (Nagrani et al., 2017) contains over 1 million video clips from YouTube videos. These clips are from over 6,000 celebrities and cover multiple languages, accents, and background noise conditions. VoxCeleb2 is often used to train cross-modal recognition systems, especially those that combine visual and audio information. We only use the English portion of the VoxCeleb2. The CMLR (Zhao et al., 2019, 2020b) dataset consists of 102, 072 spoken sentences and each sentence is up to 29 Chinese characters in length.

4.2 Metrics

The word error rate (WER) is used as the evaluation index of speech recognition, which is defined as $WER = (S + K + I)/N$, where S denotes the number of words replaced, K denotes the number of words deleted, I denotes the number of words inserted, and N denotes the total number of words in the reference text.

4.3 Data Inflation through TFG model

Full-Shot Scenario In the full-shot scenario, both the training and testing datasets are in the same domain. Consider a scenario where a large amount of video and audio data is available for a certain language. How can we improve lip-reading effectiveness without introducing new data? To solve this problem, an audio-to-lip model is used to generate the videos. These videos would then be integrated with the original footage for combined training efforts.

As shown in the Tab. 4, the LRS3 video scores 28.6, while the video generated from LRS3 audio scores 32.1. However, combining them results in a performance that outperforms both. What is the reason for this enhanced performance? This is because the training of the generative model represents a process of acquiring knowledge from an entirely new perspective, introducing fresh insights. It's similar to two people with distinct ways of thinking studying the same complex topic. Their research perspectives and focal points differ, and although they previously had an incomplete understanding

Table 1: The main results for the zero-shot scenario. The model is trained on LRS3, and tested on LRS2. The TFG model utilized in the table is Wav2Lip. We tested the effectiveness of raw data, generative data, and their combination. To further demonstrate the capability of the method, we also conducted tests on a large data scale.

Experiment	**Method**	**Authentic data (hrs)**	**Sythetic data (hrs)**	**WER(%)↓**
		Low-Resource		
I	AudioVSR	30	-	43.3
II		-	29	109.2
III		30	29	**40.8**
		High-Resource		
IV	AudioVSR	433	-	38.9
V		-	224	108.4
VI		433	224	**37.6**

of the matter, their collaborative discussions spark a collision of thoughts, resulting in an enhancement of knowledge.

Zero-Shot Scenario The traditional lip-reading task trains with video and infers with video. However, in scenarios where video information is lacking, this traditional method faces limitations. Consider this scenario: In a language setting where audio information is abundant but lacks video resources necessary for lip-reading, the TFG model is utilized to synthesize video data. Remarkably, the model is trained exclusively on audio inputs, making this a zero-shot task, as no video data in the target domain is involved in the training phase.

The (Shi et al., 2022) introduces a model that leverages audio information to enhance recognition capabilities. As shown in Tab. 1, our approach achieves performance on par with (Shi et al., 2022) in Visual Speech Recognition (VSR) and surpasses it in Audio-Visual Speech Recognition (AVSR).

Table 2: The comparison with Opensr. Our approach was compared with Opensr in Visual Speech Recognition (VSR) and Audio-Visual Speech Recognition (AVSR) tasks.

Method	**Training**		**Inference**		**WER**
	a	v	a	v	
Opensr	✓			✓	39.2
Opensr	✓		✓	✓	6.3
Our Method	✓			✓	39.8
Our Method	✓		✓	✓	5.4

Table 3: The influence of the model. On the same dataset, different models produce different results.

Dataset	TFG Model	WER
TFG: LRS2	Wav2Lip-Gan	95.0
LRS3+TFG: LRS2	Wav2Lip-Gan	41.4
TFG: LRS2	Wav2Lip	108.2
LRS3+TFG: LRS2	Wav2Lip	40.8

The Effectiveness of TFG Data Inflation To verify the effectiveness of data inflation using the audio-to-lip model, we conducted tests on two datasets, LRS2 and LRS3, at two different scales, as detailed in Tab. 1. From experiments II and V, we can find that using only generated data for lip-reading tasks yielded poor results, whereas combining synthetic with real data exceeded the performance of using real data alone. When tested within the LRS3 domain—aligned with the training environment, the outcomes were favorable, whereas a significant drop in performance in the LRS2 domain, which demonstrates that even in the same language, knowledge from different datasets is not universally applicable. Experiments I and III reveal that leveraging only the 29-hour audio data from LRS2 with the aid of the audio-to-lip generative model, along with the LRS3 dataset, resulted in a 2.3% improvement in lip-reading tasks on LRS2 (29-hour) videos. Furthermore, by comparing Exp. V, VI, we find that data inflation is also effective on large-scale datasets. As shown in Tab. 2, our result is on par with Opensr, which is trained on the English dataset.

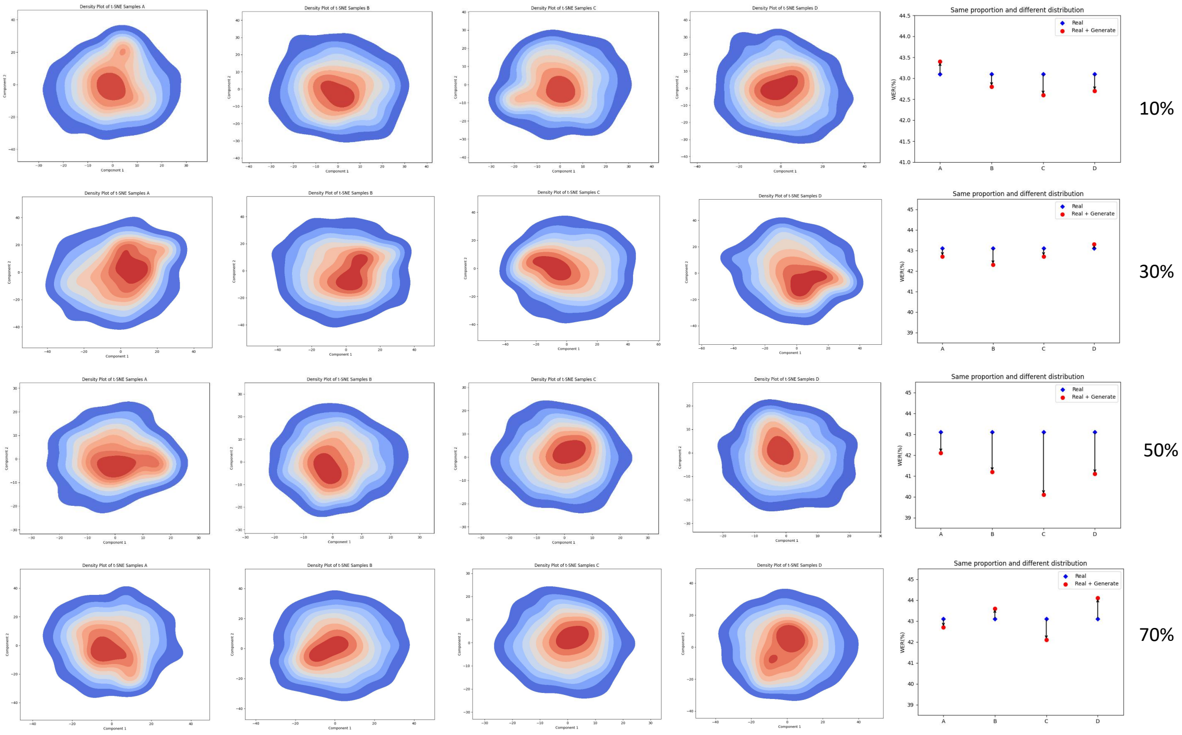

Figure 2: The distribution and effectiveness of different proportions of generated data. We tested different proportions of generative data, specifically 10%, 30%, 50%, and 70%. For each proportion, We randomly selected features four times and visualized their distribution using T-SNE, as demonstrated in A-D in the left picture. On the right, the performance improves compared to the original in all four samplings when the proportion is 50%, which demonstrates its stability.

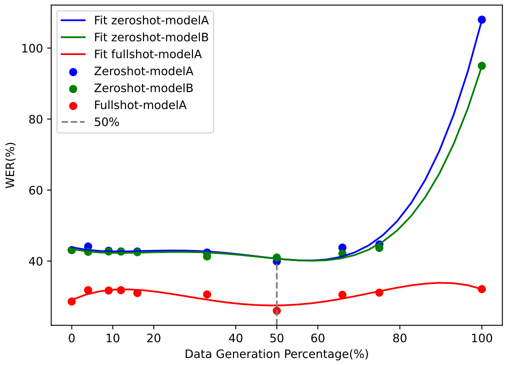

Figure 3: The influence of the proportion of the generated data. We explore various data proportions by choosing N=0, 1/3, 1/2, 1, 2, 5, 7, 10, 20, ∞ and fitting a curve to these values. It is determined that a β value of 0.5 is optimal for both zero-shot and full-shot scenarios.

Factors affecting the TFG data Inflation The performance of the model is influenced by the proportion of generated data β, data quality Q, and the type of model S, expressed as: $WER = F(\beta, Q, S)$. The data distribution P_t varies by model and task, which is proved by Tab. 3 and (Wang et al., 2024). For the TFG model applied to VSR tasks, β is typically around 0.5, and classic TFG models like Wav2lip can be used as S. As demonstrated by the fitting curve in Fig. 3, both models achieve better lip-reading outcomes when β is approximately 0.5, compared to the results obtained from the original dataset. Moreover, different sampling strategies affect the data quality Q. In Fig. 2, we examine multiple values of β, illustrating how different sampling strategies affect the resulting distributions, with β around 50% being relatively stable. In order to control variables, we keep the proportion of generated data(β) unchanged and only modify its distribution by employing random sampling, which will change the data quality Q. In practical applications, heuristic algorithms can optimize performance, but each step involves a complete training and inference process, making it time-consuming.

4.4 Audio-visual Alignment

Consider a situation where there is a lack of video information for the target language, making it impossible to train a VSR model using conven-

Table 4: The main results for the full-shot scenario. There is a phenomenon where the effects of mixed data outweigh the effects of raw data as well as generated data.

Exp	Method	Authentic Data (hrs)	Synthetic Data (hrs)	WER(%)↓
I	RNN-T AVSR (Makino et al., 2019)	31000	-	33.6
II	VSR-MLW (Ma et al., 2022)	1459	-	31.5
III	VSD (Prajwal et al., 2022)	2676	-	30.7
IV	AV-HuBERT (Seo et al., 2023)	433	-	26.9
V	AudioVSR	433	-	28.6
VI		-	433	32.1
VII		433	433	**26.0**

Table 5: The main results of the AudioVSR Exp. I and Exp. II are pre-trained on the Indo-European family of languages, but Exp. III and Exp. IV are pre-trained on the Sino-Tibetan language family.

Exp	**pre-train**	**Finetune**	**Inference**	**WER(%)↓**
I	LRS3 + VoxCeleb2	LRS2 Audio	LRS2 Video	25.0
II	LRS3 + VoxCeleb2	CMLR Audio	CMLR Video	132.9
III	CMLR + LRS3 + VoxCeleb2	LRS2 Audio	LRS2 Video	25.6
IV	CMLR + LRS3 + VoxCeleb2	CMLR Audio	CMLR Video	50.6

tional methods. In addition to the data inflation techniques mentioned in Sec. 3.2, we propose a language-independent pre-trained model, AudioVSR. We pre-train AudioVSR on many different datasets, and the results are shown in the Tab. 5. The dataset we use for the Sino-Tibetan languages contains only 61 hours of data, while the Indo-European dataset exceeds 1700 hours. To balance the representation between the Sino-Tibetan and Indo-European languages, we replicate the Sino-Tibetan data 30 times to equalize their weights. We pre-trained our model on an Indo-European dataset for five iterations, and then alternately trained on the Indo-European dataset and the processed Sino-Tibetan dataset for six iterations.

As shown in Tab. 5, we conducted four experiments to demonstrate the capabilities of AudioVSR. Comparing Exp. I and Exp. II, it was demonstrated that there are significant differences in lip shapes among different languages, highlighting the necessity to train using multiple language datasets simultaneously. This approach is based on the universal alignment of audio and video across many different languages. Comparing Exp. I and Exp. III, we observe that our model maintains good performance within the English domain. Additionally, comparing Exp. II and Exp. IV, there is a noticeable enhancement in lip-reading effectiveness on the CMLR dataset.

5 Limitation

For zero-shot and full-shot scenarios, there is a possibility that the model may not perform effectively. For cross-language scenarios, the results of AudioVSR still need improvement in the Sino-Tibetan setting compared to the Indo-European setting. Furthermore, our model only includes the two major language families and is not universally applicable to all language families. In the future, we will explore the effectiveness of our method in noisy scenarios.

6 Conclusion

We enhanced the performance of VSR across multiple scenarios. In zero-shot and full-shot situations, we employed an audio-to-lip model for data inflation, leveraging the knowledge learned from the generative model to enhance the lip-reading model. For cross-linguistic scenarios, we considered the impact of different language families on lip-reading tasks and trained an audio-lip-alignment model using self-supervised learning. We provided solutions for challenges such as insufficient video data and cross-linguistic scenarios. Our method achieved state-of-the-art (SOTA) results in zero-shot, full-shot, and cross-language tasks.

References

Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. 2022a. Deep audio-visual speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 8717–8727.

Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. 2022b. Deep audio-visual speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 8717–8727.

Triantafyllos Afouras, JoonSon Chung, and Andrew Zisserman. 2018. Lrs3-ted: a large-scale dataset for visual speech recognition. *arXiv: Computer Vision and Pattern Recognition,arXiv: Computer Vision and Pattern Recognition*.

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

Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J Fleet. 2023. Synthetic data from diffusion models improves imagenet classification. *arXiv preprint arXiv:2304.08466*.

Federico Becattini and Tiberio Uricchio. 2022. Memory networks. In *Proceedings of the 30th ACM International Conference on Multimedia*.

Oscar Chang, Hank Liao, Dmitriy Serdyuk, Ankit Shah, and Olivier Siohan. 2023. Conformers are all you need for visual speech recogntion.

Joon Son Chung and Andrew Zisserman. 2017. *Lip Reading in the Wild*, page 87–103.

Minghui Fang, Shengpeng Ji, Jialong Zuo, Hai Huang, Yan Xia, Jieming Zhu, Xize Cheng, Xiaoda Yang, Wenrui Liu, Gang Wang, Zhenhua Dong, and Zhou Zhao. 2024. Ace: A generative cross-modal retrieval framework with coarse-to-fine semantic modeling. *Preprint*, arXiv:2406.17507.

HyoJung Han, Mohamed Anwar, Juan Pino, Wei-Ning Hsu, Marine Carpuat, Bowen Shi, and Changhan Wang. 2024. Xlavs-r: Cross-lingual audio-visual speech representation learning for noise-robust speech perception. *arXiv preprint arXiv:2403.14402*.

GeoffreyE. Hinton, Oriol Vinyals, and J.Michael Dean. 2015. Distilling the knowledge in a neural network. *arXiv: Machine Learning,arXiv: Machine Learning*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Wei-Ning Hsu and Bowen Shi. 2022. u-hubert: Unified mixed-modal speech pretraining and zero-shot transfer to unlabeled modality. *Advances in Neural Information Processing Systems*, 35:21157–21170.

Shengpeng Ji, Minghui Fang, Ziyue Jiang, Rongjie Huang, Jialung Zuo, Shulei Wang, and Zhou Zhao. 2024a. Language-codec: Reducing the gaps between discrete codec representation and speech language models. *arXiv preprint arXiv:2402.12208*.

Shengpeng Ji, Ziyue Jiang, Xize Cheng, Yifu Chen, Minghui Fang, Jialong Zuo, Qian Yang, Ruiqi Li, Ziang Zhang, Xiaoda Yang, et al. 2024b. Wavtokenizer: an efficient acoustic discrete codec tokenizer for audio language modeling. *arXiv preprint arXiv:2408.16532*.

Shengpeng Ji, Ziyue Jiang, Hanting Wang, Jialong Zuo, and Zhou Zhao. 2024c. Mobilespeech: A fast and high-fidelity framework for mobile zero-shot text-to-speech. *Preprint*, arXiv:2402.09378.

Shengpeng Ji, Jialong Zuo, Minghui Fang, Ziyue Jiang, Feiyang Chen, Xinyu Duan, Baoxing Huai, and Zhou Zhao. 2024d. Textrolspeech: A text style control speech corpus with codec language text-to-speech models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10301–10305. IEEE.

Shengpeng Ji, Jialong Zuo, Minghui Fang, Siqi Zheng, Qian Chen, Wen Wang, Ziyue Jiang, Hai Huang, Xize Cheng, Rongjie Huang, and Zhou Zhao. 2024e. Controlspeech: Towards simultaneous zero-shot speaker cloning and zero-shot language style control with decoupled codec. *Preprint*, arXiv:2406.01205.

Minsu Kim, Joanna Hong, Se Jin Park, and Yong Man Ro. 2021. Cromm-vsr: Cross-modal memory augmented visual speech recognition. *IEEE Transactions on Multimedia*, 24:4342–4355.

Xubo Liu, Egor Lakomkin, Konstantinos Vougioukas, Pingchuan Ma, Honglie Chen, Ruiming Xie, Morrie Doulaty, Niko Moritz, Jachym Kolar, Stavros Petridis, et al. 2023. Synthvsr: Scaling up visual speech recognition with synthetic supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18806–18815.

Pingchuan Ma, Brais Martinez, Stavros Petridis, and Maja Pantic. 2021. Towards practical lipreading with distilled and efficient models. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Pingchuan Ma, Stavros Petridis, and Maja Pantic. 2022. Visual speech recognition for multiple languages in the wild. *Nature Machine Intelligence*, 4(11):930–939.

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

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*.

Stavros Petridis, Zuwei Li, and Maja Pantic. 2017. End-to-end visual speech recognition with lstms. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

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

KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. 2020. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, pages 484–492.

Sucheng Ren, Yong Du, Jianming Lv, Guoqiang Han, and Shengfeng He. 2021a. Learning from the master: Distilling cross-modal advanced knowledge for lip reading. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13325–13333.

Sucheng Ren, Yong Du, Jianming Lv, Guoqiang Han, and Shengfeng He. 2021b. Learning from the master: Distilling cross-modal advanced knowledge for lip reading. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Paul Hongsuck Seo, Arsha Nagrani, and Cordelia Schmid. 2023. Avformer: Injecting vision into frozen speech models for zero-shot av-asr. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22922–22931.

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

Themos Stafylakis and Georgios Tzimiropoulos. 2017. Combining residual networks with lstms for lipreading. In *Interspeech 2017*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, AidanN. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Neural Information Processing Systems,Neural Information Processing Systems*.

Jiadong Wang, Xinyuan Qian, Malu Zhang, Robby T Tan, and Haizhou Li. 2023. Seeing what you said: Talking face generation guided by a lip reading expert. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14653–14662.

Yifei Wang, Jizhe Zhang, and Yisen Wang. 2024. Do generated data always help contrastive learning? *arXiv preprint arXiv:2403.12448*.

Xiaoda Yang, Xize Cheng, Dongjie Fu, Minghui Fang, Jialung Zuo, Shengpeng Ji, Tao Jin, and Zhou Zhao. 2024. Synctalklip: Highly synchronized lip-readable speaker generation with multi-task learning. In *ACM Multimedia 2024*.

Ya Zhao, Rui Xu, and Mingli Song. 2019. 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.

Ya Zhao, Rui Xu, Xinchao Wang, Peng Hou, Haihong Tang, and Mingli Song. 2020a. Hearing lips: Improving lip reading by distilling speech recognizers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6917–6924.

Ya Zhao, Rui Xu, Xinchao Wang, Peng Hou, Haihong Tang, and Mingli Song. 2020b. Hearing lips: Improving lip reading by distilling speech recognizers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6917–6924.

Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. 2021. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4176–4186.

Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. 2020. Makelttalk. *ACM Transactions on Graphics*, page 1–15.