Uni-Dubbing: Zero-Shot Speech Synthesis from Visual Articulation

Anonymous ACL submission

Abstract

In the field of speech synthesis, there is a growing emphasis on employing multimodal speech to enhance robustness. A key challenge in this area is the scarcity of datasets that pair audio with corresponding video. We employ a methodology that incorporates modality alignment during the pre-training phase on multimodal datasets, uniquely facilitating Zero-Shot generalization through the process of freezing the video modality feature extraction component and the encoder module within the pretrained weights, thereby enabling effective cross-modal and cross-lingual transfer. We have named this method 'Uni-Dubbing'. Our method finely tunes with both multimodal and single-modality audio data. In multimodal scenarios, it achieves a reduced word error rate (WER) of 31.73%, surpassing the previous best of 33.9%. It also excels in metrics like tone quality and synchronization. With single-modality audio, it achieves a WER of 36.08%, demonstrating adaptability to limited data. Its domain generalization capabilities are proven across various language tasks in video translation and audio generation. Trained on 433 hours of audio data, it surpasses techniques using 200 hours of audio-visual data. The code and demo are available at https://diracer.github.io/unidubbing.

1 Introduction

With the widespread use of short videos and online meetings in daily life and the workplace(Gupta et al., 2023), the barrier of cross-linguistic communication has become an urgent problem, and thus multimodal technologies have attracted much attention(Yemini et al., 2023). Recently, many researchers have conducted corresponding studies in this area, such as lip reading task(Assael et al., 2016; Koumparoulis et al., 2017; Chung and Zisserman, 2016; Son Chung et al., 2017) that transfers video domain to text domain, Lip task(Prajwal et al., 2020; Kim et al., 2021; Michelsanti et al., 2021; Mira et al., 2022b) that transfers video domain to audio domain, and lip translation(Huang et al., 2023) that converts to the target language directly based on lips. In the case of the field of visual tasks, the biggest challenge for researchers is the extreme scarcity of training data. In addition, the relationship between lips and speech is not always a simple one-to-one mapping; for example, the same word may have very different lip shapes for people with different accents(Choi et al., 2023a). Therefore, maintaining accurate intonation poses a significant challenge, and this has led to the emergence of many important research findings.

For these reasons, we adopt the strategy of using discrete units as intermediate targets, i.e., transforming audio and video data into discrete units for alignment, which can effectively circumvent the disadvantage of insufficient paired audio and video data. On top of this, we employ the RVQ(Défossez et al., 2022) module thus enabling the method to achieve better timbre preservation, i.e. high fidelity, after Full-Shot training. Furthermore, in order to cope with the lack of data for contemporary visual tasks, we also use mHubert(Polyak et al., 2021) and K-means of re-combining with discrete units, which enables our model to achieve better semantic consistency and reach Zero-Shot capability. As mentioned earlier, the barriers to cross-language communication are equally significant challenges and a lot of good work has emerged, but unfortunately none of the current methods have been able to achieve Zero-Shot cross-language video translation yet. We further explored learning cross-language and cross-modal Lip2Wav mappings from the audio domain, i.e., Zero-Shot transspeech, based on the Zero-Shot Lip2Wav model, have verified that the method is capable of cross-language migration.

In summary, our goals in the current cross-language video-to-speech translation are twofold:

1) High quality and low error: the requirement to be able to recognise the gender in a video so as to generate the corresponding tones with minimal error is very challenging. 2) Zero-Shot: the ability of the reasoning process to achieve Zero-Shot is crucial for practicality when considering video translation.

Based on these two goals, in this paper, the innovation of this study lies in proposing a framework that requires only cross-linguistic audio speech training, without the need for visual speech training inputs, to achieve direct synthesis of visual speech to cross-linguistic audio speech. This framework can predict the corresponding audio speech output by analyzing an individual's lip movements, and this prediction is not limited to the language system of the input visual speech. Our method utilizes an advanced Zero-Shot learning strategy (Cheng et al., 2023) that aligns audio and visual phonemes with audio data alone during the training process, thus enabling the prediction of audio outputs in a target language that has not been seen before in seemingly impossible cross-modal scenarios. The main contributions of this paper are:

- Our cross-modal Zero-Shot transfer approach for the Lip2Wav task, trained exclusively with target audio, matches top Full-Shot models in WER, sound quality, and synchronization.
- Our method in the Lip2Wav task on the LRS3 dataset attains state-of-the-art results in WER, ESTOI, LSE-C, and LSE-D, achieving partial timbre preservation to distinguish voice characteristics of unseen speakers.
- Our cross-lingual audio generation technology creates target language audio from single-language videos, eliminating the need for dual-language video training. This streamlines training and lessens the need for extensive datasets in cross-lingual dubbing, while also reducing noise.

2 Related Work

In our paper, for the cross-language Lip2Wav synthesis task we mainly divide it into two steps: first implementing high-fidelity video-to-speech synthesis, followed by Zero-Shot cross-language video-to-speech translation. A great deal of excellent research work has preceded our study.

2.1 Video to Speech Synthesis

Video speech synthesis techniques(Cooke et al., 2006; Afouras et al., 2018b; Shi et al., 2022) that dub silent videos have received a great deal of attention from researchers in the recent past. Prajwal et al. (2020) presented the Lip2Wav, which utilizes a sequence-to-sequence architecture, enabling it to accurately capture contextual information and generate precise audio. Hong et al. (2021) trained a multimodal memory network, VV-Memory, to store and recall audio features corresponding to visual inputs so that audio information can be accessed exclusively through visual inputs during inference. Vougioukas et al. (2019) introduced an end-to-end temporal model based on GAN, capable of generating speech that synchronizes seamlessly with silent videos, presenting a convincing and difficult-to-distinguish quality. Additionally, there have been several recent papers based on GANs(Kim et al., 2021; Hong et al., 2022; Mira et al., 2022b). Most recently, a new method based on diffusion, called DiffV2S, has been proposed by Choi et al. (2023a) who introduced a novel speaker embedding extractor guided by visual information and simultaneously developed a diffusion-based video-to-speech synthesis model. Choi et al. (2023b) built upon the Lip2Wav model by incorporating quantized supervised speech representations, namely speech units, for synthesizing intelligible speech from silent videos.

However, despite the fact that all the aforementioned related methods have their own merits, the problem of lack of training data for the visual task mentioned in the previous section remains unsolved. With this in mind, we train our model by using discrete units as intermediate comparison targets in the audio and video domains, thus no longer relying on paired audio and video data.

2.2 Cross-language Translation

The task of cross-language translation is also a very challenging and important endeavour that also receives a lot of attention.(Lavie et al., 1997; Wahlster, 2000; Nakamura et al., 2006; ITU, 2016). Tjandra et al. (2019) introduced a discrete representation of the source language to target speech into the cascaded S2ST system, where this discrete representation is predicted by a separately trained VQVAE and subsequently utilized by the VQVAE decoder to generate the target speech spectrogram. Zhang et al. Zhang et al. (2021) proposed the XL-VAE model to enhance the discretization and reconstruction capabilities of VQVAE through cross-linguistic speech recognition. Lee et al. (2021) utilizes a separately trained vocoder, which includes

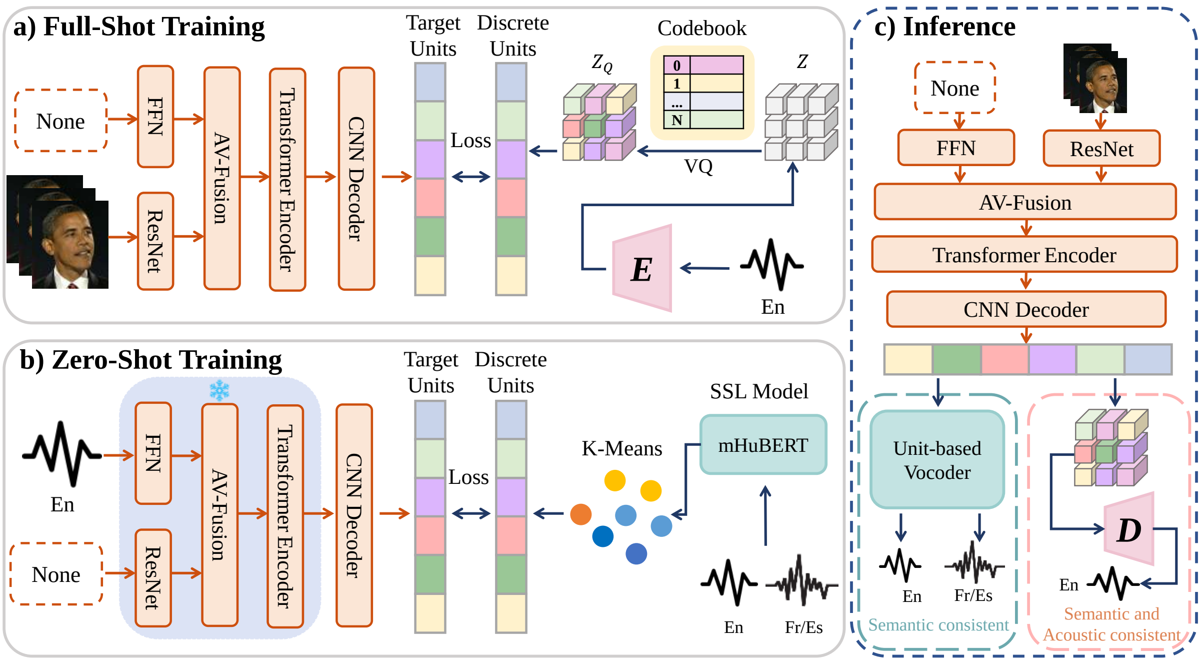


Figure 1: Uni-Dubbing Overview: In the high-fidelity Lip2Wav task, we employed a Full-Shot training approach and improved the generation of discrete units. The discrete units generated by this method capture more fine-grained acoustic information. For the cross-modal and cross-language Zero-Shot tasks, we adopted an approach similar to uHubert (Hsu and Shi, 2022), where no visual data is used during training and fine-tuning. Another distinction from the Full-Shot method is that, in Zero-Shot tasks, we froze the feature extraction and Encoder modules to prevent excessive loss of original visual knowledge during knowledge transfer. During inference, we input only visual data and use the corresponding Vocoder to generate audio through discrete units. The speech generated in the Zero-Shot manner contains only semantic information, while the Full-Shot generated speech not only includes semantic information but also retains some acoustic information.

a duration predictor, to directly predict waveforms from discrete representations. Jia et al. (2019) first introduced a model based on a sequence-to-sequence architecture capable of end-to-end training and inference. To improve translation quality and overgeneration, Jia et al. (2022) presented Translatotron2, which consists of a speech encoder, a language decoder, an acoustic synthesizer, and a single attention module that connects them together. There is also some work that attempts to introduce visual speech to enhance robustness in the translation process(Huang et al., 2023).

To the best of our knowledge, paired cross-lingual audio-video datasets are currently very sparse. This scarcity results in only one existing model capable of achieving cross-lingual Lip2Wav translation. Instead, in direct contrast with the methods mentioned above, our innovative discrete-unit-based approach can successfully cross these dataset barriers, thus learning cross-language visual-phoneme mappings with Zero-Shot cross-language lip-synthesis translation capability.

3 Method

3.1 Overview

The overview of this paper is depicted in Figure 1. Figure 1a) describes the training process for high-fidelity speech synthesis, while Figure 1b) illustrates the training flow for two tasks: cross-modal and cross-language. The main differences between these tasks lie in the modality used during training, the method for generating discrete units, and the treatment of predicted discrete units for synthesizing speech. Additionally, for Zero-Shot training, it is necessary to freeze the encoder to retain the visual knowledge acquired during the pretraining phase.

3.2 High-Fidelity Lip2Wav

While the state-of-the-art ReVISE model (Hsu et al., 2023) achieves leading performance in Lip2Wav synthesis on the LRS3 dataset, it does not preserve the speaker's timbre during speech synthesis. To address this issue, we propose a novel approach that utilizes acoustic tokens derived from the Encodec model (Défossez et al., 2022).

The Encoder model consists of an audio encoder, a Residual Vector Quantizer (RVQ), and an audio decoder: Consider an audio signal x with a length of d and sampled at a rate of sr, resulting in a total duration of $T = d/sr$.

1) Initially, the audio encoder E, comprising multiple convolutional blocks, processes the input audio. This encoder extracts features and outputs a latent representation z. 2) Subsequently, the Residual Vector Quantizer Q employs vector quantization layers to convert z into a discrete representation z_Q. In this process, the speech utterance x is encoded as a sequence of acoustic tokens $[a_1, a_2, ..., a_T]$, where each token a_i is an element of the set $\{0, 1, ..., K_2 - 1\}$, with $1 \leq i \leq T$. These acoustic tokens are the discrete units that we focus on in our training. 3) The audio decoder G reconstructs the signal $\hat{x}$ from the highly compressed latent representation z_Q. This algorithm efficiently quantizes the encoder output by iteratively refining the residual, which helps in preserving important information while reducing redundancy. Further, to address the challenges of temporal synchronicity in Lip2Wav tasks, we have innovated upon the existing AV-Hubert model. We have replaced the AV-Hubert decoder with a new structure.

Our adaptation involves a unique decoder structure, which includes three transposed convolutional layers. Each layer has a kernel size (K) of 4, a stride (S) of 2, padding (P) of 1, and output padding (O_p) of 1. This configuration is meticulously designed to more accurately align lip movements with the generated speech, thereby enhancing the synchronicity that is crucial for effective Lip2Wav synthesis. The output size (O) of each transposed convolutional layer is calculated using the formula:

$$O = ((I - 1) \times S + K - 2 \times P) + O_p \qquad (1)$$

where I denotes the input size.

3.3 Zero-Shot Lip2Wav Model Adaptation

To overcome the challenge of scarce paired audio-visual datasets, we loaded the pre-trained weights of AV-Hubert and focused on fine-tuning with pure audio data. To validate the effectiveness of our approach, we adopted the same Zero-Shot configuration on the LRS3 dataset as uHubert. The AV-Hubert model, pre-trained on paired audio-visual data, achieves multimodal alignment by mapping visual speech and audio speech to the same phoneme space. During the fine-tuning phase with pure audio data, we froze the decoder and only trained the final transposed convolution layer to preserve the multimodal alignment knowledge acquired during pre-training. In the inference process, the model processes silent lip videos, predicting the corresponding speech discrete units solely based on lip movements. This Zero-Shot learning strategy enables the model to effectively synthesize speech from unseen lip movements, enhancing its robustness in diverse scenarios.

To further validate the effectiveness of our method, we fine-tuned the model using discrete units generated in other languages (e.g., Spanish, French), which were languages not encountered during pretraining. This approach not only enables the model to generate speech from lip movements but also to translate it into different languages. For example, during inference, an English spoken video could be decoded into the audio of another language, simplifying the process of speech synthesis and translation without the need for separate models for each task.

In these two tasks, our model does not contain any speaker embeddings and is unable to implicitly acquire visual feature embeddings of the speaker during the fine-tuning phase, eliminating the need to replicate the speaker's acoustic information. Therefore, we used semantic tokens generated by the mHubert and kmeans methods as target units. Compared to acoustic information, semantic information has broader applicability, making the use of semantic tokens more conducive to generalization in cross-modal and cross-language Zero-Shot tasks.

3.4 Trainning Object

In this study, the focus is on predicting discrete units, for which the cross-entropy loss function

$$L = \sum_{t} \sum_{j=1}^{C} z_t^j \log f_t^j(\tilde{x}_a, x_v)$$

is consistently employed. This formula calculates the loss by summing over all frames (t) and across the C units in the vocabulary. The term z_t^j denotes the one-hot encoded label of the j-th unit in the t-th frame, and $f_t^j(\tilde{x}_a, x_v)$ represents the predicted probability distribution over the discrete units for the same frame and unit, as outputted by the enhancer.

Method	ESTOI ↑	LSE-C ↑	LSE-D ↓	WER ↓	MOS↑
VCA-GAN (Kim et al., 2021)	0.207	4.54	9.63	96.63	1.5±0.19
SVTS (Mira et al., 2022a)	0.244	7.08	7.04	79.83	1.96±0.24
Multi-task (Kim et al., 2023)	0.240	4.85	9.15	66.78	1.77±0.24
DiffV2s (Choi et al., 2023a)	0.284	7.28	7.27	39.2	4.06±0.21
ReVISE (Hsu et al., 2023)	0.285	7.12	7.25	33.9	4.11±0.04
Uni-Dubbing (Full-Shot)	**0.294**	**7.58**	**6.90**	**31.73**	**4.16**±0.06
Uni-Dubbing (Zero-Shot)	0.235	6.70	7.59	36.08	4.08±0.05

Table 1: The results of various methods on the test set of the LRS3 dataset are shown. The symbol ↑ indicates that higher values are better, while ↑ signifies that lower values are preferable.

4 Experiment

4.1 Datasets

LRS3 Dataset LRS3 (Afouras et al., 2018c) is an extensive and open-source benchmark collection for visual speech recognition research, commonly known as lip-reading. This dataset is the successor to the LRW (Chung and Zisserman, 2016) and LRS2 (Afouras et al., 2018a) datasets and features a vast array of labeled video content with corresponding textual transcriptions, primarily sourced from TED Talks.

LRS3-T Dataset LRS3-T (Huang et al., 2023) is a new audio-visual translation dataset that has been generated from the LRS3 dataset through a cascading process, combining Neural Machine Translation (NMT) and Text-to-Speech (TTS) technologies. This intricate processing sequence culminated in a parallel audio-visual translation dataset comprising 200 hours, encompassing both the original source videos and the translated speech in the target language.

MUSAN Dataset MUSAN (Snyder et al., 2015) is a collection of music, speech, and noise recordings suitable for audio processing tasks such as speech activity detection and machine learning applications. It features 60 hours of speech from various sources, over 42 hours of diverse music tracks, and 6 hours of environmental and technical noises. We used it to generate various types of noise which were added to the original audio, in order to test the translation task's resistance to noise interference.

4.2 Evaluation

In our study, we evaluate Lip2Wav and audio-video translation using key metrics. For semantic accuracy, we use WER , and for sound quality, we employ the Extended Short-Time Objective Intelligibility (ESTOI). Synchronization is measured using LSE-D (predicted audio-video temporal distance) and LSE-C (prediction confidence), as per SyncNet (Chung and Zisserman, 2017). Our method approximates the speaker's voice, thus we use the Mean Opinion Score (MOS) for evaluating timbre. To ensure consistency with other studies, we adopted a scoring system ranging from 1 to 5, with increments of 0.5 points. For each model, we randomly selected 50 samples for evaluation. We recommend listening to our website's audio samples for a practical understanding.

For language translation, we apply the BLEU (Papineni et al., 2002) score to evaluate the accuracy and fluency of speech generation in different languages, comparing machine-generated text to reference texts.

4.3 Results

4.3.1 High-Fidelity Video-to-speech synthesis

Unlike other datasets that may concentrate on short phrases or isolated words, LRS3 offers longer sequences of speech, enabling more complex and contextually rich lip-reading tasks. Since most speakers only give a TED talk once, the LRS3 dataset is multi-speaker, with no overlap between the speakers in the test set and those in the training set. Consequently, most methods using fixed ID speaker embeddings are ineffective for the LRS3 dataset without altering its test set. This reflects real-world application needs more accurately, as the models we train should be effective for unseen speakers. This paper focuses on speaker generalization on the original LRS3 dataset, aiming to generate audio that is perceptually credible for speakers it has never encountered before.

As shown in Table 1, DiffV2s and ReVISE significantly outperform various previous methods, with both achieving a WER below 40% and superior sound quality as evidenced by the ESTOI metric. Our results clearly surpass all prior work in these two measures, achieving a WER of 0.296 and an ESTOI of 31.96%. This is because acoustic units preserve finer details, making the generated audio easier for automatic speech recognition (ASR) systems to understand. In terms of synchronization, our model also achieved the highest rankings on the LSE-C and LSE-D metrics, surpassing all previous methods. This achievement is primarily attributed to our modifications to the original AV-Hubert decoder. We transformed it from a sequence-to-sequence model to one utilizing transposed convolutions. This change effectively ensures that the ratio between the input and output lengths of the model remains constant, thus maintaining a consistent proportional relationship between the generated audio length and the input video length. If the original AV-Hubert decoder were used, the LSE-C and LSE-D scores would be 4.65 and 9.21, respectively. Although our WER has only increased by 1.17% relative to the ReVISE, the additional fine-grained acoustic information plays a crucial role in improving synchronization. This allows our method to outperform ReVISE in terms of synchronization even when using the same transposed convolution decoder.

While quantitative metrics are important, they are not the key focus of our task. The primary contribution of our work lies in generating audio that retains partial speaker information without using the identity of the speaker. In contrast, ReVISE produces audio in a single female voice for all outputs, regardless of whether the video features a male speaker. Due to the absence of explicit speaker identity information, our method is unable to fully replicate the unique acoustic characteristics of individual speakers. However, due to its use of implicit visual embeddings and acoustic discrete units, the system is capable of generating distinct male or female voices, depending on whether the videos feature male or female speakers as protagonists. While the synthesized voices may not precisely match those of the original speakers, they do preserve certain overarching characteristics, such as gender distinctions and, to some extent, age differences. We believe this aspect is significant. In cases where humans have not seen the speaker, they cannot deduce the exact timbre from the video but can infer such general voice characteristics. The voices generated by our model align with human perception, thus meeting human expectations and requirements.Benefiting from this approach, our MOS evaluation achieved an optimal score of 4.16.

4.3.2 Zero-Shot from Audio to Video

Table 1 reveals that our method achieves impressive results even when trained solely with audio, without using any video data. The sound quality, measured by the ESTOI, is 0.235. This performance is comparable to the previous three works, ranking just behind DiffV2S and ReVISE. Surprisingly, despite the absence of video data during training, the synchronization of our generated audio is quite good, significantly surpassing the Full-Shot VCA-GAN and Multi-task methods, and comparable to other approaches. Most importantly, our method achieves a WER of 36.08%, which is only slightly inferior to ReVISE's 33.9% and better than all previous Full-Shot methods. These results indicate that our approach effectively utilizes the knowledge embedded in the pre-trained model to achieve outstanding performance, while significantly reducing data collection costs, requiring only pure audio data without corresponding lip-synced video.

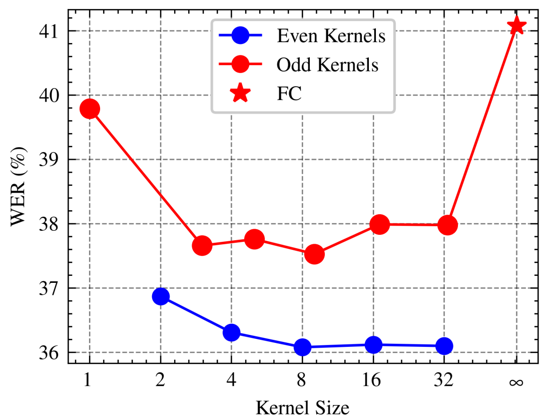


Figure 2: The curve graph illustrating the relationship between the kernel size of the last layer of transposed convolution and the corresponding WER. When the kernel size is odd, the stride is set to 1; for even kernel sizes, the stride is 2. Therefore, we have plotted two separate curves for odd and even kernel sizes to analyze the impact of stride.

Due to the mHubert audio encoder operating at 50 frames per second and the AV-Hubert video encoder at 25 frames per second, we employed a convolutional layer to align the two. It was imperative to set the stride of this transposed convolution to 2, a fixed requirement. However, the size of the convolutional kernel significantly impacted the final results. To determine the optimal kernel size, we conducted multiple experiments. For comparison, we also tried the alignment method used in AV-Hubert pre-training, which involves downscaling the audio labels' discrete units to 25 frames per second by extracting them at intervals. In this scenario, we set the stride of the transposed convolution to 1 and chose a convolutional kernel of an odd size.

As shown in Figure 2, all models using odd-numbered kernel sizes performed worse in terms of WER compared to those using even-numbered kernels. Specifically, smaller even-numbered kernels, such as 2 and 4, significantly reduced accuracy. However, the performance improvement became marginal when the kernel size increased to 8 or larger. Based on this finding, we selected a kernel size of 8, balancing optimally between temporal resolution and computational efficiency, crucial for effective synchronization between audio and video modalities. Additionally, we experimented with the original fully connected (FC) layer. The results indicated that using an FC layer instead of transposed convolutions yielded the worst outcomes, highlighting the effectiveness of transposed convolutions in extracting local information for our task.

A noteworthy observation is that methods comparable to Zero-Shot in terms of ESTOI generally have a WER exceeding 60%. This implies that Zero-Shot is capable of acquiring a substantial degree of semantic knowledge from pre-training, but it slightly lags in generating audio quality, failing to reach a level commensurate with its semantic proficiency.

4.3.3 Translate from Video

Building on the concepts discussed earlier, collecting audio and its corresponding lip-synchronized video data presents significant challenges. These challenges further escalate when the task is extended to multiple languages. Our objective is to utilize datasets composed of video-audio pairs in a single language, combined with multilingual audio datasets, to make this approach applicable to multilingual audio generation. This strategy aims to efficiently utilize existing resources while addressing the challenges of multimodal and multilingual datasets.

In our study, we compared the performance of existing Full-Shot methods with our Zero-Shot method in English to Spanish (En-Es) and English to French (En-Fr) translation tasks, with detailed results presented in Table 2. We also tested the

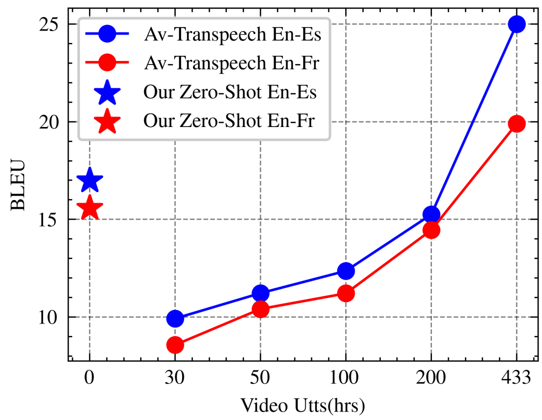


Figure 3: The comparison between Uni-Dubbing and Av-Transpeech under various sizes of visual speech data is highlighted. Remarkably, Uni-Dubbing, utilizing a Zero-Shot approach, outperforms Av-Transpeech even when the latter is fine-tuned with 200 hours of visual data.

robustness of our model under different modalities and specific noise conditions. Firstly, we found that under given noise conditions, the BLEU scores using both visual and audio modal inputs were consistently higher than those using only audio input. This demonstrates the auxiliary role of visual information in enhancing audio in noisy environments, highlighting the importance of visual data. Especially under babble noise conditions, with a signal-to-noise ratio (SNR) of -5, the BLEU score for pure audio input was even lower than that for pure visual input, further emphasizing the significance of lip-reading translation. We also provided experimental data under various noise types and intensities in the appendix. In pure visual translation, Full-Shot methods typically outperform Zero-Shot methods. However, the Zero-Shot method still performs commendably in terms of BLEU scores and MOS, achieving BLEU scores of 16.99 and 19.90, and MOS of 3.73 and 3.70, respectively.

We replicated Av-Transpeech and fine-tuned it using multimodal data of varying durations, with detailed results shown in Figure 3. The figure demonstrates that the BLEU score obtained by fine-tuning with 433 hours of pure audio data is roughly equivalent to that achieved with just 220 hours of audiovisual data. During the pre-training phase, we mapped the audiovisual data to the same phoneme space. This result indicates that the knowledge in this phoneme space is equally applicable to cross-lingual audio, enabling us to align the source language video with the target language audio through

Type	Method	Training		Eval		En-Es		En-Fr	
		A	V	A	V	BLEU ↑	MOS ↑	BLEU ↑	MOS ↑
Full-Shot	Av-Transpeech (Huang et al., 2023)	✓	✓		✓	**25.00**	3.94 ± 0.11	**19.90**	3.95 ± 0.10
		✓	✓	✓	✓	33.10	-	28.00	-
		✓	✓	✓		5.50	-	4.60	-
Zero-Shot	Uni-Dubbing (Frozen)	✓			✓	**16.99**	3.73 ± 0.12	**15.58**	3.70 ± 0.08
		✓		✓	✓	30.00	-	25.30	-
		✓		✓		7.58	-	6.31	-
	Uni-Dubbing (No Frozen)	✓			✓	**0**	-	**0**	-
		✓		✓	✓	0.94	-	1.39	-
		✓		✓		0.92	-	1.07	-

Table 2: Comparison of translation results between the Full-Shot method and our method across various modalities and noise environments. It's worth noting that babble noise with an SNR of -5 is added to all instances using the audio modality (including AV and A) during inference. Please refer to the appendix for additional experimental results on different types of noise and their intensities.

pure audio fine-tuning, resulting in the current BLEU scores. This finding not only validates the effectiveness of our method but also emphasizes the feasibility of using a large amount of pure audio data as an alternative in scenarios where it is challenging to collect extensive multimodal data.

In our study, as illustrated in Table 2, we additionally conducted an experiment to investigate the translation results obtained using our Zero-Shot method without freezing the encoder. This part of the experiment primarily aimed to assess the role of freezing the encoder in preserving pre-trained knowledge. Under this setup, we observed a significant phenomenon: the BLEU scores for model inference on pure video were zero in both En-Es and En-Fr translation tasks. This result implies that the majority of the visual knowledge acquired during the model's pre-training phase has been substantially forgotten in subsequent processes.

Furthermore, compared to models that kept the encoder frozen during the inference phase, the models with unfrozen encoders also showed lower resistance to noise. This difference not only reveals the importance of freezing the encoder for maintaining model stability but also reflects the criticality of preserving knowledge acquired during pre-training when dealing with complex and variable visual inputs. Freezing the encoder effectively retains the visual information learned during the pre-training phase, which is crucial for enhancing the model's accuracy and robustness in parsing and understanding visual data. Therefore, our study not only emphasizes the importance of managing the state of the encoder in implementing Zero-Shot learning methods but also provides valuable insights for future model design in the intersection of vision and language domains.

5 Conclusion

This paper introduces Uni-Dubbing, an innovative approach trained on multimodal audio-video datasets, which achieved the best WER, ESTOI, and synchronization metrics on the LRS3 dataset. Additionally, by utilizing implicit visual embeddings and acoustic tokens, we successfully preserved partial speaker information on the cross-speaker LRS3 dataset. We then implemented a Zero-Shot strategy, transitioning from audio to video modalities in cross-modal Lip2Wav tasks, and cross-lingual Lip2Wav translation tasks. This method significantly reduces the dependency on multimodal datasets and demonstrates potential for application in a wider range of tasks.

To further validate the practicality of this method, our research utilized only the audio portion of existing multimodal datasets. In future work, we plan to explore the use of larger single-modality audio datasets, aiming to further expand the applicability and enhance the effectiveness of this method. Through such research, we hope to deepen our understanding and utilization of single-modality audio data in multimodal tasks, thereby paving new paths for development in this field.

6 Ethics Statement

In the context of our research, we acknowledge that lip-reading technology holds considerable potential in a multitude of applications, such as facilitating silent commands in noisy environments or enhancing communication for individuals with hearing impairments. The OpenSR system is designed to democratize the development of lip-reading models, particularly for domains where resources are scarce, thereby promoting equality in technology application across different fields and languages.

However, we recognize the ethical implications surrounding the use of speech recognition technology, including the potential for unintended information exposure. It is important to note that effective lip-reading by our model demands specific video criteria, such as front-facing, high-resolution imagery with sufficient frame rates to ensure clear visibility of lip movements. Typically, such conditions are met in environments with close-range cameras or during virtual meetings, not in scenarios where video footage is obtained from a distance or without clear visibility of the mouth region, like most surveillance contexts.

Therefore, while our model advances the field of speech recognition, it is engineered with inherent limitations that naturally restrict its use in situations that could compromise individual privacy. We maintain a commitment to ethical research practices, prioritizing the beneficial impacts of our work while actively mitigating potential risks of misuse that could infringe on personal privacy or be deemed invasive. Our ongoing research includes a strong focus on developing safeguards and protocols to ensure that the technology is used responsibly and ethically.

7 Limitations

The present study is limited to the use of just two modalities: video and audio, thus neglecting the potential benefits of incorporating further modalities. Furthermore, the approach of applying single-modality Zero-Shot learning, although it minimizes reliance on extensive datasets, inherently results in the inadvertent omission of some portions of the previously acquired knowledge. Consequently, this methodology is not entirely effective in preserving the full spectrum of multimodal alignment knowledge that was initially obtained during the training phase.

References

T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman. 2018a. Deep audio-visual speech recognition. In *arXiv:1809.02108*.

Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. 2018b. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence*, 44(12):8717–8727.

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

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Yannis M Assael, Brendan Shillingford, Shimon Whiteson, and Nando De Freitas. 2016. Lipnet: End-to-end sentence-level lipreading. *arXiv preprint arXiv:1611.01599*.

Xize Cheng, Tao Jin, Linjun Li, Wang Lin, Xinyu Duan, and Zhou Zhao. 2023. OpenSR: Open-modality speech recognition via maintaining multi-modality alignment. *arXiv preprint arXiv:2306.06410*.

Jeongsoo Choi, Joanna Hong, and Yong Man Ro. 2023a. DiffV2S: Diffusion-based video-to-speech synthesis with vision-guided speaker embedding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7812–7821.

Jeongsoo Choi, Minsu Kim, and Yong Man Ro. 2023b. Intelligible lip-to-speech synthesis with speech units. *arXiv preprint arXiv:2305.19603*.

J. S. Chung and A. Zisserman. 2016. Lip reading in the wild. In *Asian Conference on Computer Vision*.

Joon Son Chung and Andrew Zisserman. 2017. 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.

Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. 2006. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120(5):2421–2424.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2022. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*.

Anchit Gupta, Rudrabha Mukhopadhyay, Sindhu Balachandra, Faizan Farooq Khan, Vinay P Namboodiri, and CV Jawahar. 2023. Towards generating ultra-high resolution talking-face videos with lip synchronization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5209–5218.

Joanna Hong, Minsu Kim, Se Jin Park, and Yong Man Ro. 2021. Speech reconstruction with reminiscent sound via visual voice memory. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3654–3667.

Joanna Hong, Minsu Kim, and Yong Man Ro. 2022. VisageSynTalk: Unseen speaker video-to-speech synthesis via speech-visage feature selection. In *European Conference on Computer Vision*, pages 452–468. Springer.

Wei-Ning Hsu, Tal Remez, Bowen Shi, Jacob Donley, and Yossi Adi. 2023. ReVISE: Self-supervised speech resynthesis with visual input for universal and generalized speech regeneration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18795–18805.

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.

Rongjie Huang, Huadai Liu, Xize Cheng, Yi Ren, Linjun Li, Zhenhui Ye, Jinzheng He, Lichao Zhang, Jinglin Liu, Xiang Yin, et al. 2023. AV − TranSpeech: Audio-visual robust speech-to-speech translation. *arXiv preprint arXiv:2305.15403*.

ITU. 2016. ITU-T F.745 : Functional requirements for network-based speech-to-speech translation services. *International Telecommunication Union*.

Ye Jia, Michelle Tadmor Ramanovich, Tal Remez, and Roi Pomerantz. 2022. Translatotron 2: High-quality direct speech-to-speech translation with voice preservation. In *International Conference on Machine Learning*, pages 10120–10134. PMLR.

Ye Jia, Ron J Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, and Yonghui Wu. 2019. Direct speech-to-speech translation with a sequence-to-sequence model. *arXiv preprint arXiv:1904.06037*.

Minsu Kim, Joanna Hong, and Yong Man Ro. 2021. Lip to speech synthesis with visual context attentional gan. *Advances in Neural Information Processing Systems*, 34:2758–2770.

Minsu Kim, Joanna Hong, and Yong Man Ro. 2023. Lip-to-speech synthesis in the wild with multi-task learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Alexandros Koumparoulis, Gerasimos Potamianos, Youssef Mroueh, and Steven J Rennie. 2017. Exploring roi size in deep learning based lipreading. In *AVSP*, pages 64–69.

Alon Lavie, Alex Waibel, Lori Levin, Michael Finke, Donna Gates, Marsal Gavalda, Torsten Zeppenfeld, and Puming Zhan. 1997. JANUS − III: Speech-to-speech translation in multiple languages. In *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 99–102. IEEE.

Ann Lee, Peng-Jen Chen, Changhan Wang, Jiatao Gu, Sravya Popuri, Xutai Ma, Adam Polyak, Yossi Adi, Qing He, Yun Tang, et al. 2021. Direct speech-to-speech translation with discrete units. *arXiv preprint arXiv:2107.05604*.

Pingchuan Ma, Alexandros Haliassos, Adriana Fernandez-Lopez, Honglie Chen, Stavros Petridis, and Maja Pantic. 2023. 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.

Daniel Michelsanti, Zheng-Hua Tan, Shi-Xiong Zhang, Yong Xu, Meng Yu, Dong Yu, and Jesper Jensen. 2021. An overview of deep-learning-based audio-visual speech enhancement and separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1368–1396.

Rodrigo Mira, Alexandros Haliassos, Stavros Petridis, Björn W Schuller, and Maja Pantic. 2022a. Svts: scalable video-to-speech synthesis. *arXiv preprint arXiv:2205.02058*.

Rodrigo Mira, Konstantinos Vougioukas, Pingchuan Ma, Stavros Petridis, Björn W Schuller, and Maja Pantic. 2022b. End-to-end video-to-speech synthesis using generative adversarial networks. *IEEE transactions on cybernetics*.

Satoshi Nakamura, Konstantin Markov, Hiromi Nakaiwa, Gen-ichiro Kikui, Hisashi Kawai, Takatoshi Jitsuhiro, J-S Zhang, Hirofumi Yamamoto, Eiichiro Sumita, and Seiichi Yamamoto. 2006. The atr multilingual speech-to-speech translation system. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(2):365–376.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux. 2021. Speech resynthesis from discrete disentangled self-supervised representations. *arXiv preprint arXiv:2104.00355*.

KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. 2020. Learning individual speaking styles for accurate lip to speech synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13796–13805.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

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*.

David Snyder, Guoguo Chen, and Daniel Povey. 2015. Musan: A music, speech, and noise corpus. *arXiv preprint arXiv:1510.08484*.

Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. 2017. Lip reading sentences in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6447–6456.

Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. 2019. Speech-to-speech translation between untranscribed unknown languages. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 593–600. IEEE.

Konstantinos Vougioukas, Pingchuan Ma, Stavros Petridis, and Maja Pantic. 2019. Video-driven speech reconstruction using generative adversarial networks. In *Interspeech*.

Wolfgang Wahlster. 2000. *Verbmobil: Foundations of Speech-to-Speech Translation*. Springer Science & Business Media.

Yochai Yemini, Aviv Shamsian, Lior Bracha, Sharon Gannot, and Ethan Fetaya. 2023. LipVoicer: Generating speech from silent videos guided by lip reading. *In arXiv:2306.03258*.

Chen Zhang, Xu Tan, Yi Ren, Tao Qin, Kejun Zhang, and Tie-Yan Liu. 2021. UWSpeech: Speech to speech translation for unwritten languages. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 14319–14327.

A Additional quantitative Results

Zero-Shot Kernel Size. The results of cross-modal Zero-Shot experiments conducted on the LRS3 dataset are closely related to the kernel size of the last layer of transposed convolution. Table 3 details these results, including ESTOI, WER, and synchronization rate metrics.

K	ESTOI ↑	LSE-C ↑	LSE-D ↓	WER ↓
2	0.228	6.54	7.77	36.87
4	0.235	6.67	7.64	36.31
8	0.235	6.70	7.59	**36.08**
16	0.234	6.75	7.61	36.12
32	0.235	6.72	7.60	36.10
1	0.211	6.22	8.05	39.79
3	0.214	6.32	7.95	37.66
5	0.214	6.37	7.91	37.76
9	0.216	6.39	7.91	**37.53**
17	0.214	6.40	7.90	37.99
33	0.214	6.39	7.90	37.98
FC	0.209	6.20	8.05	41.08

Table 3: The impact of varying kernel sizes on different metrics in audio generation. K represents the size of the kernel in the final layer of transposed convolution. FC (Fully Connected) represents a configuration where, instead of using a transposed convolution layer, a fully connected layer is employed as the final layer.

Zero-Shot Translate Data Size. For the Zero-Shot translation task, we present in Table 4 the performance of AV-Transpeech after fine-tuning with varying amounts of data. We compare the results of inference using both audiovisual data and video-only data. We found that for both AVST (Audio-Visual Synchronous Translation) and VST (Video Synchronous Translation) tasks, the effectiveness of our method is similar to that achieved by fine-tuning with a 200-hour multimodal audiovisual dataset.

Method	Utts(hrs)	En-Es		En-Fr	
		AV	V	AV	V
AV-Transpeech	433	45.2	25	33.6	19.9
	200	35.98	15.25	29.83	14.45
	100	31.59	12.36	27.64	11.21
	50	28.2	11.22	24.21	10.41
	30	24.92	9.92	15.96	8.57
Our(Zero-Shot)	433	36.53	16.99	28.94	15.58

Table 4: Translation Results of AV-Transpeech in Different Modalities After Fine-Tuning with Various Data Volumes.

Zero-Shot Translate Noise Robust. In the main text, we only present the performance of the model under partial noise conditions. Table 5 and Table 6 respectively showcase the results of the Zero-Shot model under frozen and no frozen states across various noise conditions.

Modality	Noise	Language	SNR -20	-10	-5	0	5	10	20	Average
AV	Babble	En-Es	13.45	23.61	30.00	34.15	35.40	35.55	36.07	29.75
		En-Fr	12.06	19.24	25.30	27.59	28.46	28.61	28.83	24.30
	Music	En-Es	23.93	31.73	34.54	35.09	35.81	35.56	36.18	33.26
		En-Fr	19.25	26.27	27.8	28.48	28.60	28.81	28.75	26.85
	Speech	En-Es	24.63	32.38	34.25	35.41	35.57	36.16	36.41	33.54
		En-Fr	19.83	26.21	27.69	28.72	28.92	28.55	29.30	27.03
	Average	En-Es	20.67	29.24	32.93	34.88	35.59	35.76	36.22	32.18
		En-Fr	17.05	23.91	26.93	28.26	28.66	28.66	28.96	26.06
A	Babble	En-Es	0.01	0.12	7.58	26.64	33.82	35.23	35.71	19.87
		En-Fr	0.05	0.17	6.31	21.54	27.18	28.55	29.41	16.17
	Music	En-Es	3.03	16.76	28.25	33.42	34.97	35.78	36.60	26.97
		En-Fr	3.47	15.01	22.47	27.11	28.18	28.97	29.11	22.05
	Speech	En-Es	4.11	17.97	27.88	33.89	34.79	35.53	36.09	27.18
		En-Fr	3.84	15.71	21.92	27.12	28.61	29.14	29.16	22.21
	Average	En-Es	2.38	11.62	21.24	31.32	24.53	35.51	36.13	24.68
		En-Fr	2.45	10.30	16.90	25.26	27.99	28.89	29.23	20.14
V	-	En-Es	16.99	16.99	16.99	16.99	16.99	16.99	16.99	16.99
	-	En-Fr	15.58	15.58	15.58	15.58	15.58	15.58	15.58	15.58

Table 5: Comparison of translation accuracy (BLEU score ↑) of our zero shot model between different noise configurations and input modalities. The BLEU scores for pure audio inference are lower than those for inference using only video in multiple scenarios when the noise intensity is high.

Modality	Noise	Language	SNR -20	-10	-5	0	5	10	20	Average
AV	Babble	En-Es	0.01	0.04	0.94	11.47	29.20	36.74	40.08	16.93
		En-Fr	0.11	0.14	1.39	10.26	24.33	30.93	33.94	14.44
	Music	En-Es	0.53	5.33	15.21	26.91	35.13	38.73	40.33	23.17
		En-Fr	0.40	5.31	12.91	22.63	30.19	32.67	33.70	19.69
	Speech	En-Es	0.65	7.63	16.73	28.21	34.87	38.52	40.02	23.80
		En-Fr	0.55	7.21	13.91	24.01	29.61	32.48	33.68	20.21
	Average	En-Es	0.40	4.33	10.96	22.20	33.07	38.00	40.14	21.30
		En-Fr	0.35	4.22	9.40	18.97	28.04	32.03	33.77	18.11
A	Babble	En-Es	0.01	0.01	0.92	10.60	28.76	36.96	40.01	16.75
		En-Fr	0.09	0.08	1.07	9.60	24.75	30.62	34.04	14.32
	Music	En-Es	0.48	6.92	15.61	26.06	34.37	38.40	40.04	23.13
		En-Fr	0.46	4.71	12.35	23.18	29.38	32.54	34.20	19.55
	Speech	En-Es	1.06	7.33	16.93	27.47	35.45	38.25	40.14	23.80
		En-Fr	0.66	6.53	14.50	23.46	29.82	32.30	33.83	20.16
	Average	En-Es	0.52	4.75	11.15	21.38	32.86	37.87	40.06	21.23
		En-Fr	0.40	3.77	9.31	18.75	27.98	31.82	34.02	18.01

Table 6: Comparison of translation accuracy (BLEU score ↑) of our no-frozon Zero-Shot model between different noise configurations and input modalities.

B Additional qualitative Results

LRS3 Dataset in Lip2Wav Implementation. In Figure 4, we display visualizations of four samples each from the ground truth, our Full-Shot and Zero-Shot methods, and ReVISE, to compare their respective mel-spectrogram outputs. These methods generate mel-spectrograms whose backbone structures maintain a certain degree of similarity, resulting in low WER and minimal differences in retained semantic information for the synthesized speech. However, in comparison, our Full-Shot method produces mel-spectrograms that more closely resemble real data (Ground Truth) in detail, displaying finer frequency variations and a more continuous temporal sequence structure. This indicates that the Full-Shot approach achieves higher accuracy in audio reconstruction, capturing more of the acoustic features of real speech signals beyond just semantic information. Additionally, our Zero-Shot method shows greater similarity to ReVISE, demonstrating that even when fine-tuned using only audio data, it can retain a considerable level of semantic information. This validates the effectiveness of our method in modal transfer.

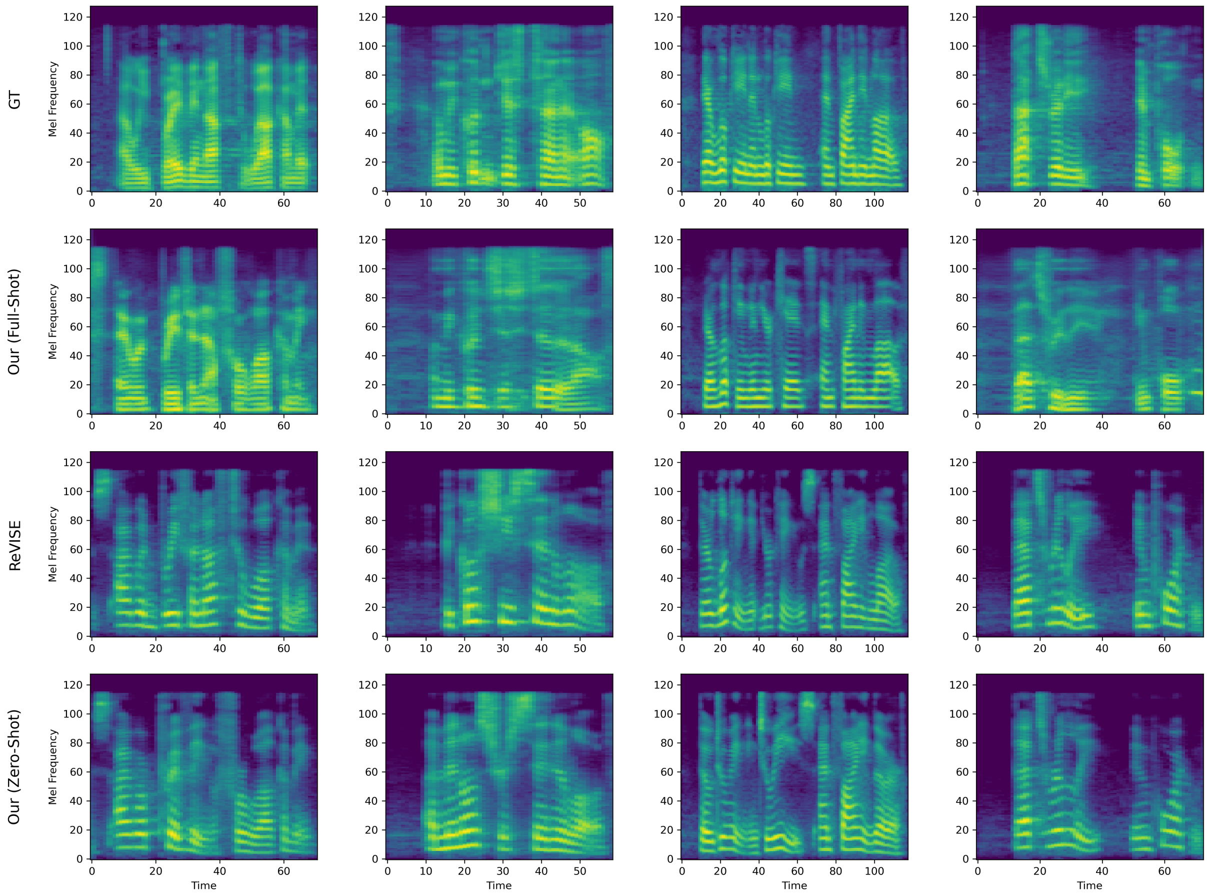
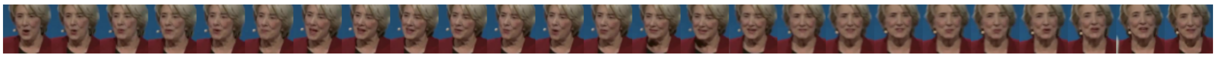


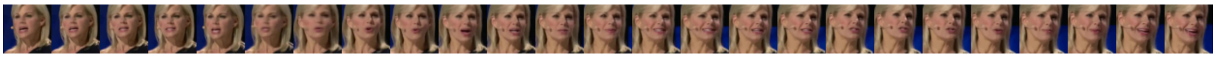
Figure 4: Sample mel-spectrogram visualizations from various methods on the LRS3 dataset.

In Table 7, we present the results of several audio samples processed through Lip2Wav and subsequently analyzed using ASR. The errors generated by these methods are largely similar, likely stemming from the inherent confusability of the Lip2Wav approach itself. This is because the majority of errors originate from phonetically similar words or phrases, which are exceedingly difficult to overcome in subsequent processing.

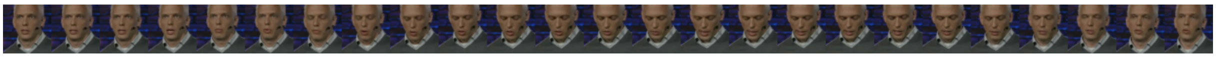
Table 7: This qualitative comparison addresses visually confusing words. 'Red words' highlighted in red indicate misidentified terms, ~~strikethroughs~~ in parentheses denote visually similar words, and (red words) within parentheses emphasize words that are absent.



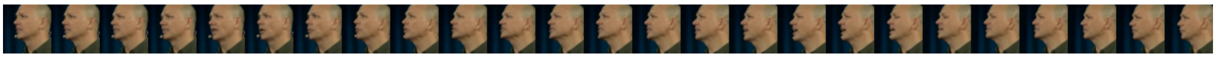
Ground Truth:	we were making what was invisible visible
Our(Full-Shot):	we were making what was invisible invisible ~~(visible)~~
ReVISE:	we were many ~~(making)~~ what was invisible invisible ~~(visible)~~
Our(Zero-Shot):	we were many ~~(making)~~ what was invisible visible



Ground Truth:	would you like to create a second one together
Our(Full-Shot):	would you like to create a successful ~~(second)~~ one together
ReVISE:	would you like to create ~~(a)~~ success when you guess ~~(second one together)~~
Our(Zero-Shot):	would you like to be in a cecil when ~~(create a second one)~~ together



Ground Truth:	african americans supported it at a higher level than had ever been recorded
Our(Full-Shot):	african americans supported it at a higher level than had ever been recorded
ReVISE:	african americans supported it at a higher level than it ~~(had)~~ ever been recorded
Our(Zero-Shot):	african americans supported it at a higher level than it ~~(had)~~ ever be ~~(been)~~ recorded



Ground Truth:	dan replies so often you won't even notice it
Our(Full-Shot):	ten ~~(dan)~~ replies so often you won't even notice it
ReVISE:	the data ~~(dan)~~ replies so often you won't even notice it
Our(Zero-Shot):	ten ~~(dan)~~ replies so often you won't even notice it

LRS3-T Dataset in Cross-Lingual Lip2Wav Translation. In Figure 5, we display the actual spectrograms for En-Es and En-Fr samples, along with the corresponding spectrograms generated by Av-Transpeech and our Zero-Shot method. The mel-spectrograms generated by Av-Transpeech show a high degree of similarity to those produced by our method, but both exhibit certain differences from the GT. This is primarily because both methods use discretized units generated in the same way as training targets, hence the information they carry is quite similar, primarily focusing on semantic information. On the LRS3-T dataset, the similarity of the mel-spectrograms generated by these two methods further confirms the Zero-Shot capabilities of our approach.

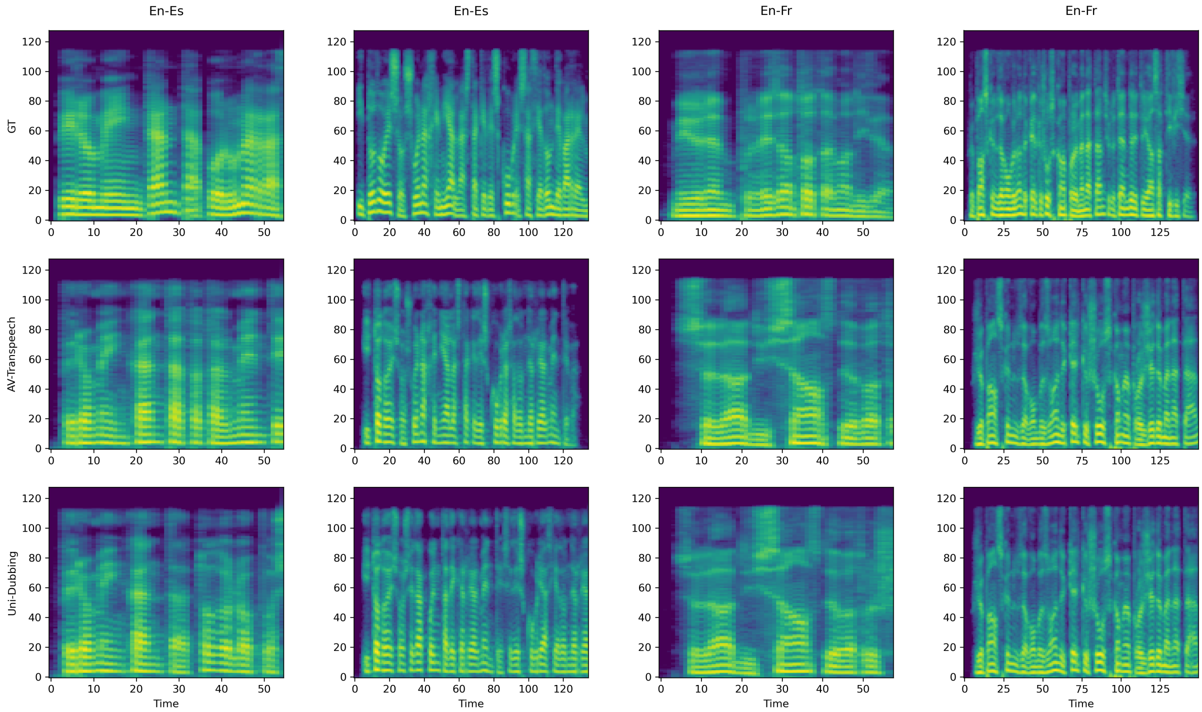
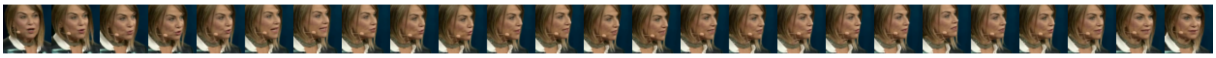


Figure 5: Sample mel-spectrogram visualizations from various methods on the LRS3 dataset.

Table 8 shows GT, Av-Transpeech, and our En-Es and En-Fr translation results. Our translations contain more erroneous words compared to Av-Transpeech, as reflected in the lower BLEU scores reported in the text. However, the locations of errors are similar for both methods, suggesting that pure audio fine-tuning might achieve semantics similar to Full-Shot for the main body of sentences, but there could be confusion in some details. Further research and exploration in this area are needed.

Table 8: This qualitative comparison addresses visually confusing words. 'Red words' highlighted in red indicate misidentified terms, ~~strikethroughs~~ in parentheses denote visually similar words, and (red words) within parentheses emphasize words that are absent. The top two samples are En-Es translations, and the bottom two are En-Fr translations.



Ground Truth:	te gustaría crear un segundo juntos
Av-Transpeech:	te gustaría crear una sensación ~~(un segundo)~~ juntos
Uni-Dubbing	te gustaría crear un sentido conjunto ~~(juntos)~~

Ground Truth:	podemos crear un parlamento mundial de alcaldes
Av-Transpeech:	podemos crear un parlamento global ~~(mudial)~~ de pares
Uni-Dubbing	necesitamos ~~(podemos)~~ crear un parlamento global ~~(mudial)~~ de c ~~(alcaldes)~~

Ground Truth:	Je te pardonne et je ne te hais pas
Av-Transpeech:	je te pardonne et je ne te déteste ~~(pas)~~
Uni-Dubbing	je te donne ~~(pardonne et)~~ je ~~(ne)~~ te déteste ~~(déteste pas)~~

Ground Truth:	donc la réponse à la deuxième question peut-on changer
Av-Transpeech:	donc la réponse à la deuxième question pouvants-nous change ~~(peut-on changer)~~
Uni-Dubbing	donc la réponse à la deuxième question pouvonts-nous ~~(peut-on)~~ changer

C Zero-Shot configuration

On the LRS3 dataset, our applied Zero-Shot configuration is consistent with that of uHubert (Hsu and Shi, 2022). One concern arises: the model might memorize audio-visual pairs from the pre-training period and associate them with unimodal data for Zero-Shot learning, as the dataset used for fine-tuning is a subset of the pre-training data. To address this issue, uHubert conducted experiments on non-LRS3 audio datasets, demonstrating the effectiveness of this configuration. Therefore, we did not seek another out-of-domain audio dataset for experimentation in this task. We directly conducted Zero-Shot experiments on LRS3-T, whose audio data is not only excluded from the pre-training but also differs in language type. Furthermore, ablation experiments regarding whether to freeze the encoder layers also validated the Zero-Shot capability of our method.

D More implementation details.

Experiment hyperparameters. Table 9 displays the training hyperparameter configurations for each task in our study, noting that audio masking was not employed in any of the tasks.

	Full-Shot	**Zero-Shot Modal**	**Zero-Shot Translate**
num. of updates	45000	20000	60000
num. of frozen	5000	20000	60000
tri-stage LR schedule	(10%,20%,70%)	(10%,20%,70%)	(33%,0%,67%)
peak learning rate	6e-05	6e-05	5e-04
batchsize /GPU	1000	1000	1000
num. of GPU	8	8	8
Adam (β_1,β_2)	(0.9,0.98)	(0.9,0.98)	(0.9,0.98)

Table 9: Experiment hyperparameters.

ASR toolkit for Evaluation. In this paper, the English ASR used is cited from (Ma et al., 2023). For Spanish and French, we utilize open-sourced ASR models within the *fairseq* framework (Ott et al., 2019) to transcribe the audios, which is consistent with the ASR used by Av-Transpeech.