# LipAdapter: Text-to-Video Alignment is All You Need for Lip-to-Speech

*Souvik Ghosh* [1], *C. V. Jawahar* [1], *Vinay Namboodiri* [2]

[1] IIIT Hyderabad, India
[2] University of Bath, United Kingdom
souvik.g@research.iiit.ac.in, jawahar@iiit.ac.in, vpn22@bath.ac.uk

## Abstract

Lip-to-Speech synthesis aims to generate intelligible, natural-sounding, and temporally synchronized speech from silent visual recordings of a speaker. Although recent end-to-end models have shown promising results, they require large amounts of training data and often lack out-of-dataset generalization. In this work, we introduce LipAdapter, a generic and efficient modular framework that adapts an existing frozen pre-trained Text-to-Speech model into a lip-synchronized speech generator. LipAdapter bridges the modularity gap between frozen Visual Speech Recognition and Text-to-Speech models by aligning phoneme representation and lip embeddings via a novel text-to-video alignment module. Our model achieves state-of-the-art performance across multiple benchmarks in English while using 15× less training data than previous methods. We also demonstrate the model's zero-shot capabilities in multilingual lip-to-speech synthesis in French, German, Spanish and Portuguese. Code and demo are available at: `https://lipadapter.github.io/`.

**Index Terms**: Lip-to-speech, video guided speech synthesis

## 1. Introduction

Lip-to-speech synthesis seeks to reconstruct intelligible, natural-sounding, and lip-synced speech from silent visual recordings of individuals speaking. It has applications in assistive technologies for people with speech impairments, communication in noisy environments, and the restoration of speech in corrupted or silent videos. Recent advances have led to powerful end-to-end models for silent video-to-speech synthesis, which directly map video frames to audio waveforms [1, 2, 3].

Although these systems have demonstrated good performance, their speech quality often deteriorates when applied to real-world conditions. The limitations may arise due to their speech generation components being trained on datasets curated for tasks such as Visual Speech Recognition (VSR), audiovisual speech recognition (AVSR). These datasets lack high-quality, clean speech samples on which modern Text-to-Speech (TTS) models are trained, leading to lower naturalness and expressiveness in generated speech. In addition, directly feeding raw visual information into the speech synthesis stage can introduce irrelevant variability and effectively act as noise, further degrading output quality. Other methods such as Lipvoicer [4] and Accurate L2S [5] predict text from a silent video using a VSR model, and train the speech generation pipeline by conditioning on both visual and textual guidance. They achieve a lower word error rate (WER) than video-based models but still suffer from limited generalization as their speech component is trained from scratch. Moreover, training both video-only and text-video Lip-to-speech systems require over 430 hours of data, making it difficult for future research in low-resource languages.

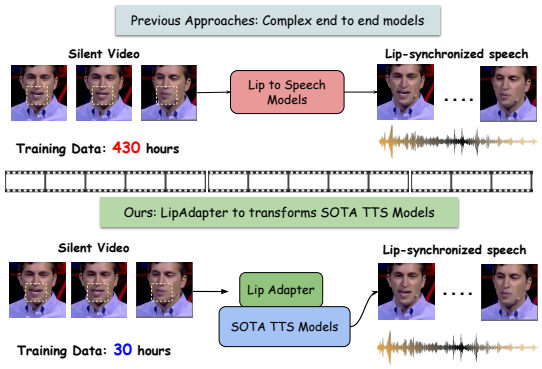


Figure 1: *Previous approaches rely on end-to-end models to map silent video to speech. In contrast, we propose LipAdapter that transforms a SOTA Text to Speech model to generate speech that is temporally synchronized with the speaker's lip movements.*

In contrast, recent TTS models [6, 7] already generalize well across languages, styles and speakers. We propose treating lip-to-speech not as a separate end-to-end problem but as adapting a pretrained TTS through a lightweight LipAdapter network. Our adapter serves as an information bottleneck between the pretrained frozen TTS and VSR modules by introducing a novel Text-to-Video Alignment mechanism between the text embeddings fed into the TTS model and lip embeddings extracted from silent videos. This provides phoneme-level durations aligned with the lip movements, enabling the TTS to generate speech with precise synchronization. Therefore, unlike prior methods that train the full speech components from scratch, we only focus on training the alignment, allowing us to achieve robust performance with less than 30 hours of video data. Our contributions are:

1. We adapt an off-the-shelf frozen pre-trained TTS model for Lip-to-Speech using LipAdapter that utilizes a novel text-to-video monotonic alignment network, which predicts lip-guided phoneme durations enabling the TTS to generate speech that is synchronized with the silent video.
2. Our experiments demonstrate that our proposed method achieves competitive, and in some cases superior, performance in terms of speech quality, word error rate (WER), and lip-sync accuracy on English lip-to-speech tasks, using LRS3 and MultiVSR. It also demonstrates stable zero-shot transfer to unseen languages in multilingual lip-to-speech synthesis, while being trained on only 30 hours of LRS3 videos instead of 430 hours.

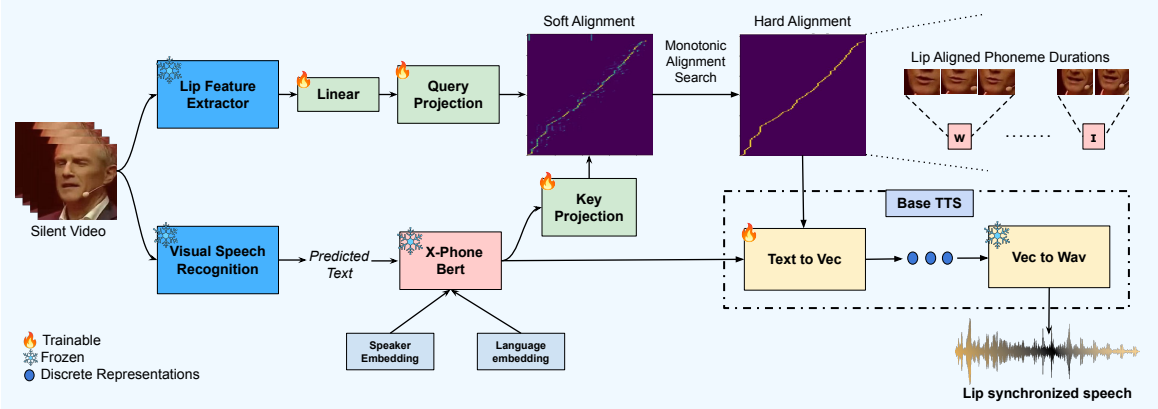


Figure 2: *Given a silent video, we first extract the transcribed text using a pretrained Visual Speech Recognition model and obtain lip embeddings from a pretrained lip feature extractor. A text-to-video alignment network is then trained to compute the soft alignment between phoneme representations and lip embeddings. Monotonic Alignment Search is applied to soft alignment to derive hard phoneme durations aligned with video frames. Finally, the lip-aligned phoneme representations are passed to the base TTS model to synthesize speech that is temporally synchronized with the video. Here, the green blocks represent LipAdapter modules while the yellow blocks show the base TTS model.*

## 2. LipAdapter

As discussed, we solve lip-to-speech generation using an alignment between the text and visual embeddings. The alignment energy between the visual $t$ and phoneme $s$ features is computed as:

$$e_{t,s} = -\tau \sum_{d=1}^{d_a} (\mathbf{q}_t^{\text{enc}}[d] - \mathbf{k}_s^{\text{enc}}[d])^2 \tag{1}$$

where $\tau$ is the temperature and $\mathbf{q}_t^{\text{enc}}$ and $\mathbf{k}_s^{\text{enc}}$ denote the encoded visual and phoneme features, respectively. These energies are converted to soft alignment after incorporating an external alignment prior $P$ to encourage diagonal alignment:

$$A_{t,s} = \text{softmax}\left(e_{t,s} + \log(P_{t,s} + \epsilon)\right) \tag{2}$$

Although soft alignment $A$ provides differentiability for training, we require a hard monotonic path for duration extraction. We employ Monotonic Alignment Search (MAS), a dynamic programming algorithm that finds the optimal monotonic path through the attention matrix. The algorithm computes cumulative log probabilities as follows:

$$\log p(i,j) = \log A_{i,j} + \max \begin{cases} \log p(i-1, j-1) & \text{(advance)} \\ \log p(i-1, j) & \text{(stay)} \end{cases} \tag{3}$$

with initialization $\log p(0,0) = \log A_{0,0}$ and $\log p(0,j) = -\infty$ for $j > 0$. Backtracking from the final position yields a binary alignment matrix $\hat{A} \in \{0,1\}^{T_v \times T_t}$ where exactly one entry per row equals 1. The duration of each phoneme is computed as $d_s = \sum_{t=1}^{T_v} \hat{A}_{t,s}$, ensuring temporal consistency with $\sum_{s=1}^{T_t} d_s = T_v$.

We implement the alignment using the pipeline shown in Fig. 2. A silent video $\mathbf{x} \in \mathbb{R}^{T \times H \times W \times 3}$ with $T$ frames is processed by the pretrained VTP feature extractor [8] to produce lip embeddings $\in \mathbb{R}^{T \times D}$. Phoneme embeddings are obtained from XPhoneBERT [9], a multilingual BERT-based model. Lip embeddings are linearly projected to the phoneme dimension and upsampled to match the temporal resolution of the speech (for $n$ video frames the speech yields $2n-1$ Wav2Vec2 codebook indices). The upsampled visual features $Q = [\mathbf{q}_1, \ldots, \mathbf{q}_{T_v}] \in \mathbb{R}^{T_v \times d_v}$ and phoneme features $K = [\mathbf{k}_1, \ldots, \mathbf{k}_{T_t}] \in \mathbb{R}^{T_t \times d_t}$ are then projected into a shared attention space via the key and query projection networks consisting of three non linear convolutional blocks : $\mathbf{q}_t^{\text{enc}} = f_q(\mathbf{q}_t)$ and $\mathbf{k}_s^{\text{enc}} = f_k(\mathbf{k}_s)$. The alignment energy is subsequently computed between these encoded representations.

We employed a novel two-stage training approach to ensure stable training and differentiability. For the first 20K training steps, we use soft alignment directly to create a differentiable alignment, $K^{\text{soft-aligned}} = A_{\text{soft}} \cdot X$. where $A_{\text{soft}} \in \mathbb{R}^{T_v \times T_t}$ represents the soft alignment matrix, $X$ denotes the phoneme embedding and the matrix multiplication produces temporally weighted text embeddings $K^{\text{soft-aligned}} \in \mathbb{R}^{T_v \times d_t}$. This soft alignment maintains full differentiability, allowing audio reconstruction losses to backpropagate through the attention mechanism and optimize the alignment parameters. After 20K steps, we transition to using the hard attention durations $d_s$ with the Length Regulator for discrete alignment:

$$K^{\text{hard-aligned}} = \text{concat}(\text{repeat}(\mathbf{x}_1, d_1), \ldots, \text{repeat}(\mathbf{x}_{T_t}, d_{T_t})) \tag{4}$$

These temporally regulated phoneme representations serve as inputs to ZMM TTS, a state-of-the-art multilingual TTS model. It comprises of two modules: Text2Vec and Vec2Wav. Text2Vec converts the aligned phoneme representations into a sequence of discrete code indices $V$ derived from a pretrained multilingual wav2vec 2.0 (XLSR-53) model. Subsequently, these indices are mapped through the XLSR-53 codebook $C$ to obtain the corresponding discrete representations $R$, which serve as inputs to the Vec2Wav module which transforms the discrete representations into an audio waveform. Thus, our novel text-to-video alignment network effectively serves as a lip-guided phoneme duration predictor, enabling TTS to synthesize speech that is both content-accurate and temporally synchronized with the visual input. We train the alignment module using a monotonic alignment likelihood objective [12] inspired by CTC [13], together with a path-concentration regularizer that sharpens soft attention along the MAS-derived monotonic trajectory. To better adapt the model, we also lightly fine-tuned the Text2Vec module along with the text-to-video aligner network. However,

| Dataset | Method | Input | Training Hours | WER ↓ | STOI-Net ↑ | DNSMOS ↑ | LSE-C ↑ | LSE-D ↓ |
|---|---|---|---|---|---|---|---|---|
| LRS3 | VCA-GAN [10] | V | 500 | 90.6% | 0.63 | 2.27 | 4.54 | 9.63 |
| | Multitask L2S [11] | | 296 | 57.4% | 0.67 | 2.36 | 5.19 | 8.89 |
| | Intelligible [3] | | 430 | 29.8% | 0.90 | 2.85 | 7.732 | **6.761** |
| | Diff V2S [2] | | 430 | 39.2% | 0.931 | 3.13 | 7.019 | 7.463 |
| | V2SFlow [1] | | 430 | 28.5% | **0.948** | **3.24** | 7.752 | 7.015 |
| | Accurate L2S [5] | V+T | 660 | 23.1% | 0.77 | 2.79 | **7.886** | 6.850 |
| | LipVoicer [4] | | 430 | 21.4% | 0.92 | 3.11 | 5.803 | 8.637 |
| | **LipAdapter (Ours)** | | **30** | **21.2%** | 0.944 | 3.13 | 6.165 | 8.250 |
| MultiVSR | V2SFlow [1] | V | 430 | 61.09% | 0.924 | **3.42** | **8.272** | **7.165** |
| | LipVoicer [4] | V+T | 430 | 35.95% | 0.922 | 3.14 | 7.105 | 8.141 |
| | **LipAdapter (Ours)** | V+T | **30** | **34.44%** | **0.928** | 3.27 | 7.145 | 8.126 |

Table 1: *Performance comparison of LipAdapter on LRS3 and MultiVSR. V in the input modality represents video-only input, while V+T includes visual and predicted text as input.*

| Method | Intelligibility | Naturalness | Sync |
|---|---|---|---|
| Ground Truth | 4.8 | 4.5 | 4.6 |
| Accurate L2S [5] | 3.6 | 3.5 | 4.0 |
| Intelligible [3] | 3.4 | 3.7 | 3.8 |
| DiffV2S [2] | 3.9 | 4.0 | 4.2 |
| VS2Flow [1] | 4.2 | **4.3** | **4.4** |
| LipVoicer [4] | 4.0 | 4.1 | 4.1 |
| **LipAdapter(Ours)** | **4.3** | 4.2 | 4.2 |

Table 2: *Mean Opinion Score (MOS) based on human evaluations of different methods on the LRS3 test set.*

| Language | WER ↓ | STOI ↑ | DNSMOS ↑ | LSE-C ↑ | LSE-D ↓ |
|---|---|---|---|---|---|
| French | 108.4 / **62.09** | 0.71 / **0.94** | 2.85 / **3.85** | 3.42 / **4.16** | 12.8 / **10.17** |
| German | 101.6 / **64.60** | 0.69 / **0.92** | 2.74 / **3.70** | 3.98 / **4.83** | 12.1 / **9.82** |
| Portuguese | 112.3 / **61.69** | 0.66 / **0.90** | 2.81 / **3.74** | 4.02 / **5.18** | 12.6 / **9.52** |
| Spanish | 103.7 / **68.36** | 0.68 / **0.90** | 2.77 / **3.68** | 3.85 / **4.70** | 12.3 / **9.71** |

Table 3: *Zero-shot multilingual comparison (LipVoicer / LipAdapter).*

we show that we can choose to train only the alignment alone and still generate results close to the joint training.

# 3. Experimental Setup

**Dataset:** LipAdapter is compared against the baselines on LRS3 [14], a highly challenging in-the-wild English dataset. To evaluate cross-dataset generalizability, we test LipAdapter and other state-of-the-art approaches on 500 English samples from MultiVSR [15], a recently released multilingual dataset that includes videos sourced and transcribed from diverse YouTube content, making it particularly suitable for testing out-of-distribution robustness. We use the MultiVSR testset to assess the zero-shot multilingual performance of the model.
**Metrics:** We evaluate the effectiveness of our model using both quantitative metrics (i.e. WER [16], LSE-C and LSE-D [17]) and human subjective assessments via Mean Opinion Scores (MOS). To objectively quantify speech quality and intelligibility, we adopt the same metrics used in LipVoicer, specifically DNSMOS [18] and STOI-Net [19].

# 4. Experimental Results

## 4.1. Baselines

We compare LipAdapter with prior methods that report results on LRS3 [14]. We consider two categories of models: (i) video-only methods, which generate speech directly from silent video input [10, 11, 3, 2, 1], and (ii) video and text methods [5, 4], which incorporate predicted text as additional guidance alongside the visual input. Our approach belongs to the latter category, as we leverage a TTS system guided by alignment with video-derived features. For evaluating V2SFlow [1], DiffV2S [2], and Intelligible L2S [3], we directly use the generated samples available in the V2SFlow GitHub repository[1]. For LipVoicer [4] and Accurate L2S [5], we perform inference on the standard LRS3 testset comprising of 1321 samples.

## 4.2. Results in English

Table 1 summarizes our results. Notably, while prior models typically rely on 430 hours of training data, we achieve comparable and in some cases superior performance with only 30 hours of data. In terms of speech quality and Word Error Rate (WER), our method achieves state-of-the-art results among comparable systems. We also achieve strong lip-synchronization performance, surpassing several previous approaches on LRS3. We observe that video-only models tend to produce good synchronization scores, because their audio output is directly conditioned on the visual speech signal. However, these methods frequently exhibit higher WERs because mapping visemes to speech is inherently ambiguous, leading to mispronunciations or incorrect words.
The most significant challenge for video-only architectures arises when these models are applied to datasets like MultiVSR, which are not seen during training. We take the best-performing model from the Video-only set and the best performing model from the Video+Text set to quantify the results. Table 1 shows that although V2SFlow achieves perfect synchronization and clear speech on this dataset, its WER has increased significantly making the pipeline unsuitable for being used. We notice that text modality greatly helps Lipvoicer to achieve comparable results to us. However even with just 30 hours of data we surpass Lipvoicer in synchronization, speech quality and WER. Although Lipvoicer is the closest alternative in terms of generalizability, its diffusion backbone makes it slow, producing 5 s of audio takes 50 s on an NVIDIA 2080 Ti, compared to 6 s with our pipeline. **Human Evaluation:** We conduct a human subjective study using Mean Opinion Scores (MOS) across three dimensions: naturalness, intelligibility, and lip synchronization. To do this, 50 randomly selected LRS3 test samples of different methods, along with the ground truth are shuffled and rated by 20 annotators on a 1–5 scale. As shown in Table 2, LipAdapter

[1] https://github.com/kaistmm/V2SFlow

| Data | WER | STOI | LSE-C |
|---|---|---|---|
| 3h | 27.4 | 0.905 | 5.574 |
| 10h | 25.2 | 0.914 | 5.782 |
| 30h | **23.6** | **0.923** | **5.922** |

Table 4: *Impact of training data hours.*

| Modules | WER | STOI | LSE-C |
|---|---|---|---|
| Aligner | 23.6 | 0.923 | 5.922 |
| Aligner+Text2Vec | **21.2** | **0.944** | **6.165** |

Table 5: *Impact of trainable modules.*

| Backbone | WER ↓ | LSE-C ↑ | STOI-Net ↑ |
|---|---|---|---|
| ZMM-TTS | 23.6 | 5.9 | 0.92 |
| FastSpeech2 | 31.4 | 4.7 | 0.74 |

Table 6: *Backbone generalization experiment. LipAdapter is evaluated with two different length-regulated TTS backbones.*

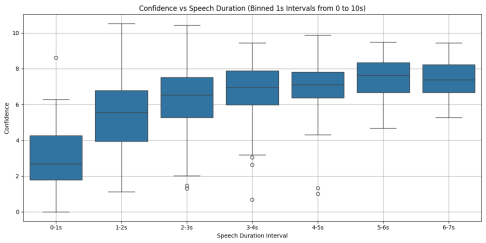


Figure 3: *Box plot showing LSE-C vs the video duration.*

| Language | VSR (%) | ASR (%) | WER (%) |
|---|---|---|---|
| French | 23.6 | 25.72 | 62.09 |
| German | 24.8 | 39.50 | 64.60 |
| Portuguese | 22.9 | 30.01 | 61.69 |
| Spanish | 25.1 | 26.35 | 68.36 |

Table 7: *Error rates of the transcription modules used in multilingual evaluation.*

surpasses the existing methods in intelligibility and achieves results comparable to those of the diffusion-based approaches in naturalness and sync.

### 4.3. Zero-shot multilingual generalization

Our approach further demonstrates the ability to generalize to unseen languages without additional training. As shown in Table 3, our model trained on only 30 h of English videos generalizes without any additional training in French, German, Portuguese and Spanish. These findings indicate that the text-to-video alignment module learns a robust cross-modal mapping between visual speech articulations and their corresponding phonemes that generalize to language combinations not encountered during training. To the best of our knowledge, this is the first lip-to-speech system to demonstrate stable zero-shot multilingual transfer under such limited training data. For comparison, we also perform zero-shot multilingual inference using LipVoicer and report the corresponding results. Despite employing a multilingual tokenizer, LipVoicer exhibits substantially higher WER across all languages. WER exceeds 100% in all cases, indicating that the number of insertion errors in the ASR transcription of the synthesized speech surpasses the number of words in the reference transcript.

### 4.4. Ablation Study

**Impact of Frozen TTS and Backbone Generalization:** As shown in Table 5, freezing the entire TTS backbone and training only the alignment module yields performance comparable to end-to-end training, demonstrating that LipAdapter operates independently of the internal TTS parameters and modifies only phoneme durations. To further validate backbone generality, we replace ZMM-TTS with a FastSpeech2-style length-regulated TTS model trained on XPhoneBERT phoneme embeddings. LipAdapter integrates without architectural modification, supplying only phoneme durations to the new backbone. Although performance decreases relative to the stronger ZMM-TTS setup, largely due to the simpler TTS model and non-neural vocoder, the results confirm that LipAdapter remains structurally compatible with different length-regulated TTS backbones.

**Effect of less training data:** Table 4 shows that our model learns reliable alignments with as little as 3 hours of data, making it well-suited for low-resource or underrepresented languages.

### 4.5. Error Analysis and Limitations

We analyze the principal failure modes of LipAdapter on in-the-wild videos, where errors can accumulate across multiple stages of the pipeline. These include (i) incorrect VSR transcriptions, (ii) text-to-video misalignment and (iii) transcription noise introduced during WER computation due to imperfections in the Automatic Speech Recognition model(Whisper-Large [20]) used for evaluation. Table 7 reports the error breakdown of the VSR and ASR components in multilingual scenarios. The ASR error is computed by running the ASR on the ground truth audio. The high ASR error for these audio files indicates that the measured WER is a conservative estimate, and the true intelligibility of our outputs may be higher.

We further analyze how synchronization confidence varies with input video length. As shown in Figure 3, very short utterances (below 1 s) exhibit noticeably lower LSE-C scores, indicating weaker lip-sync. This degradation is expected, because short clips provide a limited temporal context for the monotonic alignment module to reliably estimate phoneme durations. Performance improves substantially for utterances longer than 2 s, where the alignment becomes more stable and synchronization confidence increases. This suggests that LipAdapter benefits from sufficient temporal evidence and that handling extremely short segments remains an open challenge.

## 5. Conclusion

We introduced LipAdapter, a modular and generalized lip-to-speech framework that connects state-of-the-art visual speech recognition and text-to-speech models via a novel text-to-video alignment module. Experiments on LRS3 and MultiVSR show that LipAdapter delivers strong results in terms of speech quality, intelligibility, and synchronization while using far less training data than prior works. The model is capable of performing zero-shot multilingual lip-to-speech paving the way for future research in low-resource languages. We believe that LipAdapter opens a promising future direction towards video-guided speech generation in accessibility, broadcasting, and human-computer interaction. Additional analyses, including oracle text evaluation, VSR error propagation, alignment robustness, extended ablations and visual results are provided in the supplementary material.

## 6. Acknowledgment

This work is supported by MeitY, Government of India, through the NLTM-Bhashini project.

## 7. Use of Generative AI Disclosure

Generative AI tools were used solely for language refinement and editorial assistance during manuscript preparation. The authors reviewed, verified, and edited all generated content and take full responsibility for the accuracy, originality, and integrity of the work. No generative AI tools were used in the design of the methodology, analysis of results, interpretation of findings, or generation of experimental data.

## 8. References

[1] J. Choi, J.-H. Kim, J. Li, J. S. Chung, and S. Liu, "V2sflow: Video-to-speech generation with speech decomposition and rectified flow," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[2] J. Choi, J. Hong, and Y. M. Ro, "Diffv2s: Diffusion-based video-to-speech synthesis with vision-guided speaker embedding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 7812–7821.

[3] J. Choi, M. Kim, and Y. M. Ro, "Intelligible lip-to-speech synthesis with speech units," *arXiv preprint arXiv:2305.19603*, 2023.

[4] Y. Yemini, A. Shamsian, L. Bracha, S. Gannot, and E. Fetaya, "Lipvoicer: Generating speech from silent videos guided by lip reading," in *The Twelfth International Conference on Learning Representations*, 2024.

[5] S. Hegde, R. Mukhopadhyay, C. Jawahar, and V. Namboodiri, "Towards accurate lip-to-speech synthesis in-the-wild," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 5523–5531.

[6] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, "Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone," in *International conference on machine learning*. PMLR, 2022, pp. 2709–2720.

[7] C. Gong, X. Wang, E. Cooper, D. Wells, L. Wang, J. Dang, K. Richmond, and J. Yamagishi, "Zmm-tts: Zero-shot multilingual and multispeaker speech synthesis conditioned on self-supervised discrete speech representations," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

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

[9] L. T. Nguyen, T. Pham, and D. Q. Nguyen, "Xphonebert: A pre-trained multilingual model for phoneme representations for text-to-speech," *arXiv preprint arXiv:2305.19709*, 2023.

[10] M. Kim, J. Hong, and Y. M. Ro, "Lip to speech synthesis with visual context attentional gan," *Advances in Neural Information Processing Systems*, vol. 34, pp. 2758–2770, 2021.

[11] ——, "Lip-to-speech synthesis in the wild with multi-task learning," in *2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2023, pp. 1–5.

[12] R. Badlani, A. Łańcucki, K. J. Shih, R. Valle, W. Ping, and B. Catanzaro, "One tts alignment to rule them all," in *ICASSP 2022-2022*. IEEE, 2022, pp. 6092–6096.

[13] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd International Conference on Machine Learning*, ser. ICML '06. New York, NY, USA: Association for Computing Machinery, 2006, p. 369–376. [Online]. Available: https://doi.org/10.1145/1143844.1143891

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

[15] K. R. Prajwal, S. Hegde, and A. Zisserman, "Scaling multilingual visual speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[16] P. Ma, S. Petridis, and M. Pantic, "Visual Speech Recognition for Multiple Languages in the Wild," *Nature Machine Intelligence*, vol. 4, pp. 930–939, 2022. [Online]. Available: https://doi.org/10.1038/s42256-022-00550-z

[17] K. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar, "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*, 2020, pp. 484–492.

[18] C. K. Reddy, V. Gopal, and R. Cutler, "Dnsmos p. 835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2022, pp. 886–890.

[19] R. E. Zezario, S.-W. Fu, C.-S. Fuh, Y. Tsao, and H.-M. Wang, "Stoi-net: A deep learning based non-intrusive speech intelligibility assessment model," in *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2020, pp. 482–486.

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