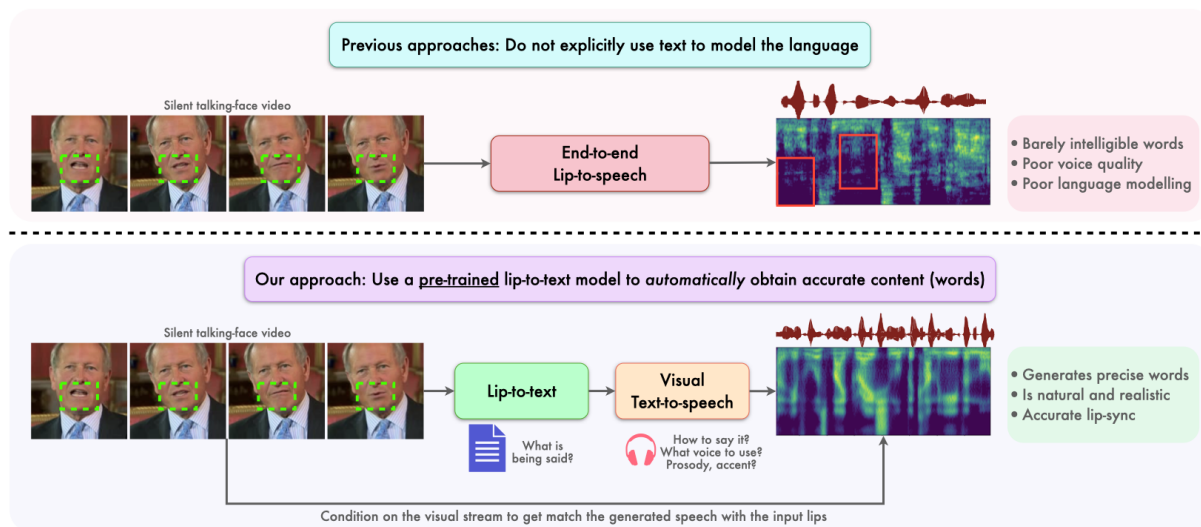# Towards Accurate Lip-to-Speech Synthesis in-the-Wild

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

Rudrabha Mukhopadhyay*
radrabha.m@research.iiit.ac.in
CVIT, IIIT Hyderabad
India

C.V Jawahar
jawahar@iiit.ac.in
CVIT, IIIT Hyderabad
India

Vinay Namboodiri
vpn22@bath.ac.uk
University of Bath
United Kingdom

Figure 1: We propose a novel approach for multi-speaker lip-to-speech synthesis in the wild. Prior works try to learn a language model directly from raw speech, which only provides weak supervision due to the presence of other acoustic variations such as voice, accents, and prosody. We solve this problem by relying on recent advancements in lip-to-text generation. We condition on the noisy text outputs and lip video to generate natural speech with clearly pronounced words.

ABSTRACT

In this paper, we introduce a novel approach to address the task of synthesizing speech from silent videos of any in-the-wild speaker solely based on lip movements. The traditional approach of directly generating speech from lip videos faces the challenge of not being able to learn a robust language model from speech alone, resulting in unsatisfactory outcomes. To overcome this issue, we propose incorporating noisy text supervision using a state-of-the-art lip-to-text network that instills language information into our model. The noisy text is generated using a pre-trained lip-to-text model, enabling our approach to work without text annotations during inference. We design a visual text-to-speech network that utilizes the visual stream to generate accurate speech, which is in-sync with the silent input video. We perform extensive experiments and ablation studies, demonstrating our approach's superiority over the current state-of-the-art methods on various benchmark datasets. Further, we demonstrate an essential practical application of our method in assistive technology by generating speech for an ALS patient who has lost the voice but can make mouth movements. Our demo video, code, and additional details can be found at http://cvit.iiit.ac.in/research/projects/cvit-projects/ms-l2s-itw.

---

*Both authors have contributed equally to this research.
†This work was done while pursuing MS by Research at IIIT Hyderabad.

CCS CONCEPTS

• **Computing methodologies → Neural networks**; **Computer vision tasks**.

KEYWORDS
Lip-reading, Lip-to-Speech, Assistive Technology, Speech Generation

1 INTRODUCTION

As multi-sensory beings, humans have the ability to experience the world through various senses, adapting to make the most of what we perceive. When one sense is diminished, like hearing, the brain compensates by relying more on other senses, like vision, to maintain independence and navigate the environment. One of the most prevalent ways for deaf individuals to comprehend speech is through lip reading - the process of understanding spoken content purely from silent lip movements. Therefore, lip reading has numerous applications, ranging from searching through old silent films to being used as an assistive technology for people who are unable to speak. For example, people with Amyotrophic Lateral Sclerosis (ALS) often lose their ability to speak due to issues with their vocal cords but can mouth words. Therefore, developing automated lip reading technologies can improve the communication abilities of people with ALS. In the past few years, previous works have shown that it is far easier for deep learning models to perform lip reading than it is for humans [7]. There has been a growing interest in the research community [2, 3, 7, 8] to solve the task of transcribing silent lip movements into text. While giant strides are being made to solve lip-to-text generation, only a handful of works go a step further to generate speech from silent lip movements. The widening gap between these two closely related tasks is becoming increasingly evident. In other words, while lip-to-text models are reaching word error rates as low as 17% WER [32], there are no practically usable lip-to-speech systems that generate natural, meaningful speech for in-the-wild identities.

Applications for lip-to-speech synthesis. Lip-to-Speech generation is not only more exciting and challenging but also has more far-reaching applications. For instance, having a live video call with a person who has lost their voice is more engaging than reading the text transcription of their silent lip movements. Speech can convey more: it is possible to express emotions through silent lip movements if it is converted to speech rather than text. Speech is also more instantaneous for the listener: it is easier and faster to hear than to read text. The immersive user experience is a significant impetus for developing algorithms that translates lip movements directly to speech. Lip-to-Speech can also be applied in forensic investigations, where silent footage can be analyzed to determine what was being said or generate speech from archival film footage. Our work showcases the practical impact of lip-to-speech generation by generating speech for the silent lip movements of an ALS patient, a feat not previously demonstrated by other models in this field. We also show results on reading mouth movements of people suffering from hearing loss to demonstrate the effectiveness

of accurate lip-to-speech synthesis further. More details about the applications can be found in Section 4.3.

The challenges in lip-to-speech synthesis. However, synthesizing speech from silent talking-face videos is far more challenging than lip-to-text generation. Not only does the model first needs to distill the content from the lip movements, but it also needs to generate the final speech by accurately modeling the speaker attributes like voice, accent, and style. Owing to these difficulties, until recently, most of the existing works [12, 27, 36] focused on building either single-speaker models or constrained multi-speaker models with limited corpus size as well as limited vocabulary. While more recent efforts like [27] have extended lip-to-speech to "in-the-wild" environments, most of these models are speaker-specific in nature, i.e., they only work on speakers they are trained on. The speaker-independent models [17, 20, 21, 23] suffer from numerous weaknesses and fail to accurately learn language and speech attributes like voice, prosody, etc. Due to the sheer difficulty of the task, they turn out to be well below expectations for an end-user application.

1.1 Our Contributions

First, we would like to clearly state that our goal is to be able to generate speech for a silent lip video of *any* speaker in the wild and in *any* desired target voice. The fundamental premise of this paper is that current models trying to solve this task struggle to learn language and speech attributes because they try to learn both of these solely from speech supervision. They inadvertently disregard the advancements that have been made in the sibling task of the lip-to-text generation. We argue that learning lip-to-speech without any understanding/supervision of text is very challenging to achieve, which can be observed in all the current models, where they fail to produce clear and intelligible outputs in unconstrained settings. Our key idea is to use the fact that lip-to-text models already have learned the task of extracting content from lips, as they are trained with text supervision. We show that we can achieve accurate, natural, and high-quality speech outputs by using the silent video input and the noisy text transcriptions from a pre-trained lip-to-text model. We train a visual text-to-speech model to synthesize speech that syncs with the input video. Our approach outperforms the existing multi-speaker lip-to-speech models by a significant margin in terms of both qualitative and quantitative evaluations. We highlight our key contributions below.

(1) Lip-to-Speech generation is challenging with only speech supervision, and current models struggle to learn language and speech attributes. To address this, we propose using a pre-trained lip-to-text model's output to aid in lip-to-speech generation.

(2) Our approach involves training a visual text-to-speech model to synthesize speech that syncs with the input video. Using this approach, we outperform existing multi-speaker lip-to-speech models in qualitative and quantitative evaluations.

(3) Our proposed approach enables accurate and natural speech output for silent lip videos of any speaker in any desired target voice.

2 RELATED WORKS

Lip-to-Speech. Generating speech directly from lip movements has always been a challenge for researchers due to the ill-posed nature of the task. While recognizing content from lip movements itself is challenging due to the one-to-many relation present between visemes and phonemes, lip-to-speech synthesis also needs to model speech-related variations like voice, accents, and prosody. Therefore, it was initially attempted in laboratory setups [10] using datasets with a minimal vocabulary. The initial works to [4, 11, 12, 36] used fully convolutional neural networks to learn a mapping between lip movements and speech. These models were also trained on speaker-specific data and did not work for unseen speakers. Moreover, such models were incapable of handling "in-the-wild" settings. To mitigate this issue, a sequence-to-sequence model, Lip2Wav [27], was proposed that was trained on large amounts of speaker-specific data taken from in-the-wild YouTube videos. Lip2Wav worked reasonably well on speakers seen during training but did not extend to handling unseen speakers. Several more recent studies [17, 20, 21, 23] have attempted to develop end-to-end lip-to-speech synthesis models on large datasets containing data from hundreds of speakers. For instance, in [17], authors proposed a variational approach that matches the distributions of lip movements and speech segments to project them into a shared space, which allows for handling the high variations of in-the-wild speakers to some extent. Meanwhile, both [20, 23] utilized a transformer-based approach to convert lip-to-speech synthesis into a sequence-to-sequence problem, where a sequence of lip movements is translated into a sequence of speech tokens. Additionally, [21] treated lip-to-speech synthesis as a multi-tasking problem, where the model predicts text transcripts in addition to speech forcing the model to learn better content information. However, all of these models produce sub-par outputs with words that are barely uttered and often unintelligible with unnatural voice and prosody. We argue that these works are limited because they do not use the recent advancements in the sibling task, lip-to-text, which we discuss below.

Lip-to-Text. Lip reading has become synonymous with lip-to-text, and several advancements have been made to achieve highly accurate text transcriptions of silent lip videos. The very first works [5] used datasets [10, 15] collected in a laboratory setup and trained simple RNN-based architectures. One of the initial approaches [8] handling the "in-the-wild" setup posed the problem in a word-level classification setting and continued using RNN-based models. Subsequent approaches [7] have attempted to tackle this task by posing it as a sequence-to-sequence problem. The goal here is to translate a sequence of lip movements into a sequence of characters. Approaches like [2] compared multiple sequence-to-sequence architectures, including transformers, to achieve favorable word error rates (WER). A more recent work [26] uses a variant of transformer and specific attention to the mouth region through visual transformer pooling and achieves state-of-the-art results in lip reading. Other works such as AV-HuBERT [34] have also improved the state-of-the-art for in-the-wild lip reading, paving the way for practical applications. Our work demonstrates the importance of pre-trained lip-to-text models in achieving precise lip-to-speech synthesis in uncontrolled environments.

Text-to-speech. Our proposed approach uses text as an intermediate output and converts it into speech. Text-to-Speech has long interested the speech community and has historically seen a lot of progress. Modern deep learning architectures [25, 30, 33] enable several industrial and multimedia applications. They are accurate in terms of uttered content and have natural speaking styles, prosody, and voice. The efforts slowly shifted towards multi-speaker scenario, which is far more challenging due to the sheer variation in prosody, voice, and other speaker-related attributes like pitch and tone. Most of these methods are provided with a separate voice token [19] containing identity information about the target speaker. The voice tokens are generated from a short speech segment of a target speaker and often only capture the voice information while failing to capture the prosody. Thus, there has also been a push for using additional information like lip movements [16, 18] to help TTS models generate higher-quality outputs. Since lip-to-speech synthesis requires us to generate accurate prosody and speaking style, our approach is inspired by this line of work. We design a visual text-to-speech model that works uses both text and visual features from state-of-the-art lip reading network [27] to generate natural, accurate speech. Our Visual TTS is integral to generating speech that perfectly syncs with the input silent lip video.

3 LIP-TO-SPEECH SYNTHESIS IN THE WILD

We start the discussion by highlighting the key issues and challenges in the existing lip-to-speech works, leading to their poor performance in unconstrained settings. We then present our framework with a detailed description of the modules involved.

3.1 Issues and challenges in existing works

3.1.1 Learning language from speech. As discussed previously, all the current lip-to-speech works directly generate speech from the lips. It is known that learning a language model is crucial for reading the lips accurately. However, the current multi-speaker lip-to-speech models are sub-par because they try to learn a language model in the speech modality, which contains a large diversity of speaker identities, styles, accents, and prosody. Thus, we argue that we need some other way of incorporating language knowledge.

3.1.2 The missing block of lip-to-speech: lip-to-text. Lip-to-Speech synthesis models have two tasks: (i) inferring the content from lips and (ii) inferring the style in which that content is spoken. If we have the content that is being spoken, then our task is now reduced to just generating the speech that matches the silent lip video. This is the premise of our paper. But how do we get this text information, especially when we only have a silent lip video as input? We show that we can get this text information from pre-trained lip-to-text models: a class of models closely related to our task at hand but have been mainly ignored in previous works on lip-to-speech synthesis. We design an approach that can build upon the current works [2, 26] in lip reading, i.e., pre-trained lip-to-text models, and generate far more accurate speech outputs.

3.1.3 Achieving accurate lip-sync. Now that we have the text, the next step is to generate the "right" kind of speech output. That is, the generated speech must match the input lip sequence. There are many ways to utter the same sentence, but only one will match the
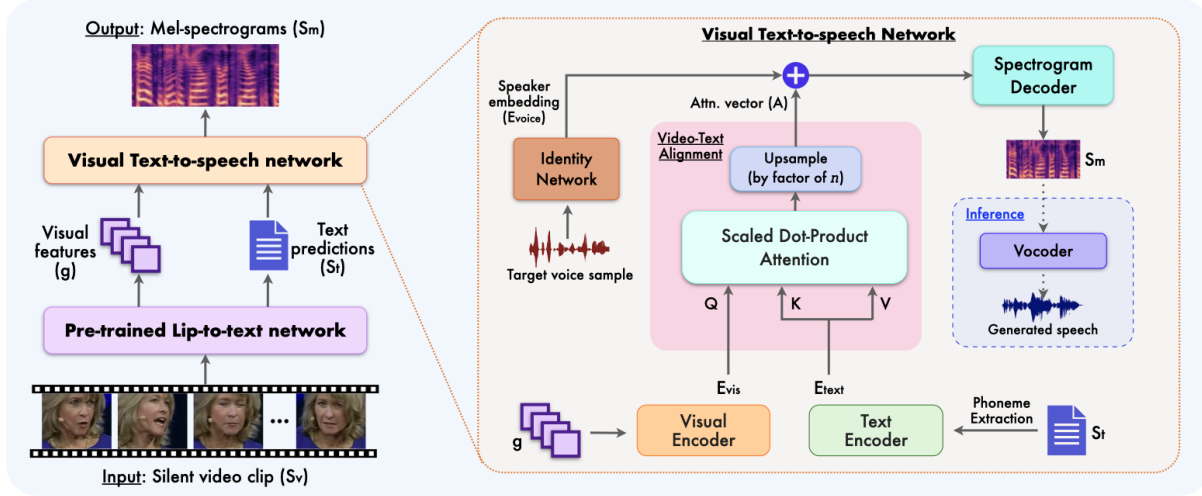
Figure 2: Overview of our approach. We first extract the visual features and the text predictions from a pre-trained lip-to-text network. Using a visual text-to-speech (TTS) model, we can generate speech outputs that sync with the silent video input. The visual TTS encodes the visual and textual (in the form of phonemes) inputs and aligns them in time using the scaled dot-product attention. For each query video time-step, we retrieve the phoneme to utter at that time using this attention mechanism. After adding the speaker identity embedding, these are then upsampled and decoded into melspectrograms. The melspectrograms are converted into natural waveforms using a pre-trained vocoder.

input lip video. Thus, it is worth noting that a trivial text-to-speech will not serve our task. Instead, we need a text-to-speech model that is also conditioned on video input.

3.2 Our Approach

The task of lip-to-speech synthesis is formulated as follows: given a silent video frame sequence of a person talking $S_v = \{I_1, I_2, ..., I_p\}$, the goal is to generate the corresponding speech melspectrogram sequence, $S_m = \{M_1, M_2, ..., M_q\}$. An overview of our proposed two-stage framework is depicted in Figure 2. We use a state-of-the-art lip-to-text model [26] to obtain the lip features and noisy text transcriptions for the given silent lip video S_v. We design a visual text-to-speech model conditioned on (i) the noisy text and (ii) the lip features to produce high-quality speech outputs S_m that are in sync with the input silent lip video. This solves the task of lip-to-speech synthesis.

3.2.1 Pre-trained Lip-to-Text.
Lip-to-Text networks ingest the silent talking face videos as input and transcribe what is being uttered. The lip-to-text models [1–3, 7, 8, 26] are now capable of generating text transcripts with word error rates as low as $20 - 30\%$. One such recent work trained on public datasets is VTP [26]. Specifically, this work stands out due to its two unique aspects: (i) it has been shown to be data efficient, (ii) it contains a strong visual backbone that can attend to and extract accurate lip features. These visual features have been shown to work well in several other tasks such as visual keyword spotting [29] and even spotting mouth movements in sign language [24]. The sub-word units considered to learn the text representations help to model the ambiguities of the task significantly better than using characters. Also, the sub-word tokens are semantically meaningful and provide a language prior, thereby resulting in a significant performance boost. In addition, the visual

representations learned using the backbone network track and aggregate the spatio-temporal lip movement features, mainly due to the strong attention-based pooling mechanism. We thus propose to adopt this model as our first module to generate text and visual features from silent lip videos.

Overview of VTP [26]: The model inputs a sequence of 5 consecutive frames, $S_v \in \mathbb{R}^{T \times H \times W \times 3}$ ($T = 5$) and extracts low-level features using a spatio-temporal residual CNN block. These individual frame-wise features are then processed using the visual transformer pooling block, which consists of a series of transformer layers. The output from this block, z_t is a self-attended feature map, which is used along with a learnable query vector, Q_{att} to obtain the spatially weighted average of z_t. These compact per-frame visual representations are stacked along the temporal dimension to obtain a temporal embedding sequence, $g \in \mathbb{R}^{T \times fd}$, where fd is the transformer feature dimension. Finally, the learned visual representations are passed to the transformer encoder-decoder network to predict the text outputs in the form of sub-word tokens in an auto-regressive fashion. The model uses beam search decoding [13] and language model rescoring [6] strategies to obtain the final sentence outputs S_t. We refer the reader to [26] for more details on the architecture and training strategies.

Adopting VTP for our task: For our task at hand, we use the pre-trained VTP network to obtain: (i) text predictions - the final decoded output of the model S_t, and (ii) per-frame visual representations - g. The text predictions directly act as input to our speech generation module. The visual representations serve as a condition for speech generation, which is crucial to obtain speech that is in-sync with the silent input video.

3.2.2 Visual Text-to-Speech.
Once we have the accurate text predictions, the next step is to generate the corresponding speech

sequence S_m, which is in-sync with the input video clip S_v. As explained in Section 3.1.3, directly using state-of-the-art TTS models to synthesize speech from text inputs will generate out-of-sync speech that does not match the input video. We thus design a TTS network by conditioning the model on the input video features. Our visual TTS network majorly comprises five components: (i) Text Encoder, (ii) Visual Encoder, (iii) Visual-Text Attention, (iv) Speaker Embedding, and (v) Spectrogram Decoder. We delve into each of these components below.

Text Encoder: As followed in most of the TTS networks [30], we extract phoneme representations from text input, which is then given as input to the Transformer encoder layers. Our text encoder block is similar to the one used in FastSpeech2 [30], which consists of a positional encoding layer and Feed-Forward Transformer (FFT) layers. The phonemes are transformed to encode the semantic representation and output the text embedding vectors, E_{text} of dimension: $N \times d$, where d is the transformer feature dimension.

Visual Encoder: The input to our visual encoder is the visual feature sequence g obtained from the lip-to-text network. We highlight that these visual features, which capture the lip shape and motion, play a crucial role in generating speech that syncs with the input video. Since these representations were also learned using text supervision, they are likely to reflect accurate content information. This starkly contrasts previous works that directly learn visual representations from speech supervision only, which can lead to sub-par visual representations that might contain other unnecessary information, such as the input face identity. The superiority of these representations is one of the critical reasons for our overall network's performance. The extracted $N \times T \times fd$ dimensional representations are given as input to the Transformer encoder layers as shown in Figure 2. Similar to the text encoder network, the visual encoder consists of a positional encoding followed by a series of FFT blocks. The encoder network outputs the learned visual embeddings, E_{vis} of dimension: $N \times T \times d$.

Visual-Text Attention: Once we obtain the text and the visual embeddings, the next and the most important step is to find the alignment between these embeddings in time: which phoneme must be uttered when? The generated speech must take the content from the text embeddings and simultaneously, it should also temporally align (sync) it with the video frames. In order to achieve this, we employ a scaled-dot product attention [35] mechanism to learn the correspondence between text and video frames. Specifically, the visual embeddings E_{vis} act as query, and the text embeddings E_{text} act as keys and values.

$$Attention(Q, K, V) = Scaled - Dot\ Product$$

$$Attention(E_{vis}, E_{text}, E_{text}) \in \mathbb{R}^{T \times d} \quad (1)$$

Through this attention module, the network learns the video-text temporal alignment, which synchronizes the generated speech with the input video frames. Now between the video sequence and melspectrograms, it is known that there exists a natural temporal alignment. The length of the melspectrograms is a constant n times the length of the video. Thus, the attention output A is up-sampled n times to directly obtain the melspectrogram duration. This eliminates the need to train a separate duration predictor as done in

FastSpeech2 [30]. In other words, the text-to-video alignment network already determines the duration of each phoneme in the speech output.

Speaker Embedding: Unlike single-speaker models, we want our model to generate speech for any arbitrary speaker in the wild. Thus, our model also needs the voice input of the target speaker to generate the speech in his/her voice. We consider a random one-second audio segment of the target speaker and extract the speaker embedding vector E_{voice} using the pre-trained identity network[1].

Spectrogram Decoder: The speaker embedding vector E_{voice} is added to the upsampled attention output A to obtain voice-aware content representation. The spectrogram decoder, consisting of transformer decoder layers, ingests this representation and generates the melspectrogram sequence S_m. To further improve speech quality, as done in most of the TTS networks, we adopt a pre-trained neural vocoder model BigVGAN [14], to synthesize the speech from the melspectrogram output. Note that this step is only used during inference to obtain high-quality speech outputs.

3.3 Datasets and Training Settings

3.3.1 Datasets. We evaluate both constrained and unconstrained datasets to analyze the model's performance. The first corpus we experiment with is the TCD-TIMIT [15] lip speaker dataset, which comprises lab-recorded videos of 3 speakers. Next, we consider the word-level LRW [8] dataset, consisting of around 150 hours of single-word utterances from hundreds of speakers. We then move on to the more challenging large-scale datasets: LRS2 [7] and LRS3 [3]. The LRS2 data comprises thousands of speakers from BBC programs with a vocabulary of $59k$ and around 230 hours of video clips (both "train" and "pre-train" sets together). The LRS3 dataset, on the other hand, is also a large-scale dataset with a total of approximately 430 hours ("train" and "pre-train" sets) of video data with $150k$ utterances. It consists of thousands of spoken sentences from TED and TEDx talks in English. We train and test the performance of our network using the official splits of LRW, LRS2 and LRS3 datasets and use the train-test split proposed in Lip2Wav [27] for TIMIT dataset [15].

3.3.2 Data pre-processing. We sample the video frames at 25 FPS and follow the pre-processing procedure of VTP [26] to obtain the face crops. For the speech segments, we compute STFT and then melspectrograms of 80 mel-bands, with a hop length of $10ms$ and a window length of 25ms, sampled at 16kHz. We use an open-source grapheme-to-phoneme tool for text processing to obtain the phoneme inputs for our Visual TTS model.

3.3.3 Model configuration and training. Our Visual TTS model comprises 4 FFT blocks in the text and visual encoders and 6 FFT blocks in the spectrogram decoder network. The visual embeddings obtained from VTP are 512-dimensional embeddings for each frame. In the video-text attention sub-network, the upsample factor, n is set to 4. For the speaker embedding, the identity network outputs a 256 dimensional vector for each speech sample. For the lip-to-text network, we use the publicly released pre-trained model[2] (trained

---

[1]github.com/CorentinJ/Real-Time-Voice-Cloning
[2]https://github.com/prajwalkr/vtp

on LRS2 [7] and LRS3 [3]). Our visual TTS model is trained on a single NVIDIA 2080 Ti GPU. We use the Adam optimizer [22] with $\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\epsilon = 10^{-9}$ and follow the same learning rate schedule as done in [26]. We set the batch size to be 16 for all the datasets and train the model using L_1 reconstruction loss for approximately $900k$ steps (until convergence). We also train BigVGAN vocoder [14] and use it during inference to generate the speech from the output melspectrograms.

4 EXPERIMENTS

We present our approach's quantitative results and comparisons with the existing methods. As automatic speech metrics are imperfect, we also show MOS scores using human evaluation. Finally, we demonstrate a real-world application of lip-to-speech for the first time by voicing the silent lip movements of an ALS patient.

4.1 Quantitative Evaluations

Metrics: We measure the quality of the generated speech using the standard speech metrics: Perceptual Evaluation of Speech Quality (PESQ), Short-Time Objective Intelligibility measure (STOI), and its extended version (ESTOI). PESQ measures the clarity and overall perceptual quality of speech and STOI and ESTOI measure the intelligibility of speech. Further, as discussed previously, it is very crucial to generate speech that is in sync with the input video. We use the lip-sync metrics, Lip-Sync Error - Confidence (LSE-C) and Lip-Sync Error - Distance (LSE-D) [9] to evaluate whether the output speech matches the input lip movements. We use the public implementations of all the above metrics for a fair comparison.

4.1.1 Speech Synthesis in Constrained Settings.

Comparisons: To evaluate lip-to-speech methods on the constrained single-speaker TCD-TIMIT dataset, we compare four existing approaches: (i) GAN-based [36], (ii) Lip2Wav [27], (iii) VAE-GAN [17], and (iv) VCA-GAN [20]. We adopt the same settings as Lip2Wav [27] and report the scores [27] and VCA-GAN [20]. We use the Wav2Lip [28] repository to compute the LSE-C and LSE-D scores for each method. We have excluded the metrics for a particular model that were not mentioned in the original papers or for which no publicly available pre-trained checkpoint exists.

Results: Table 1 contains the results on the TCD-TIMIT dataset. We observe that our approach achieves comparable results to previous methods in constrained settings with minimal data of only 3 speakers. However, the significant benefits of our approach can be seen in unconstrained settings, which we describe below.

4.1.2 Speech Synthesis in Unconstrained Settings.

Comparisons: In order to assess the performance of lip-to-speech methods in unconstrained scenarios, we employ three datasets: word-level LRW [8], sentence-level LRS2 [7], and LRS3 [3]. While the authors of VAE-GAN [17] have re-trained the GAN-based [36] and Lip2Wav [27] models in a multi-speaker context, we present the scores from their original study for comparison. For VCA-GAN [20], SVTS [23], and Multi-task Lip-to-Speech synthesis [21], we adopt the speech metric (PESQ, STOI and ESTOI) scores from [21]. Further, we use publicly accessible pre-trained checkpoints for VCA-GAN[3]

and Multitask-L2S[4] to generate speech on LRS2 and LRS3 test sets for the former, and LRS2, LRS3, and LRW test sets for the latter. We utilize these generations to compute the LSE metrics for both techniques, wherever applicable. Lastly, we include results for a baseline approach that leverages lip-to-text conversion followed by multi-speaker TTS without a visual stream in the TTS model. Our evaluation does not include all metrics that were not originally reported in the papers or for which no pre-trained model is publicly available.

Results: We present the results on the challenging LRW, LRS2, and LRS3 datasets in Table 1. Our model consistently outperforms the existing methods by a significant margin on all these datasets. Since GAN-based [36] model was proposed to work for constrained laboratory recorded datasets, we can observe that extending this model in unconstrained settings does not yield satisfactory results. Lip2Wav [27] performs decently on the word-level LRW dataset; however, it fails to learn the audio-visual alignment on the LRS2 dataset, thus leading to very poor performance. We discard this model for further comparison on the LRS3 dataset. VAE-GAN [17], VCA-GAN [20], SVTS [23] and Multitask-L2S [21] generate speech that is in-sync with the input video; however, they fail to synthesize accurate content. The quality of the generated speech is often non-intelligible and leads to lower scores in speech quality metrics. The Lip-to-Text + TTS baseline model is on the opposite spectrum, where the model generates the content well but fails to capture the lip-sync, mainly because the model cannot infer the speed, prosody, and accents of speakers just from the text input. Our model, on the other hand, is capable of generating both the actual spoken content as well as maintaining precise lip synchronization. As we can see from the table, we outperform the previous methods in all the speech quality metrics, indicating the robustness and superiority of our approach. In Figure 3, we depict how our model temporally aligns video and text sequences in the process of generating speech. We encourage the reader to view our demo video comprising multiple qualitative samples and comparisons.
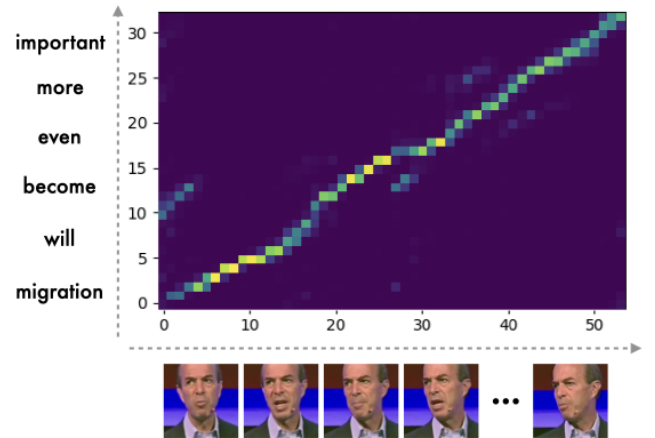
Figure 3: We visualize the video-text alignment from the scaled dot product attention step of our model. We observe that the model learns a strong monotonic near-diagonal attention, as expected.

Table 1: We compare against state-of-the-art methods on several standard multi-speaker benchmarks using standard metrics. The generated outputs from our model are most natural (PESQ), most accurate (STOI, ESTOI) and in perfect sync with the video input (LSE-C, LSE-D) in the in-the-wild videos of LRW [8], LRS2 [7] and LRS3 [3].

Dataset	Method	PESQ↑	STOI↑	ESTOI↑	LSE-C↑	LSE-D↓
TCD-TIMIT [15]	GAN-based [36]	1.22	0.51	0.32	-	-
	Lip2Wav [27]	1.35	0.56	0.36	6.610	7.815
	VCA-GAN [20]	**1.43**	0.58	0.40	-	-
	VAE-GAN [17]	1.35	0.55	0.35	-	-
	Ours	1.34	**0.61**	**0.42**	**6.623**	**6.901**
LRW [8]	GAN-based [36]	0.72	0.10	0.02	1.983	9.426
	Lip2Wav [27]	1.19	0.54	0.34	2.526	8.286
	VAE-GAN [17]	0.78	0.15	0.03	2.538	8.173
	VCA-GAN [20]	1.33	0.56	0.36	-	-
	SVTS [23]	1.49	0.64	0.48	-	-
	Multi-task L2S [21]	1.56	0.64	0.47	4.876	8.102
	Lip-to-Text + TTS baseline	0.69	0.10	0.01	1.993	12.872
	Ours	**1.61**	**0.71**	**0.56**	**6.812**	**6.974**
LRS2 [7]	Lip2Wav [27]	0.58	0.28	0.11	1.874	11.48
	VAE-GAN [36]	0.60	0.34	0.17	2.507	8.155
	VCA-GAN [20]	1.24	0.40	0.13	4.016	7.914
	SVTS [23]	1.34	0.49	0.29	-	-
	Multi-task L2S [21]	1.36	0.52	0.34	4.001	8.192
	Lip-to-Text + TTS baseline	0.53	0.19	0.02	2.013	15.891
	Ours	**1.47**	**0.65**	**0.47**	**8.083**	**6.586**
LRS3 [3]	VAE-GAN [36]	0.51	0.30	0.15	2.063	8.256
	VCA-GAN [20]	1.23	0.47	0.20	3.905	8.392
	SVTS [23]	1.25	0.50	0.27	-	-
	Multi-task L2S [21]	1.31	0.48	0.26	3.876	8.677
	Lip-to-Text + TTS baseline	0.42	0.16	0.01	1.771	17.882
	Ours	**1.39**	**0.58**	**0.37**	**7.886**	**6.850**

4.2 Human Evaluations

To evaluate the applicability of our method in real-world scenarios, we conduct subjective human evaluations. We ask 25 volunteers to assess the quality of speech generations. The participant group has an almost equal male-female ratio, spanning an age group of $20-45$ years. We randomly select 10 long sentences (10 seconds or longer) from the test set of LRS3 [3] and present the results from different methods to the participants. We ask them to rate the samples on a scale of $1-5$ based on the following criteria: (A) Intelligibility (is the speech meaningful?), (B) Content clarity (are the words clear?), (C) Sync Accuracy, (D) Overall perceptual quality of the talking head video + audio. We report the mean opinion scores in Table 2. On par with the quantitative evaluations, our method is highly rated over other approaches in all the criteria listed above. As expected, the speech intelligibility is slightly rated higher for the Lip-to-Text + TTS baseline. However, the overall perceptual quality for this baseline sharply falls due to the lack of sync between the spoken content and the lip movements. Overall, Table 2 clearly signifies that our network is able to generate speech with more clarity, which sounds more natural and is of considerably higher quality.

4.3 Applications in Assistive Technology

Lip-to-Speech synthesis has a host of applications in a world that is becoming increasingly digital. Simple applications such as performing video calls in quiet environments, filling in the audio interruptions due to technical issues, eliminating unwanted background

Table 2: (A) Intelligibility, (B) Content clarity, (C) Sync Accuracy, (D) Overall perceptual quality. Our model produces natural and realistic speech outputs that is largely preferred by the users in comparison to other approaches.

Method	(A)	(B)	(C)	(D)
GAN-based [36]	2.05	1.87	1.99	2.12
Lip2Wav [27]	1.01	1.03	1.34	1.01
VAE-GAN [17]	1.07	1.33	2.18	2.57
VCA-GAN [20]	2.18	1.88	2.97	2.54
Multi-task L2S [21]	2.19	1.85	3.01	2.64
Lip-to-Text + TTS baseline	**3.61**	2.87	1.01	2.96
Ours	3.49	**3.52**	**3.82**	**3.31**

chatter, etc., can be made possible with accurate lip-to-speech. We believe the most significant application of lip-to-speech can be in assistive technologies. It can revolutionize the current assistive systems used to improve the communication ability of people suffering from various disorders affecting their speech. Patients suffering from vocal cord disabilities can mouth words to communicate naturally with the world around them. The synthesized speech can be personalized and also be in sync with the speaker's lip movements.

Generating Speech for a Patient suffering from ALS:. A recent work [31] proposed a lip reading technique for a patient suffering from ALS. The patient has feeble vocal cord movements but can mouth words silently. The authors of the paper collected limited

amounts of data from the patient and trained a model to recognize words and sentences from his lip movements. As a result of our significant improvement in this task, we can now demonstrate the benefit of using lip-to-speech as a future assistive technology. Through the help of authors of [31], we evaluate our lip-to-speech system on the patients' data. Please note that the data was anonymized and was used only for research purposes. We find that the lip-to-text module generates fairly accurate text (WER of $\approx$ 37%), and the visual TTS model generates clear speech in sync with the patient's lip movements. This is the first demonstration of automatic lip-to-speech synthesis for an unseen speaker in an entirely out-of-domain real-world application. Further, we also test our model on the other deaf speakers studied in [31] and observe accurate performance.
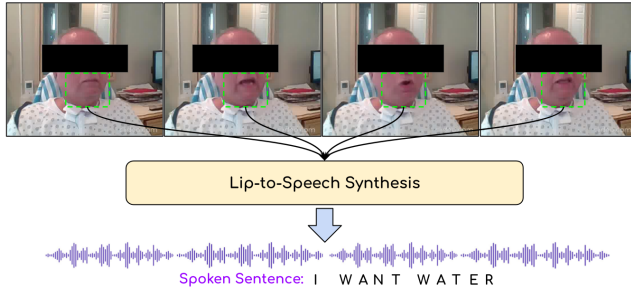
Figure 4: We demonstrate our model on an ALS patient who cannot voice words but can mouth them. We can generate the speech corresponding to the silent lip movements. Lip-to-Speech can thus be a cheap and non-invasive method to assist someone who has lost their voice.

Ethical Considerations. We acknowledge that our work has the potential to generate synthetic speech for videos, given that we only require a 1−second voice sample from any target speaker. However, since the video provides context and constrains the output speech, the generated speech will likely follow the original content closely. We recognize the importance of ethical considerations regarding using such models and ensure that our models will only be shared with users who consent to limit their usage to research-oriented and ethically valid tasks.

5 ABLATION STUDIES

In this section, we perform several ablation studies to understand the effect of different components of our model. All the ablation experiments are performed on the LRS2 [7] test set.

Effect of different pre-trained lip-to-text models. Additional experiments were conducted using other pre-trained lip-to-text models, specifically DeepLR [2] and AV-HuBERT [34]. The former had a WER of 51.3 on the LRS2 test set, while the latter had a WER of 46.6. The input text to our pre-trained Visual TTS module was taken from different Lip-to-Text models. Additionally, we also directly provide the ground-truth text from the LRS2 test set. As shown in Table 3, our approach recovered speech that was somewhat accurate, despite the presence of noisy text transcripts from both models. This can be attributed to two factors: (i) the correction of errors

made by the lip-to-text network by our pipeline to some extent (demonstrated in the demo video); and (ii) the reduced difference in scores between homonyms such as "ship and sheep" or "berth and birth" in the audio domain.

Table 3: Comparison of using generated text from different lip-to-text network in our pipeline. We also report the WER of the lip reading model (L2T-WER) on the LRS2 test set as a reference.

Method	L2T-WER	PESQ↑	STOI↑	ESTOI↑	LSE-C↑	LSE-D↓
Deep Lip Reading [2]	51.3	1.17	0.40	0.22	7.847	6.904
AV-HuBERT [34]	46.1	1.27	0.53	0.40	7.960	7.003
VTP (Ours)	22.6	1.47	0.65	0.47	8.083	6.586
GT text	-	1.51	0.69	0.50	8.781	6.106

Effect of different visual representations. We train our proposed Visual Text-to-Speech module with RGB face crops instead of the VTP embeddings to generate speech conditioned on text and lip movements. Based on our observations from Table 4, VTP embeddings are the most suitable for this task because they excel in localizing and representing the shape of the speaker's lips.

Table 4: We present the effect of using different visual representations for training the Visual TTS module.

Method	PESQ↑	STOI↑	ESTOI↑	LSE-C↑	LSE-D↓
Face crops	1.17	0.40	0.22	7.847	6.904
VTP (Ours)	1.47	0.65	0.47	8.083	6.586

6 LIMITATIONS

In our work, we address the problem of lip-to-speech networks not being able to learn a language model directly from speech supervision. We do so by using a pre-trained lip-to-text network. While our model does not require ground-truth text annotations, the lip-to-text model which we build upon has been trained with text supervision. However, recent efforts in self-supervised pre-training have led to a sharp decrease in the number of text annotations required for training accurate lip-to-text [34], making it easier to extend such models to lip-to-speech using our approach. Currently, we have only tested our model in English, and it remains to be validated in other languages.

7 CONCLUSION

Our research presents an innovative approach to unconstrained multi-speaker lip-to-speech synthesis that outperforms previous methods by incorporating language and visual information from a highly accurate lip-to-text model. We demonstrate significant improvements in lip-to-speech synthesis, generating high-quality outputs that seamlessly synchronize with silent lip video. Our study can potentially open up exciting avenues for future research. We are particularly encouraged by the success of our approach in assistive technology, where we have shown that our method can generate accurate speech from silent lip movements of individuals with speech impairments. Overall, we are optimistic about the possibilities of our approach to improve communication and enhance the quality of life for people with speech impairments, and we look forward to seeing our work drive further progress in this field.

REFERENCES

[1] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. 2018. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence* (2018).

[2] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. 2018. Deep Lip Reading: a comparison of models and an online application. In *INTERSPEECH*.

[3] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. 2018. LRS3-TED: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496* (2018).

[4] Hassan Akbari, Himani Arora, Liangliang Cao, and Nima Mesgarani. 2017. Lip2Audspec: Speech Reconstruction from Silent Lip Movements Video. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2017), 2516–2520.

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

[6] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 4960–4964. https://doi.org/10.1109/ICASSP.2016.7472621

[7] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. 2017. Lip reading sentences in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 3444–3453.

[8] Joon Son Chung and Andrew Zisserman. 2016. Lip reading in the wild. In *Asian Conference on Computer Vision*. Springer, 87–103.

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

[10] 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 (2006), 2421–2424.

[11] Ariel Ephrat, Tavi Halperin, and Shmuel Peleg. 2017. Improved Speech Reconstruction from Silent Video. *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)* (2017), 455–462.

[12] Ariel Ephrat and Shmuel Peleg. 2017. Vid2Speech: speech reconstruction from silent video. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

[13] Markus Freitag and Yaser Al-Onaizan. 2017. Beam Search Strategies for Neural Machine Translation. In *NMT@ACL*.

[14] Sang gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sung-Hoon Yoon. 2022. BigVGAN: A Universal Neural Vocoder with Large-Scale Training. *ArXiv* abs/2206.04658 (2022).

[15] Naomi Harte and Eoin Gillen. 2015. TCD-TIMIT: An audio-visual corpus of continuous speech. *IEEE Transactions on Multimedia* 17, 5 (2015), 603–615.

[16] Michael Hassid, Michelle Tadmor Ramanovich, Brendan Shillingford, Miaosen Wang, Ye Jia, and Tal Remez. 2022. More than Words: In-the-Wild Visually-Driven Prosody for Text-to-Speech. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), 10577–10587.

[17] Sindhu B. Hegde, K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar. 2022. Lip-to-Speech Synthesis for Arbitrary Speakers in the Wild. In *Proceedings of the 30th ACM International Conference on Multimedia* (Lisboa, Portugal) *(MM '22)*. Association for Computing Machinery, New York, NY, USA, 6250–6258. https://doi.org/10.1145/3503161.3548081

[18] Chenxu Hu, Qiao Tian, Tingle Li, Yuping Wang, Yuxuan Wang, and Hang Zhao. 2021. Neural Dubber: Dubbing for Videos According to Scripts. In *NeurIPS*.

[19] Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, and Yonghui Wu. 2018. Transfer Learning from Speaker Verification to Multispeaker Text-to-Speech Synthesis. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18)*. Curran Associates Inc., 4485–4495.

[20] Minsu Kim, Joanna Hong, and Yong Man Ro. 2022. Lip to Speech Synthesis with Visual Context Attentional GAN. In *Neural Information Processing Systems*.

[21] Minsu Kim, Joanna Hong, and Yong Man Ro. 2023. Lip-to-Speech Synthesis in the Wild with Multi-task Learning. *ArXiv* abs/2302.08841 (2023).

[22] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. *CoRR* abs/1412.6980 (2015).

[23] Rodrigo Mira, Alexandros Haliassos, Stavros Petridis, Björn Schuller, and Maja Pantic. 2022. SVTS: Scalable Video-to-Speech Synthesis. In *Interspeech*.

[24] Liliane Momeni, Hannah Bull, Prajwal K R, Samuel Albanie, Gül Varol, and Andrew Zisserman. 2022. Automatic dense annotation of large-vocabulary sign language videos. In *ECCV*.

[25] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O. Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. 2018. Deep Voice 3: 2000-Speaker Neural Text-to-Speech. In *International Conference on Learning Representations*. https://openreview.net/forum?id=HJtEm4p6Z

[26] K R 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 (CVPR)*. 5162–5172.

[27] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar. 2020. Learning Individual Speaking Styles for Accurate Lip to Speech Synthesis. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[28] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. 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* (Seattle, WA, USA) *(MM '20)*. Association for Computing Machinery, New York, NY, USA, 484–492. https://doi.org/10.1145/3394171.3413532

[29] Prajwal K R, Liliane Momeni, Triantafyllos Afouras, and Andrew Zisserman. 2021. Visual Keyword Spotting with Attention. In *BMVC*.

[30] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2020. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558* (2020).

[31] Bipasha Sen, Aditya Agarwal, Rudrabha Mukhopadhyay, Vinay Namboodiri, and CV Jawahar. 2021. Personalized One-Shot Lipreading for an ALS Patient. *arXiv preprint arXiv:2111.01740* (2021).

[32] Dmitriy Serdyuk, Otavio Braga, and Olivier Siohan. 2022. Transformer-Based Video Front-Ends for Audio-Visual Speech Recognition for Single and Muti-Person Video. In *Proc. Interspeech 2022*. 2833–2837. https://doi.org/10.21437/Interspeech.2022-10920

[33] Jonathan Shen, R. Pang, Ron J. Weiss, M. Schuster, Navdeep Jaitly, Z. Yang, Z. Chen, Yu Zhang, Yuxuan Wang, R. Skerry-Ryan, R. A. Saurous, Yannis Agiomyrgiannakis, and Y. Wu. 2018. Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2018), 4779–4783.

[34] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdel rahman Mohamed. 2022. Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction. *ArXiv* abs/2201.02184 (2022).

[35] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. *ArXiv* abs/1706.03762 (2017).

[36] Konstantinos Vougioukas, Pingchuan Ma, Stavros Petridis, and Maja Pantic. 2019. Video-Driven Speech Reconstruction using Generative Adversarial Networks. *arXiv preprint arXiv:1906.06301* (2019).