Lip-to-Speech Synthesis for Arbitrary Speakers in the Wild

Sindhu B Hegde*
sindhu.hegde@research.iiit.ac.in
International Institute of Information Technology, Hyderabad, India

K R Prajwal†*
prajwal@robots.ox.ac.uk
University of Oxford
United Kingdom

Rudrabha Mukhopadhyay*
radrabha.m@research.iiit.ac.in
International Institute of Information Technology, Hyderabad, India

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

C. V. Jawahar
jawahar@iiit.ac.in
International Institute of Information Technology, Hyderabad, India

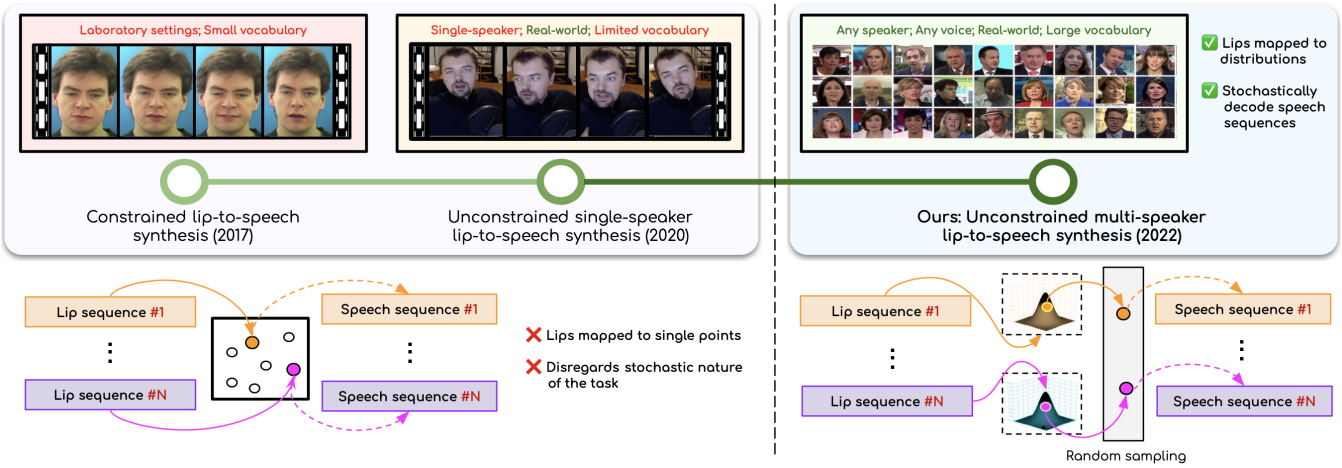

Figure 1: We address the problem of generating speech from silent lip videos for any speaker in the wild. Previous works train either on large amounts of data of isolated speakers or in laboratory settings with a limited vocabulary. Conversely, we can generate speech for the lip movements of arbitrary identities in any voice without additional speaker-specific fine-tuning. Our new VAE-GAN approach allows us to learn strong audio-visual associations despite the ambiguous nature of the task.

ABSTRACT

In this work, we address the problem of generating speech from silent lip videos for any speaker in the wild. In stark contrast to previous works, our method (i) is not restricted to a fixed number of speakers, (ii) does not explicitly impose constraints on the domain or the vocabulary and (iii) deals with videos that are recorded in the wild as opposed to within laboratory settings. The task presents a host of challenges, with the key one being that many features of the desired target speech, like voice, pitch and linguistic content, cannot be entirely inferred from the silent face video. In order to handle these stochastic variations, we propose a new VAE-GAN architecture that learns to associate the lip and speech sequences amidst the variations. With the help of multiple powerful discriminators that guide the training process, our generator learns to synthesize speech sequences in any voice for the lip movements of any person. Extensive experiments on multiple datasets show that we outperform all baselines by a large margin. Further, our network can be fine-tuned on videos of specific identities to achieve a performance comparable to single-speaker models that are trained on 4× more data. We conduct numerous ablation studies to analyze the effect of different modules of our architecture. We also provide a demo video that demonstrates several qualitative results along with the code and trained models on our website[1].

CCS CONCEPTS

• **Computing methodologies** → **Reconstruction**; **Neural networks**.

KEYWORDS

lip-to-speech, speech synthesis, hybrid vae-gan, talking-face videos

*All three authors contributed equally to this research.
†Work done at IIIT Hyderabad

MM '22, October 10–14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3548081

[1]http://cvit.iiit.ac.in/research/projects/cvit-projects/lip-to-speech-synthesis

ACM Reference Format:
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 (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3503161.3548081

1 INTRODUCTION

As the world's communication becomes increasingly digital, it is also becoming increasingly visual. From video calls to movies to YouTube videos, there is a surge in video content consumption. Naturally, understanding and enabling applications for talking-face videos [2, 3, 13, 20, 36, 37] has been an active area of research in recent years. Tasks such as speech/text-based lip synthesis [27, 30, 37] have witnessed tremendous advancements. The opposites of these tasks, namely, lip-to-text generation and lip-to-speech generation, both falling under the umbrella of "lip-reading", have proven far more challenging. For the task of lip-to-text generation, multiple impressive works have pushed the boundaries with models that work for any speaker in the wild. However, its sibling task, lip-to-speech synthesis, has not yet witnessed a similar advancement in such unconstrained settings.

Lip-to-Speech Synthesis for Arbitrary Identities: The goal of lip-to-speech synthesis is to generate meaningful speech for a silent talking-face video. Previous works in this space have focused on training models that work for a fixed set of speakers. They achieve impressive results but rely on videos recorded in laboratory settings [17, 23] or require tens of hours of single-speaker data [36] when working with real-world videos. This makes the previous methods hard to scale to the large number of identities in the wild.

Our goal in this work is to perform lip-to-speech synthesis for silent videos of any identity. This allows us to produce results on any speaker at test time. We also show that we can further fine-tune on videos of a single speaker, if necessary, and achieve similar performance to single-speaker models but with 4× lesser data.

Overarching Challenges: The set of challenges in our task can be divided into two major groups: (i) challenges of lip-to-speech generation and (ii) challenges in handling the large variations in identities. In lip-to-speech generation, deciphering the uttered words is ambiguous, e.g. lip movements of "pat", "bat" and "mat" are the same but map to different speech outputs. The second set of challenges is unique to the task in this work and was not faced by previous single-speaker models. The diversity in voices, accents and speaking styles makes it difficult to learn the lip-speech correspondences. Generating continuous speech data is also far more challenging when there are no constraints on identity and voice. Finally, more speakers usually mean more variations in terms of content and topics spoken. Given this second set of challenges, we must ask ourselves: *Can single-speaker methods be directly extended for unconstrained, real-world multi-speaker lip-to-speech synthesis?*

Overview of this Work: The key idea of this work is to allow the model to handle the highly stochastic nature of the task. Unlike the constrained single-speaker case, the model has limited knowledge about the topic being spoken. It needs to determine which voice to generate in, the pitch, accent, emotion, prosody and tone. All these aspects vary stochastically due to inadequate priors in the input. Existing single-speaker models trained using an L_1 reconstruction loss enforces a highly constrained one-to-one correspondence between the input and the output, which, as we will see, is detrimental to this task.

In this work, we propose a novel VAE-GAN architecture whose core idea is to map the input lip sequence to an output distribution of plausible speech sequences. We are the first to handle the issue of ambiguities in lip-to-speech synthesis explicitly. Through extensive quantitative and qualitative comparisons, we show that this is very helpful in such an unconstrained setting. In addition to the variational architecture, we add a variety of perceptual loss functions to ensure the realism and style of the generated speech. We show that these discriminators (GAN discriminator to enforce the generation quality and voice discriminator to enforce the voice quality) play an essential role when learning such an unconstrained task. Our model not just handles videos of arbitrary speakers but also makes the single-speaker lip-to-speech synthesis task much more scalable. We show that we can match the quality of the current state-of-the-art single-speaker models while using 4× lesser training data. Our key contributions/claims in this work are:

- We address the problem of lip-to-speech synthesis in the wild, with no explicit constraints on the number of speakers and vocabulary. This allows us to, for the first time, generate speech for any person's silent lip movements in any voice.
- We distill the content information from speech sequences and align them with the corresponding lip movements using our novel VAE-GAN architecture.
- Our pre-trained model can be further fine-tuned and personalized to specific speakers in a data-efficient manner compared to the current single-speaker models trained from scratch. We show that our network achieves comparable performance while using only 25% of the training data.

2 RELATED WORK

2.1 Constrained Multi-speaker Lip-to-speech

The problem of lip-to-speech synthesis has been receiving growing attention in recent times. One of the initial works [21] in constrained laboratory settings used a 2D convolution-based encoder-decoder architecture to learn a mapping between lip movements and LPC features of the corresponding speech. The network is trained on the GRID corpus [17], containing lab-recorded videos from 34 speakers. There have been follow-up works [5, 19, 31, 32, 41] that train in similar settings [17, 23], containing speakers with limited head movements and a small vocabulary. However, we observe that the performance of these works severely degrades when directly extended to unconstrained settings comprising in-the-wild videos with large variations in vocabulary, speakers and head movements.

2.2 Single-speaker Lip-to-speech

In order to perform lip-to-speech synthesis for in the wild silent videos, a more recent work, Lip2Wav [36] proposes the idea of learning a model by training on large amounts of single-speaker data. Lip2Wav demonstrates impressive results in real-world settings by utilizing $\approx$ 20 hours of data for isolated speakers. The sheer amount

of data per speaker allows the model to learn fine-grained speaker-specific attributes. The same work also shows preliminary results on word-level multi-speaker lip-to-speech using LRW [14] dataset. In Table 1, we contrast our task against the previous works. Most of the earlier works function under one or more constraints. As we will show later, these methods do not scale to the case of "generating speech for any identity in any voice". We discuss the reasons and describe how our novel approach addresses these issues.

Table 1: Major differences between our approach and the existing approaches. Our work deals with the most challenging task in this space.

Approach	vocab. size	natural setting?	training data per spkr (in mins.)	zero-shot gen. (unseen spkrs.)
Vid2Speech [21]	56	×	48	×
Ephrat et.al [19]	82	×	30	×
GAN based [41]	82	×	30	×
Lip2AudSpec [5]	56	×	48	×
Lip2Wav [36]	$\approx 5K$	✓	1200	×
Ours	**50K+**	✓	**3**	✓

2.3 Lip-to-text Generation

A closely related task to lip-to-speech generation is lip-to-text generation, usually referred to as "lip reading". Early works focused on obtaining single-word labels by posing it as a classification problem [14, 39]. More recent works can produce sentence-level predictions using different models with losses like CTC [8, 42] and models ranging from LSTMs [13] to Transformers [1, 26].

Synthesizing speech from lips is far more challenging than generating text from lips due to the following reasons: (i) Lip-to-text models only need to transcribe the content (words), whereas in lip-to-speech, along with the content, other speaker attributes like voice and prosody also need to be modelled; (ii) Lip-to-speech deals with continuous outputs (harder for learning), whereas lip-to-text has the luxury of generating discrete tokens. Thus, although there has been tremendous progress in unconstrained lip-to-text generation, lip-to-speech generation in unconstrained settings, which is the focus of our work, still has a large room for improvement.

3 VAE-GAN ARCHITECTURE

Given a sequence of lip movements $L = (L_1, L_2, ..., L_T)$ and a speaker identity vector V, the goal is to generate speech segment $S = (S_1, S_2, ..., S_{T'})$ corresponding to the lip movements L and in the voice of V. We start our discussion by examining the issues in previous methods and propose appropriate changes to enable learning in a significantly more unconstrained multi-speaker setting.

3.1 Fundamental issues in Previous Works

3.1.1 Stochastic Nature in Lip-to-Speech Synthesis. All of the previous works aim to map the input lip sequence to a single speech sequence, i.e., they do not account for stochastic nature of the task. The stochasticity arises due to inadequate priors, i.e., the speech cannot be entirely inferred from the lip movements due to the homophene ambiguity. But additional ambiguities are introduced as we move from laboratory settings to utterances in real-world videos [36], where even the single-speaker case becomes challenging when the speech is "freely uttered": it can have varying decibel levels (no concrete correlation to lips), stress on particular phonemes, and even transient lip motion during pauses.

3.1.2 Scaling to Multi-speaker Lip-to-Speech. As we move further into the multi-speaker case, the task becomes severely ill-posed and extremely challenging. Given only the lip movements and a voice token, there are many stochastically varying factors that cannot be inferred from either of the inputs. In addition to the ambiguities mentioned in Section 3.1.1, each speaker can have distinct speaking styles and lip shapes in the multi-speaker setting. The large variation in voices and accents also influences how the phonemes are uttered. Such variations cannot be adequately captured in the voice token input. As none of the existing models handle these issues, even in the single-speaker case, they do not scale well to multi-speaker lip-to-speech.

3.1.3 What "Space" is Right to Learn these Ambiguous Audio-Visual Correspondences? Finally, the existing models struggle to learn in the unconstrained multi-speaker scenario because their learning signal comes from the level of raw spectrograms. This is because most of the current models use a *visual encoder - speech decoder* approach with only an L_1 reconstruction loss. Given a large amount of stochastic variations in both the visual and speech modalities, we argue that it is beneficial to learn speech-lip correspondences in the feature space, whose benefits have been well-studied in the literature [10, 11, 25, 34, 35]. The intuition is that the low-level variations are more meaningfully represented in the feature space. For example, matching the lip shapes of "ma" with its instances in the speech in different voices will be far easier in a feature space that is voice invariant and contains only the content information from the speech sequences. We build upon this intuition to arrive at our core idea.

3.2 Our Core Idea

Our core idea is two-fold. Firstly, we want to match the distributions (Figure 1) of (i) the lip sequence and (ii) the content from the speech sequence in the latent feature space to allow the model to handle the stochastic variations mentioned above. Secondly, we want to learn a decoder that decodes meaningful speech samples from this latent space while also conditioning on a speaker identity embedding that provides the voice information.

Concretely, we first represent each input lip sequence as a distribution (instead of a single vector) and match it to the corresponding speech content distribution. The intuition for matching at the level of distributions is that it allows the ambiguities to be meaningfully represented by allowing a "one-to-many" correspondence. Once we have such a shared latent space, the second step is to decode samples from this latent distribution and generate meaningful speech. We realize these ideas in the following manner. We use a standard automatic speech recognition (ASR) model to extract content information from the input speech sequences. A variational auto-encoder [28] then maps the speech content information to a shared latent space and decodes the samples from this latent space to real speech sequences. We have an additional visual encoder that maps the lip sequences to the same shared latent space. We tie these two latent distributions together using the Kullback-Leibler Divergence

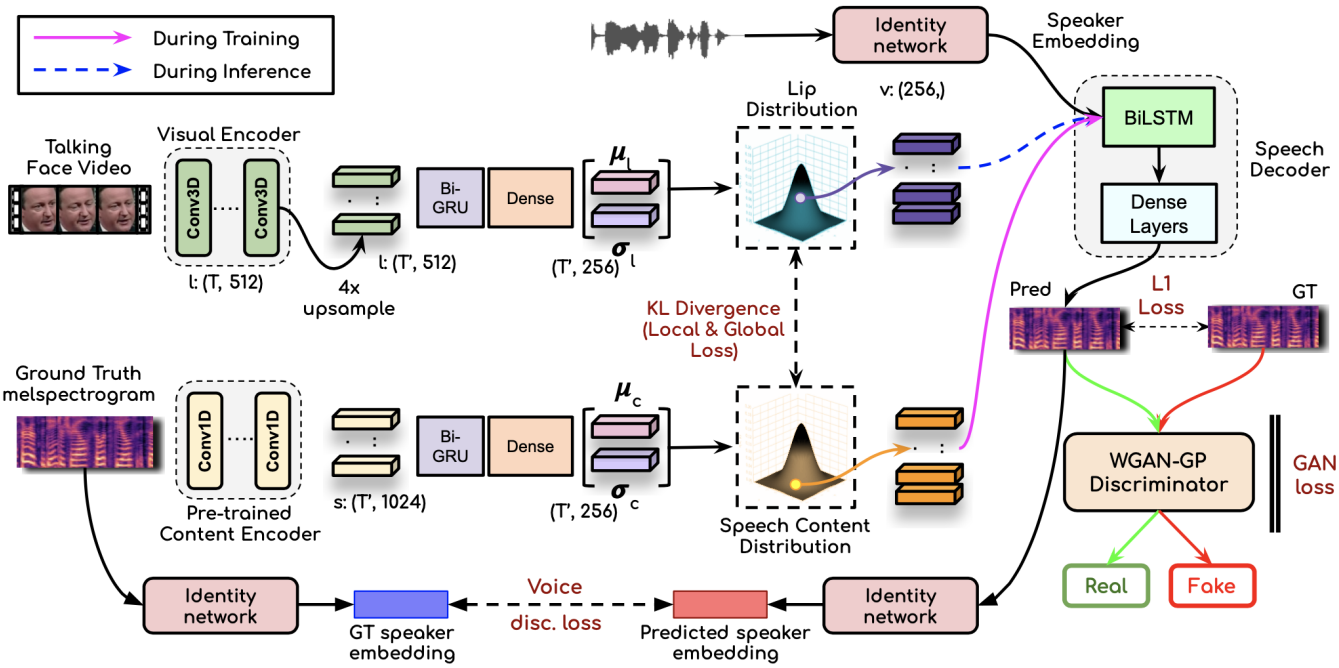

Figure 2: We propose a novel VAE-GAN architecture for our task. While previous approaches enforce a one-to-one mapping between lip and speech sequences, we deal with the task's ambiguities differently. We map the speech content (ASR representations) and the lip sequence to similar distributions and use a decoder to generate realistic speech outputs from this latent space. Additional discriminators enable high-fidelity generation in such unconstrained settings.

(KL) loss [29] as illustrated in Figure 2. Finally, we sample points from these distributions and feed them to a speech decoder along with the speaker identity embedding to generate intelligible speech sequences. We delve into each of the modules below.

3.2.1 Visual Encoder. We adopt the visual encoder used in several previous models that aim to learn audio-visual correspondence [2, 4, 15, 16]. Our visual encoder consists of 3D convolutional layers, with only the first layer having a temporal receptive field of 5 frames. It provides a good trade-off between speed and capturing short-range temporal information. The visual encoder inputs a spatio-temporal volume $L : (T, 96, 96, 3)$ and outputs 1D embeddings $l : (T, 512)$ at each time-step. We perform 4× temporal upsampling using nearest-neighbor interpolation on l to match the speech time-steps T'.

3.2.2 Speech Content Encoder. We believe that lip movements primarily represent the content information present in a speech sequence. Thus, before matching with the lip distribution, we need to distill the content from speech segment. We achieve this using a standard pre-trained ASR network [6]. The melspectrogram is passed through this frozen encoder to generate a $T' \times 1024$ dimensional embedding denoted by c. Thus, we separate the voice information from the speech representations, which, as we will see later, is crucial to our training strategy.

3.2.3 Variational Auto Encoder Based Approach & Latent Distribution Matching. We now map both, lip and speech content embeddings, l and c to Gaussian distributions with a diagonal covariance matrix: $\mathcal{N}(\mu_l, \sigma_l)$ and $\mathcal{N}(\mu_c, \sigma_c)$, where $(\mu_l, \sigma_l), (\mu_c, \sigma_c)$ are obtained using two projection modules P_l and P_c. Both these modules contain a bi-directional GRU [12] followed by ReLU-activated fully-connected layer. The bi-directional GRU helps to capture contextual information in both directions at each time-step. Now that we have two distributions, one corresponding to the speech content, another corresponding to the lips, random points c_p and l_p are sampled from these distributions using the re-parametrization trick [28]. We can clearly see that we no longer have a "single value" for each input lip or speech sequence, but rather two probability distributions for these inputs.

Our final step is to tie these distributions together, i.e., we want the lip distribution $\mathcal{N}(\mu_l, \sigma_l)$ to be close to the speech content distribution $\mathcal{N}(\mu_c, \sigma_c)$. If we do this, then we can train a decoder that decodes speech samples from the content distribution $\mathcal{N}(\mu_c, \sigma_c)$ and also use it to decode from points in the lip distribution. Thus, we minimize the Kullback-Leibler Divergence (KL) loss [29] between these two distributions:

$$L_{\text{kl}_{\text{global}}} = \frac{1}{N}\sum_{i=1}^{N} KL[\mathcal{N}(\mu_c, \sigma_c)||\mathcal{N}(\mu_l, \sigma_l)] \tag{1}$$

We term $L_{kl_{global}}$ as the global KL-divergence loss since the distributions are created by considering the complete sequence of speech and lip movements. To further improve the alignment, inspired from [18], we take random corresponding temporal segments of the distributions and align them by minimizing a "local" KL-divergence loss (Figure 3). We choose $R = 10$ small temporal segments for each batch sample and use its (μ^r, σ^r) to minimize Equation 2:

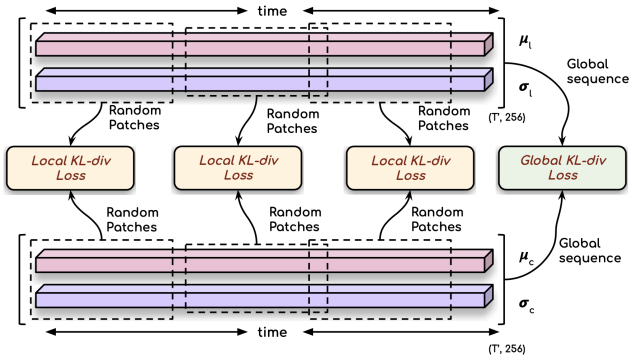

Figure 3: In addition to using a global KL-divergence loss to tie the lip and speech content distributions, we also enforce these distributions to be temporally aligned by minimizing a local KL-divergence loss on random smaller time segments. The intuition is that lips and speech are locally aligned in time, in the form of visemes and phonemes.

$$L_{\text{kl}_{\text{local}}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{r=1}^{R} KL[\mathcal{N}(\mu_c^r, \sigma_c^r)||\mathcal{N}(\mu_l^r, \sigma_l^r)] \tag{2}$$

Here, $\mathcal{N}(\mu^r, \sigma^r)$ are r^{th} random patches sampled from the lip and content distributions along the temporal dimension. Binding the distributions at the local level is crucial as phoneme-viseme mappings occur locally rather than globally. We show the importance of employing both local and global KL-divergence losses in Table 6. Since the two distributions are aligned using the KL-divergence loss, we can sample from the lip distribution during inference while sampling from the speech content one during training.

3.2.4 Speaker Embedding. While the visual/content encoder specifies "what to utter", we also need an input for "which voice to utter in". While this can be done by representing each speaker in the dataset with a one-hot vector, it does not generalize to new speakers during inference. Instead, we adopt a recent advancement [24] in training multi-speaker text-to-speech models, where a pre-trained identity network[2] containing the embedding with voice information is used. The speaker embedding can be obtained for any voice, given just one second of the voice sample. For each video in the dataset, we generate a 256-dimensional speaker embedding using a random one-second segment of the audio. We apply a ReLU-activated fully-connected layer to this pre-trained speaker embedding input V before feeding it to the decoder.

3.2.5 Speech Decoder. Our final step is to train a module that can generate speech segments given points sampled from the above created joint latent space. It is clear that we need to feed points from the lip distribution at test time, as we do not have the speech. During training, we have three ways of sampling the points: (i) only from the lip distribution, (ii) only from the speech content distribution and (iii) alternately sample from both the distributions. We hypothesize that learning with points from (ii) is far easier and allows the network to learn excellent latent representations of the speech content. As the distributions are being matched in the latent space, learning accurate, meaningful representations of one of them can be quite beneficial for learning the joint space. Indeed, we observed that good convergence and intelligible speech both at training and test time could be achieved only by training the decoder on points sampled from the speech content distribution. The points are sampled using the re-parametrization trick [28] and are of the shape $(T', 256)$. Along with the sampled points from content (c_p) or lip distribution (l_p), the decoder ingests the speaker embedding input V, which is concatenated with the sampled points. The concatenated content-voice feature vectors $[(c_p|l_p); V] : (T', 512)$ contain information on both "what to utter" and "the voice to utter in". We use a bi-directional LSTM layer followed by 4 dense layers to decode the melspectrogram segment $(T', 80)$ from the concatenated feature vector. During training, the generator G ingests speech content c and speaker embedding V and minimizes the L_1 reconstruction loss between the generated speech and the ground-truth speech S:

$$L_{\text{r}} = \frac{1}{N}\sum_{i=1}^{N} ||G(c, V) - S||_1 \tag{3}$$

Note that during training, the generator network is essentially a VAE for the speech with an additional KL loss constraint on its latent space. Because of the KL loss, we can sample from the speech distribution during training, and during inference, when the speech is absent, we can sample and decode points from the lip distribution. Therefore, during inference, we predict $G(l, V)$ as our output. Since we employ a content encoder during training, the decoder is forced to condition on the speaker embedding for the voice information. The content encoder distills only the content information from a speech sequence and does not leak the voice information, allowing us to maintain good voice quality even at test time when we decode from the lip distribution. We now describe additional discriminators that we will use along with our generator to improve the quality and accuracy of the generated speech outputs.

3.2.6 Enforcing Realism with a VAE + GAN. In our experiments, we observed that generating realistic samples for such a diverse set of voices, accents and speaking styles, using a plain L_1 reconstruction loss produced unrealistic, unintelligible samples (Table 5). We hypothesize that this occurs because of the known issue of the L_1 loss regressing to the mean. Multiple works in the past [33, 38] also point to the benefits of using a GAN along with a VAE. We found that it is highly beneficial to train a WGAN-GP [22] critic in a GAN setup along with our VAE architecture. The critic consists of a series of 1D convolutional layers that takes an audio spectrogram segment of shape $T' \times 80$ as input and outputs a single number as the score. The generator G and the critic D optimize the Wasserstein objective [7] along with the gradient penalty [22] in the equations below, where $\hat{S}$ contains all linear interpolates between S and $G(L, V)$:

$$L_{\text{adv}} = \mathbb{E}_{x\sim S}[D(x)] - \mathbb{E}_{x'\sim G(L,V)}[D(x')] \tag{4}$$

$$L_{\text{gp}} = \mathbb{E}_{\hat{x}\sim\hat{S}}[(\|\nabla_{\hat{x}} D(\hat{x})\| - 1)^2 \tag{5}$$

3.2.7 Improving Voice of the Generated Speech. To ensure that the model learns the voice and other style attributes, we use our pre-trained identity network as described in Section 3.2.4 to penalize the generated speech segments if they do not match the voice/style attributes of the ground-truth speech segment. We train the discriminator network to maximize the cosine similarity L_{voice} between

[2] github.com/CorentinJ/Real-Time-Voice-Cloning

the embeddings of the generated (V_{gen}) and the ground-truth (V_{GT}) speech segments.

$$L_{\text{voice}} = \frac{1}{N}\sum_{i=1}^{N}\frac{V_{gen}\cdot V_{GT}}{max(\|V_{gen}\|_2\cdot\|V_{GT}\|_2,\epsilon)} \tag{6}$$

3.3 Training Settings & Inference

The complete loss function to train our network is the weighted summation of all the above losses:

$$L_{\text{G}} = \lambda_r L_{\text{r}} + \lambda_{k_{global}} L_{\text{KL}_{\text{global}}} + \lambda_{k_{local}} L_{\text{KL}_{\text{local}}} + \lambda_{\text{voice}} L_{voice} + \min_{G}\max_{D} L_{\text{adv}} + \lambda_{\text{g}} L_{\text{gp}} \tag{7}$$

In our experiments we set $\lambda_r = 10$, $\lambda_{k_{global}} = 5$, $\lambda_{k_{local}} = 5$ and $\lambda_{voice} = 5$. We follow the pre-processing procedures of Lip2Wav [36] to detect and extract face crops from training videos. We create video inputs by randomly sampling a window of $T = 25$ (1 sec.) contiguous face crops resized to 96×96. The corresponding audio segment is sampled at 16kHz. We compute STFT with a hop length of $10ms$ and a window length of $25ms$. We finally obtain melspectrograms with 80 mel-bands and $T' = 100$ mel time-steps (1 sec.). We use a batch size of 32 and RMSProp [40] optimizer with an initial learning rate of 0.00005 for both the generator and the discriminator, which is advised for training a WGAN model [22]. The generator is trained every five discriminator iterations following [22]. As this is a WGAN, the discriminator loss shows the progress of training and correlates with the quality of the generated samples. Hence, we stop the training once the discriminator loss does not improve for 10 epochs. During inference, we feed the speaker embedding and the lip distribution to the decoder instead of the content distribution. Since our model can take a variable number of time steps as input, it can directly generate for any length of video without any further changes.

3.3.1 Datasets and Training Strategy. Our primary focus is to synthesize speech for silent lip videos in unconstrained settings; we intend to make our model identity-agnostic and work for a larger vocabulary. But, for the sake of comparison with previous works, we also train our model on the lab-recorded constrained GRID [17] and TCD-TIMIT [23] datasets. We use the speaker-independent train-test setting as [41] for GRID and single-speaker lip-to-speech setting as [36] for TCD-TIMIT dataset. For unconstrained evaluation, we first train our model on the word-level LRW data [14]. Next, we use the complete LRS2 dataset [13] (both train and pre-train sets), which contains sentences and phrases as opposed to specific words. The LRS2 data comprises thousands of speakers from BBC programs with a vocabulary of $59k$ and $2M$ word instances. A large number of speakers and vast vocabulary covered in both of these datasets encourage our model to be speaker agnostic and pose no limitations on the vocabulary size.

4 EXPERIMENTS

We evaluate our model against various baseline methods in two settings: (i) laboratory setting videos and (ii) in the wild videos.

4.1 Evaluation in Constrained Settings

4.1.1 Baselines. We compare our model with the following existing lip-to-speech methods: (i) Improved Vid2Speech [19], (ii) GAN-based [41] and (iii) Lip2Wav [36]. Note that since we train using the same settings as Lip2Wav [36] for TCD-TIMIT dataset, we report the paper scores for all the comparison methods. Similarly, we take the paper scores from [41] for GRID dataset (speaker-independent training settings).

4.1.2 Metrics. We evaluate our model using the standard speech metrics: Perceptual Evaluation of Speech Quality (PESQ) and short-time objective intelligibility measure (STOI). PESQ measures the overall perceptual quality of speech and STOI correlates with the intelligibility of speech. Also, to specifically evaluate the voice quality of the generated samples, we measure the distance (L_1) between the speaker embeddings of the generated and the ground-truth samples (termed speaker embedding distance (SED)).

4.1.3 Results. Table 2 shows the results of different models on GRID and TCD-TIMIT datasets. We see that our approach designed specifically for the unconstrained scenario performs slightly better or comparable to other methods when used in constrained settings. Also, we can observe that in terms of voice quality (SED metric), our method beats the existing approaches, thus indicating that we are able to preserve the voice of the identity to a large extent.

Table 2: Quantitative results on the constrained GRID [17] and TCD-TIMIT [23] datasets.

Dataset	GRID [17]			TCD-TIMIT [23]		
Method	PESQ↑	STOI↑	SED↓	PESQ↑	STOI↑	SED↓
Imp. Vid2Speech [19]	n/a	n/a	n/a	1.23	0.49	n/a
GAN-based [41]	1.24	0.44	n/a	1.22	0.32	n/a
Lip2Wav [36]	1.20	0.38	4.38	1.35	**0.56**	4.64
Ours	**1.28**	**0.45**	**3.76**	**1.35**	0.55	**4.36**

4.2 Evaluation in Unconstrained Settings

4.2.1 Baselines. As no prior works in the multi-speaker lip-to-speech synthesis train on such unconstrained datasets, we extend previous models [19, 36, 41] with the same speaker embedding we use in our model and train all of them on the same dataset as ours. On the LRW dataset, we evaluate the publicly released multi-speaker Lip2Wav model. Additionally, to highlight the importance of our novel modules and facilitate more direct comparison, we implement the following baselines: (i) a non sequence-to-sequence encoder-decoder architecture, (ii) a sequence-to-sequence model with only L_1 reconstruction loss and without the VAE-GAN setup, (iii) A standard speech encoder trained from scratch instead of our pre-trained content encoder and (iv) Lip-to-text [26] followed by text-to-speech (TTS) [24] model.

4.2.2 Metrics. Explicitly modeling the stochastic nature of the problem is one of the major contributions of our work. Naturally, this allows our model to generate speech samples, which can differ from the original ground-truth. Thus, along with the standard speech evaluation metrics (PESQ, STOI) and our voice quality metric (SED), that directly evaluate the generated speech against a fixed ground-truth, we also evaluate our model using the perceptual metrics. Specifically, following the recent GAN-based TTS systems [9],

Table 3: All models are pre-trained on LRW dataset and then trained on LRS2. We can see that we outperform all the competitive methods, especially on the challenging LRS2 data, which contains unseen speakers, words, poses and a large vocabulary.

Dataset	LRW [14]							LRS2 [13]						
Method	PESQ↑	STOI↑	SED↓	FDSD↓	KDSD↓	LSE-C↑	LSE-D↓	PESQ↑	STOI↑	SED↓	FDSD↓	KDSD↓	LSE-C↑	LSE-D↓
Imp. Vid2Speech [19]	0.65	0.09	6.01	5.645	10.2	1.782	10.43	0.59	0.30	6.25	4.275	3.1	2.009	8.424
GAN-based [41]	0.72	0.10	5.90	5.189	9.1	1.983	9.426	0.80	0.40	6.13	3.626	1.8	2.503	8.489
Lip2Wav [36]	1.19	0.54	5.73	1.831	1.1	2.526	8.286	0.58	0.28	6.22	10.71	15.5	1.874	11.48
Seq2seq baseline	1.01	0.50	6.16	4.306	7.7	2.396	8.412	0.97	0.43	6.47	3.840	2.8	1.991	8.532
Non seq2seq baseline	1.05	0.49	6.17	4.112	7.1	2.282	8.441	0.96	0.44	6.43	3.803	2.3	2.078	8.536
Ours w/o Content Encoder	0.53	0.14	5.89	2.941	3.4	2.531	8.205	0.45	0.32	6.02	2.856	1.2	2.385	8.230
Lip-to-text [26] + TTS [24]	0.60	0.09	5.91	1.056	0.5	2.181	14.160	0.48	0.11	6.17	0.984	0.1	2.024	19.012
Ours	0.78	0.15	**5.65**	**1.638**	**0.8**	**2.538**	**8.173**	0.60	0.34	**5.95**	**1.273**	**0.2**	**2.507**	**8.155**

we propose to use: *Frechet DeepSpeech Distance (FDSD)* and *Kernel DeepSpeech Distance (KDSD)*, to evaluate the perceptual quality and the linguistic aspect of the generated speech. Note that we multiply KDSD scores with 10^3 for better readability. Further, we also evaluate whether the output speech matches the lip movements using LSE-C (measures the confidence of lip-syncing) and LSE-D (measures an embedding level distance between the speech and lip-movements) metrics of [37]. We use the public implementations of these metrics for reliable comparison and reproducibility.

4.2.3 Results. Table 3 compares our model with the different methods on the LRW and LRS2 datasets. We outperform existing approaches [19, 41] and the baseline methods by a significant margin in perceptual metrics. Thus, although we under-perform in standard speech metrics (PESQ and STOI), we argue that our method is superior because perceptual metrics are more correlated to human judgement of intelligibility and speech quality. We further support this fact by providing qualitative results in the demo video on our website and conducting a human evaluation (Table 4). The standard metrics PESQ and STOI enforce one-to-one mapping, and thus are not ideal for evaluating our method. Also, GRID and TIMIT are constrained datasets with very less variations. On the other hand, LRW and LRS2 are more challenging and unconstrained datasets and our method is more effective on such challenging data. On the LRS2 dataset, where single-speaker methods such as Lip2Wav [36] fail to learn the audio-visual alignment, we achieve state-of-the-art perceptual metric scores. We encourage the reader to view the demo video for qualitative comparisons demonstrating the superiority of our approach.

Lip-to-text + TTS baseline: An additional baseline would be to use a state-of-the-art lip-to-text model [26] and convert the predicted text transcripts to speech using a multi-speaker TTS model [24]. We can deduce the following from the scores reported in Table 3. The lip-to-text model trained on text transcripts is naturally far more accurate in predicting the word tokens than any lip-to-speech model. Thus, it achieves the best results in terms of intelligibility and perceptual quality metrics such as FDSD, KDSD. The fact that our lip-to-speech model comes close to the lip-to-text baseline for the same metrics shows that our approach captures the speech content most accurately.

For other metrics like LSE-D that measures if the generated speech is in-sync with the video, we see that the output of lip-to-text baseline is not in-sync with the lip movements. The same content can be uttered in different ways (speeds, accent, prosody, voice), and the lip-to-text + TTS baseline cannot capture this. All the lip-to-speech models inherently achieve this to different extents and is an essential condition for the lip-to-speech task. Thus, ours is the best approach for the task of lip-to-speech synthesis.

Figure 4: Activation maps of the visual encoder. Our model strongly attends to the lip region while generating speech, despite variations in head poses and the lip location.

Qualitative results: In Figure 4, we plot the activation maps from the visual encoder to highlight that the model predominantly attends to the lip region. We encourage the reader to check our supplementary and demo video on our website for qualitative samples.

Computation cost: We train our network using 4 NVIDIA 2080 Ti GPUs. The network consists of 18M parameters and takes 0.5-seconds to generate 1-second of speech.

4.3 Human Evaluations

We perform human evaluations with the help of 20 participants. The participant group spans members of $22-40$ years with an almost equal male-female ratio. We choose 15 random samples from the LRS2 dataset [13] and generate the results for all the comparison models. Participants rate the speech segments on a scale of $1-5$ based on: (A) Intelligibility (is the speech meaningful?) (B) Perceptual Quality (C) Sync Accuracy (is the generated speech in-sync with lip movements?) and (D) Voice Match. Table 4 summarizes the mean scores of all the participants. Inline with the quantitative evaluations, the speech generated by our approach is of considerably higher quality and is more legible and natural. We also perform a Student's T-Test for Table 4 and compute the p-value to be ≈ 0.035, indicating that the differences are statistically significant.

4.4 Adapting to Single-Speaker Lip-to-Speech

Our model can generate speech for arbitrary speakers, which is highly beneficial for applications where there is almost no training data available for that speaker. However, in a few applications, it is possible to obtain some data of a target speaker for fine-tuning. Current single-speaker models need nearly 20 hours of data to produce impressive results. This is really difficult to obtain in many scenarios. Our multi-speaker model can resolve this issue to a large extent

Table 4: (A) Intelligibility (is the speech meaningful?), (B) Perceptual Quality, (C) Sync Accuracy, (D) Voice Match. Our approach outputs meaningful, intelligible speech that matches lip movements and voice of the target person.

Method	(A)	(B)	(C)	(D)
Imp. Vid2Speech [19]	2.02	1.98	1.74	1.13
WGAN-based [41]	2.17	2.43	2.19	2.01
Lip2Wav [36]	1.07	1.02	1.25	1.03
Seq2seq baseline	1.98	2.10	1.86	1.83
Non seq2seq baseline	2.01	2.23	1.92	1.84
Ours w/o Content Encoder	2.51	2.62	2.01	1.76
Ours	**3.22**	**2.98**	**2.28**	**2.69**

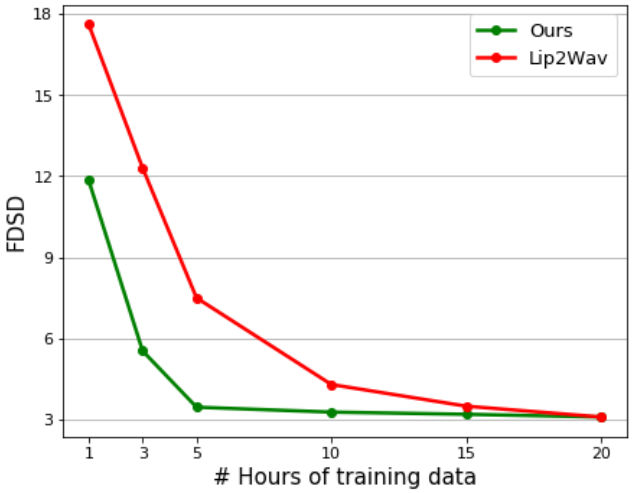

Figure 5: Fine-tuning our pre-trained multi-speaker model consistently outperforms the current best single-speaker model (FDSD lower is better) in the low data regime.

- we can fine-tune our pre-trained multi-speaker model on a small amount of speaker-specific data and achieve impressive personalized results. By using only 25% of the training data (5 hours), we can nearly match the single-speaker model's performance trained with 20 hours.

We fine-tune our network on the speakers in the Lip2Wav dataset [36]. We vary the number of hours in the train set and train the current single-speaker state-of-the-art Lip2Wav model [36] and fine-tune our multi-speaker model. We plot the variation of the FDSD metric with the training data size in Figure 5. We can clearly see that in the low data regime, pre-training on multi-speaker data vastly outperforms the best single-speaker model trained from scratch.

5 ABLATION STUDIES

We perform ablations to assess the effect of our key design choices. The results are on the unseen LRS2 test set.

5.1 Impact of Each Discriminator

We use two discriminators in our final model, one for enforcing better voice and style attributes and another for enforcing realistic speech. We assess the importance of using each of them in Table 5. We can see that despite getting a minor improvement in lip-sync metrics, both the discriminators enforce better overall speech generation as observed by the speech metrics.

Table 5: The discriminators enforce our model to produce meaningful and realistic speech outputs.

Method	FDSD↓	KDSD↓	LSE-C↑	LSE-D↓
Ours w/o both Discs	4.055	2.9	2.188	8.199
Ours w/o WGAN	3.916	2.7	2.294	8.194
Ours w/o Voice Disc	4.310	3.6	2.319	8.189
Ours	**1.273**	**0.2**	**2.507**	**8.155**

5.2 Importance of Local and Global Alignment

Table 6 shows that optimizing both local and global KL-divergence together improves the alignment between lip and content distributions, thus improving the overall performance. Training with either of the losses in isolation leads to inferior results.

Table 6: Optimizing both the global and local KL-divergence loss improves the overall quality of the results.

Method	FDSD↓	KDSD↓	LSE-C↑	LSE-D↓
Ours w/o local KL div	2.883	3.2	2.340	8.249
Ours w/o global KL div	5.040	6.8	2.003	8.937
Ours	**1.273**	**0.2**	**2.507**	**8.155**

6 LIMITATIONS AND FUTURE DIRECTIONS

Unconstrained lip-to-speech synthesis is far from a solved problem - there is still a considerable room for improvement. Ours is the first attempt to design a model which can generate speech for any speaker in any voice. We specifically deal with the issue of learning speech-lip correspondences and handling the homopheme ambiguities. However, there are still multiple unresolved problems. For example, our model struggles when there is a drastic movement of the head while speaking and if the head is non-frontal. Another issue is that we can get output sounds that do not form the right words or phrases. This is because it is hard to learn a language model in the speech modality compared to lip-to-text models that train on text transcripts. We hope our efforts in this work lead to new future directions that tackle some of the aforementioned issues.

7 CONCLUSION

In this work, we address the problem of the unconstrained lip-to-speech synthesis for the first time. We extensively discuss the challenges this problem presents, due to which the existing approaches fail to scale to such unconstrained settings. To tackle these challenges, we propose a VAE-GAN model trained to explicitly handle the stochastic nature of the task. Our approach produces significantly more intelligible, realistic speech outputs compared to all other models. We justify the use of different parts of our architecture with numerous ablation studies. We believe that our core idea of handling stochasticity can encourage future efforts to enable advanced versions of lip-to-speech and lip-to-text generation systems for arbitrary languages and speakers in the wild.

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. The Conversation: Deep Audio-Visual Speech Enhancement. *ArXiv* abs/1804.04121 (2018).

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

[4] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. 2020. Self-Supervised Learning of Audio-Visual Objects from Video. In *European Conference on Computer Vision*.

[5] 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.

[6] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, and Jingliang Bai. 2016. Deep Speech 2: End-to-End Speech Recognition in English and Mandarin. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48* (New York, NY, USA) *(ICML'16)*. JMLR.org, 173–182.

[7] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein Generative Adversarial Networks *(Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, International Convention Centre, Sydney, Australia, 214–223.

[8] 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).

[9] Mikołaj Bińkowski, Jeff Donahue, Sander Dieleman, Aidan Clark, Erich Elsen, Norman Casagrande, Luis C. Cobo, and Karen Simonyan. 2020. High Fidelity Speech Synthesis with Adversarial Networks. In *International Conference on Learning Representations*. https://openreview.net/forum?id=r1gfQgSFDr

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709* (2020).

[11] Jun-Ho Choi, Jun-Hyuk Kim, Manri Cheon, and Jong-Seok Lee. 2020. Deep learning-based image super-resolution considering quantitative and perceptual quality. *Neurocomputing* 398 (2020), 347–359.

[12] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*.

[13] 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.

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

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

[16] S. Chung, J. S. Chung, and H. Kang. 2019. Perfect Match: Improved Cross-modal Embeddings for Audio-visual Synchronisation. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 3965–3969. https://doi.org/10.1109/ICASSP.2019.8682524

[17] 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.

[18] U. Demir and G. Ünal. 2018. Patch-Based Image Inpainting with Generative Adversarial Networks. *ArXiv* abs/1803.07422 (2018).

[19] 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.

[20] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T. Freeman, and Michael Rubinstein. 2018. Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation. 37, 4, Article 112 (July 2018), 11 pages. https://doi.org/10.1145/3197517.3201357

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

[22] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. 2017. Improved Training of Wasserstein GANs. In *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc., 5767–5777. https://proceedings.neurips.cc/paper/2017/file/892c3b1c6dccd52936e27cbd0ff683d6-Paper.pdf

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

[24] 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.

[25] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*.

[26] Prajwal K. R, Triantafyllos Afouras, and Andrew Zisserman. 2021. Sub-word level lip reading with visual attention. *arXiv preprint arXiv:2110.07603* (2021).

[27] Prajwal K R, Rudrabha Mukhopadhyay, Jerin Philip, Abhishek Jha, Vinay Namboodiri, and C V Jawahar. 2019. Towards Automatic Face-to-Face Translation. In *Proceedings of the 27th ACM International Conference on Multimedia* (Nice, France) *(MM '19)*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3343031.3351066

[28] Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.).

[29] Solomon Kullback. 1959. *Information Theory and Statistics*. Wiley, New York.

[30] Rithesh Kumar, J. Sotelo, K. Kumar, A. D. Brébisson, and Yoshua Bengio. 2018. ObamaNet: Photo-realistic lip-sync from text. *ArXiv* abs/1801.01442 (2018).

[31] Yaman Kumar, Mayank Aggarwal, Pratham Nawal, Shin'ichi Satoh, Rajiv Ratn Shah, and Roger Zimmermann. 2018. Harnessing AI for Speech Reconstruction Using Multi-View Silent Video Feed. In *Proceedings of the 26th ACM International Conference on Multimedia* (Seoul, Republic of Korea) *(MM '18)*. Association for Computing Machinery, New York, NY, USA, 1976–1983. https://doi.org/10.1145/3240508.3241911

[32] Yaman Kumar, Rohit Jain, Khwaja Mohd Salik, Rajiv Ratn Shah, Yifang Yin, and Roger Zimmermann. 2019. Lipper: Synthesizing thy speech using multi-view lipreading. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 2588–2595.

[33] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. 2016. Autoencoding beyond pixels using a learned similarity metric. In *International conference on machine learning*. PMLR, 1558–1566.

[34] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. 2017. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 105–114. https://doi.org/10.1109/CVPR.2017.19

[35] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

[36] 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)*.

[37] 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

[38] Mihaela Rosca, Balaji Lakshminarayanan, David Warde-Farley, and Shakir Mohamed. 2017. Variational approaches for auto-encoding generative adversarial networks. *arXiv preprint arXiv:1706.04987* (2017).

[39] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. 2017. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (ToG)* 36, 4 (2017), 1–13.

[40] T. Tieleman and G. Hinton. 2012. Lecture 6.5—RmsProp: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning.

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

[42] Kai Xu, Dawei Li, Nick Cassimatis, and Xiaolong Wang. 2018. LCANet: End-to-end lipreading with cascaded attention-CTC. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 548–555.

A SUPPLEMENTARY MATERIAL

A.1 Additional Experiments and Ablation Studies

In this section, we report additional experiments and ablation studies to get further insights and better understand the different aspects of our network. The experimental setup is identical to that of the ablation studies in the main paper.

We also encourage the reader to view the demo video containing the results and comparisons.

*A.1.1 Comparisons on LRS3 [**?**] Dataset.* We compare our work on the LRS3 [**?**] test split. LRS3 dataset was collected from TedX videos and consists of a large vocabulary and majorly profile face videos. Further, the videos also have very different lighting conditions and extreme head-motions when compared to LRS2 [13] dataset. Please note that, we do not fine tune our model on LRS3 dataset; thus evaluating on it highlights our model's generalization ability and robustness on completely unseen speakers. As seen in Table 7, our method performs well on LRS3 videos containing unseen vocabulary, identities, voices, and profile views, clearly indicating the robustness of our approach.

Table 7: Quantitative comparison on LRS3 dataset [?]. We can see that we outperform all the competitive methods, even in a very different setting in LRS3, which contains unseen speakers, words, and a large number of profile views. Note that our model is not fine-tuned on the LRS3 dataset.

Dataset	LRS3 [?]			
Method	FDSD↓	KDSD↓	LSE-C↑	LSE-D↓
Imp. Vid2Speech [19]	5.286	5.4	1.929	8.435
GAN-based [41]	4.589	4.2	2.031	9.372
Lip2Wav [36]	12.663	16.2	1.632	11.465
Seq2seq baseline	4.821	4.8	1.880	8.568
Non seq2seq baseline	4.766	4.2	1.874	8.579
Ours w/o Content Encoder	4.054	2.9	2.041	8.312
Ours	**3.148**	**1.8**	**2.063**	**8.256**

A.1.2 Near Frontal vs. Non-frontal Videos. We study the extent to which the performance deteriorates if the face view moves towards a non-frontal profile view. As expected and also observed in past lip-reading works [3], our model also has room for improvement when handling non-frontal talking faces.

Table 8: Similar to other lip-reading models [3], our model also has room for improvement when dealing with non-frontal views.

Method	LSE-C↑	LSE-D↓	FDSD↓	KDSD↓
Near frontal	2.608	8.016	1.351	0.3
Non-frontal	2.491	8.058	3.473	1.0

A.1.3 What Kind of Visual Input is the best? We compare different forms of visual inputs, such as feeding only the lower-half of the face and pre-trained face embeddings [**?**]. Providing the full face crop performs the best, as shown in Table 9 and also reflected by activation maps near the eye regions in Figure 4 of the main paper.

Table 9: Feeding the full face crop produces the best results.

Method	LSE-C↑	LSE-D↓	FDSD↓	KDSD↓
Facenet emb [**?**]	1.931	**7.664**	8.641	8.7
Lower half (ours)	2.338	8.173	3.224	2.8
Full face (ours)	**2.507**	8.155	**1.273**	**0.2**

A.1.4 Sampling strategy of VAE at train-time. In 3.2.4 of the main paper, we mention that there are three ways of sampling the points to decode from during the training: (i) only from the lip distribution, (ii) only from the speech content distribution, (iii) alternately sample from both the distributions. We mentioned that we obtain good convergence and intelligible results only by following (ii), because it allows the network to learn excellent latent representations of the speech, which can help overall learning. We experimentally verify this in Table 11, by showing that sampling from other distributions, i.e. (i) Lip distributions and (iii) Alternately sampling from both the distributions, leads to far worse results.

Table 10: Sampling solely from the speech distribution during training enables the decoder to learn to generate realistic, accurate outputs.

Method	LSE-C↑	LSE-D↓	FDSD↓	KDSD↓
Lip dist.	2.261	8.195	4.036	2.7
Speech content dist.	**2.507**	**8.155**	**1.273**	**0.2**
Alternate sampling	2.241	8.320	3.487	2.4

A.1.5 Auto-encoder vs. VAE. We assess the importance of mapping the lip and the content distributions using a VAE. In Table 11, we show that removing the variational aspect and using just a naive auto-encoder approach results in poor speech generation. Interestingly, while the network learns comparable audio-visual correspondence, the speech is not intelligible or meaningful, as indicated by the speech metrics.

Table 11: Using a VAE enables the model to generate meaningful, high-fidelity speech outputs.

Method	FDSD↓	KDSD↓	LSE-C↑	LSE-D↓
Auto-encoder	6.015	5.3	2.002	8.431
Ours (VAE)	**1.273**	**0.2**	**2.507**	**8.155**

A.1.6 Model's variation across speaker attributes. In Table 12, we evaluate the performance of our model across gender of the identities. We automatically classify the LRS2 test set into male and female speakers using a gender detection tool [**?**]. From the table, we can clearly observe that there is no significant variation in performance across gender of the identities.

Table 12: There is no distinctive variation of performance across gender of the speakers.

Gender	LSE-C↑	LSE-D↓	FDSD↓	KDSD↓
Female	2.549	8.138	1.633	0.8
Male	2.424	8.233	1.703	0.8

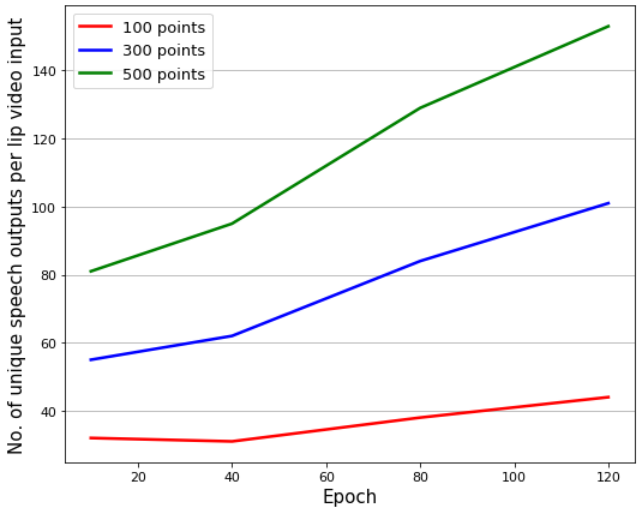

Figure 6: We plot the average number of unique speech outputs generated by our model for every input lip video. We show this "generative strength" [?] at different stages of training, and we can see that our model captures more variations in the latent space as the training progresses.

A.2 Generative Strength of our Lip-to-Speech Model

Lip to speech synthesis is a highly ambiguous task, and many speech outputs are possible for the same input lip sequence. For instance, the variations in voice, speech amplitude, intonation, prosody, and emotion do not clearly correlate with the lips. Moreover, the content to generate is also ambiguous due to the presence of homophenes. Our VAE-GAN model is the first architecture that is capable of modeling these variations in the latent space. In fact, for the same input lip sequence, we can generate different output speech sequences, by sampling different points from the lip distribution.

Evaluating the generative capabilities of a VAE/GAN has been explored in a previous work [**?**], and we adopt the same metric here. The "Generative Strength" metric determines the average percentage of unique speech samples generated for every input lip sequence. To compute this metric, we sample N points ($N = 100, 300, 500$) for each input lip video from the test set of LRS2. Hence, we obtain N speech outputs for every LRS2 test video. We consider a generated spectrogram output as "unique", if its L2 distance from the remaining $N - 1$ spectrograms is at least $\delta = 0.5$. Note that this is a high enough L2 threshold, as the generated audio samples are clearly distinct to hear. In Figure 6, we plot the average count of unique speech outputs per input lip video at different stages of the model training. We can see that our model captures more variations for the same input lip sequence as the training progresses.

A.3 Lip2Wav fails to learn the attention alignment

Lip2Wav [36] is a sequence-to-sequence model with attention, and was proposed for the *single-speaker* lip-to-speech task. The attention mechanism allows the decoder to look at the correct frame's lip movements while decoding. The model learns diagonal attention [36] upon convergence.

We observed that when training on challenging sentence-level, multi-speaker datasets such as LRS2 [13] with a vast number of voices and vocabulary, it fails to learn the temporal attention alignment. We mention this in the main paper, but we also add the final alignment plots of the trained model in Fig 7 to clearly show the failure of this model to learn audio-visual correspondence in such unconstrained settings. The public code[3] provided by the authors is used for this implementation.

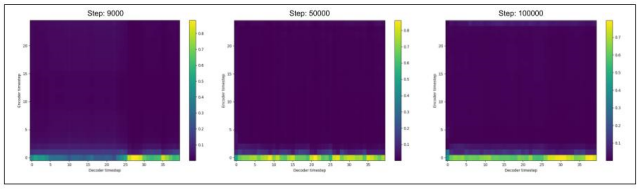

Figure 7: Attention alignment plots at various training stages indicating that Lip2Wav [36] fails to learn temporal attention in unconstrained multi-speaker setting.

[3] github.com/Rudrabha/Lip2Wav