Learning Individual Speaking Styles for Accurate Lip to Speech Synthesis

by

K R Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, C V Jawahar

in

Computer Vision and Pattern Recognition,CVPR
: 1
-10

Report No: IIIT/TR/2020/-1

Centre for Visual Information Technology
International Institute of Information Technology
Hyderabad - 500 032, INDIA
May 2020

Learning Individual Speaking Styles for Accurate Lip to Speech Synthesis

K R Prajwal* IIIT, Hyderabad
Rudrabha Mukhopadhyay* IIIT, Hyderabad
Vinay P. Namboodiri IIT, Kanpur
C V Jawahar IIIT, Hyderabad

{prajwal.k,radrabha.m}@research.iiit.ac.in, vinaypn@iitk.ac.in, jawahar@iiit.ac.in

Abstract

Humans involuntarily tend to infer parts of the conversation from lip movements when the speech is absent or corrupted by external noise. In this work, we explore the task of lip to speech synthesis, i.e., learning to generate natural speech given only the lip movements of a speaker. Acknowledging the importance of contextual and speaker-specific cues for accurate lip-reading, we take a different path from existing works. We focus on learning accurate lip sequences to speech mappings for individual speakers in unconstrained, large vocabulary settings. To this end, we collect and release a large-scale benchmark dataset, the first of its kind, specifically to train and evaluate the single-speaker lip to speech task in natural settings. We propose a novel approach with key design choices to achieve accurate, natural lip to speech synthesis in such unconstrained scenarios for the first time. Extensive evaluation using quantitative, qualitative metrics and human evaluation shows that our method is four times more intelligible than previous works in this space.

1. Introduction

Babies actively observe the lip movements of people when they start learning to speak [24]. As adults, we pay high attention to lip movements to "visually hear" the speech in highly noisy environments. Facial actions, specifically the lip movements, thus reveal a useful amount of speech information. This fact is also exploited by individuals hard of hearing, who often learn to lip read their close acquaintances over time [15] to engage in more fluid conversations. Naturally, the question arises as to whether a model can learn to voice the lip movements of a speaker by "observing" him/her speak for an extended period. Learning such a model would only require videos of people talking with no further manual annotation. It also has a variety of practical applications such as (i) video conferencing in silent environments, (ii) high-quality speech recovery from background noise [1], (iii) long-range listening for surveillance and (iv) generating a voice for people who cannot produce voiced sounds (aphonia). Another interesting application would be "voice inpainting" [41], where the speech generated from lip movements can be used in place of a corrupted speech segment.

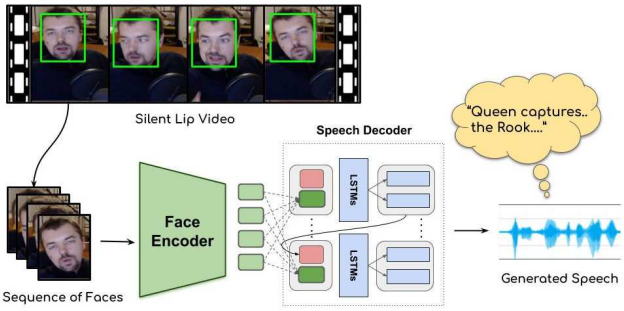

Figure 1. We propose "Lip2Wav": a sequence-to-sequence architecture for accurate speech generation from silent lip videos in unconstrained settings for the first time. The text in the bubble is manually transcribed and is shown for presentation purposes.

Inferring the speech solely from the lip movements, however, is a notoriously difficult task. A major challenge [5, 10] is the presence of homophenes: multiple sounds (phonemes) that are auditorily distinct being perceptually very similar with almost identical lip shapes (viseme). For instance, the lip shape when uttering the phoneme /p/ (park) can be easily confused with /b/ (bark), and /m/ (mark). As a matter of fact, only 25% to 30% of the English language is discernible through lip-reading alone [8, 15, 25, 26]. This implies that professional lip readers do not only lip-read but also piece together multiple streams of information: the familiarity with their subjects, the topic being spoken about, the facial expressions and head gestures of the subject and also their linguistic knowledge. In contrast to this fact, contemporary works in lip to speech and lip to text take a drastically different approach.

Recent attempts in lip to text [2, 5] learn from unconstrained, large vocabulary datasets with thousands of speakers. However, these datasets only contain about 2 minutes

*Both authors have contributed equally to this work.

of data per speaker which is insufficient for models to learn concrete speaker-specific contextual cues essential for lip-reading. The efforts in lip to speech also suffer from a similar issue, but for a different reason. These works are constrained to small datasets [7] with narrow vocabulary speech in artificially constrained environments.

In this work, we explore the problem of lip to speech synthesis from a unique perspective. We take inspiration from the fact that deaf individuals or professional lip readers find it easier to lip read people who they frequently interact with. Thus, rather than attempting lip to speech on random speakers in the wild, we focus on learning speech patterns of a specific speaker by simply observing the person's speech for an extended period. We explore the following question from a data-driven learning perspective: *"How accurately can we infer an individual's speech style and content from his/her lip movements?"*.

To this end, we collect and publicly release a 120-hour video dataset of 5 speakers uttering natural speech in unconstrained settings. Our Lip2Wav dataset contains $800\times$ more data per speaker than the current multi-speaker datasets [2] to facilitate accurate modeling of speaker-specific audio-visual cues. The natural speech is spread across a diverse vocabulary[1] that is about $100\times$ larger than the current single speaker lip to speech datasets [7, 13]. To the best of our knowledge, our dataset is the only publicly available large-scale benchmark to evaluate single-speaker lip to speech synthesis in unconstrained settings. With the help of this dataset, we develop Lip2Wav, a sequence-to-sequence model to generate accurate, natural speech that matches the lip movements of a given speaker. We support our results through extensive quantitative and qualitative evaluations and ablation studies. Our key contributions are as follows:

- We investigate the problem of silent lip videos to speech generation in large vocabulary, unconstrained settings for the first time.
- We release a novel 120-hour person-specific Lip2Wav dataset specifically for learning accurate lip to speech models of individual speakers. Each speaker contains $80\times$ more data with $100\times$ larger vocabulary than its counterparts. The speech is uttered in natural settings with no restriction to head pose or sentence lengths.
- Our sequence-to-sequence modeling approach produces speech that is almost $4\times$ more intelligible in unconstrained environments compared to the previous works. Human evaluation studies also show that our generated speech is more natural with rich prosody.

We release the data, code, trained models publicly for future research along with a demonstration video here[2]. The rest of the paper is organized as follows: In Section 2, we survey the recent developments in this space. Following this, we describe our novel Lip2Wav dataset in Section 3. Our approach and training details are explained in Sections 4 and 5. We evaluate and compare our model with previous works in Section 6. We perform various ablation studies in Section 7 and conclude our work in Section 8.

[1] only words with frequency > 4 are considered

2. Related Work

2.1. Lip to Speech Generation

While initial approaches [20, 23] to this problem extracted the visual features from sensors or active appearance models, the recent works employ end-to-end approaches. Vid2Speech [10] and Lipper [22] generate low-dimensional LPC (Linear Predictive Coding) features given a short sequence of K frames ($K < 15$). The facial frames are concatenated channel-wise and a 2D-CNN is used to generate the LPC features. We show that this architecture is very inadequate to model real-world talking face videos that contain significant head motion, silences and large vocabularies. Further, the low dimensional LPC features used in these works do not contain a significant amount of speech information leading to robotic, artificial sounding speech.

A follow-up work [9] of Vid2Speech does away with LPC features and uses high-dimensional melspectrograms along with optical flows to force the network to explicitly condition on lip motion. While this can be effective for the GRID corpus that has no head movements, optical flow could be a detrimental feature in unconstrained settings due to large head pose changes. Another work [36] strives for improved speech quality by generating raw waveforms using GANs. However, both these works do not make use of the well-studied sequence-to-sequence paradigm [31] that is used for text-to-speech generation [30]; thus leaving a large room for improvement in speech quality and correctness.

Finally, all the above works show results primarily on the GRID corpus [7] which has a very narrow vocabulary of 56 tokens and very minimal head motion. We are the first to study this problem in a large vocabulary setting with thousands of words and sentences. Our datasets are collected from YouTube video clips and hence contain a significant amount of natural speech variations and head movements. This makes our results directly relevant to several real-world applications.

2.2. Lip to Text Generation

In this space as well, several works [6, 28, 37, 38] are limited to narrow vocabularies and small datasets, however, unlike lip to speech, there have been multiple works [2, 5]

[2] cvit.iiit.ac.in/research/projects/cvit-projects/speaking-by-observing-lip-movements

Figure 2. Our Lip2Wav dataset contains talking face videos of 5 speakers from chess analysis and lecture videos. Each speaker has about 20 hours of YouTube video content spanning a rich vocabulary of 5000+ words.

that specifically tackle open vocabulary lip to text in-the-wild. They employ transformer sequence-to-sequence [35] models to generate sentences given a silent lip movement sequence. These works also highlight multiple issues in the space of lip-reading, particularly the inherent ambiguity and hence the importance of using a language model. Our task at hand is arguably harder, as we not only have to infer the linguistic content, but also generate with rich prosody in the voice of the target speaker. Thus, we focus on extensively analyzing the problem in a *single-speaker* unconstrained setting, and learning precise individual speaking styles.

2.3. Text to Speech Generation

Over the recent years, neural text-to-speech models [27, 30] have paved the way to generate high-quality natural speech conditioned on any given text. Using sequence-to-sequence learning [31] with attention mechanisms, they generate melspectrograms in an auto-regressive manner. In our work, we propose Lip2Wav, a modified version of Tacotron 2 [30] that conditions on face sequences instead of text.

3. Speaker-specific Lip2Wav Dataset

The current datasets for lip to speech (or) text are at the two opposite ends of the spectrum: (i) small, constrained narrow vocabulary like GRID [7], TCD-TIMIT [13] or (ii) unconstrained, open vocabulary multi-speaker like LRS2 [2], LRW [6] and LRS3 [3]. The latter class of datasets contains only about 2 - 5 minutes of data per speaker, making it significantly harder for models to learn speaker-specific visual cues that are essential for inferring accurate speech from lip movements. Further, the results would also be directly affected by the existing challenges of multi-speaker speech synthesis [11, 19]. In the other extreme, the single-speaker lip to speech datasets [7, 13], do not emulate the natural settings as they are constrained to narrow vocabularies and artificial environments. Thus, both of these extreme cases do not test the limits of unconstrained single-speaker lip to speech synthesis.

We introduce a new benchmark dataset for unconstrained lip to speech synthesis that is tailored towards exploring the following line of thought: *How accurately can we infer an individual's speech style and content from his/her lip movements?* To create the Lip2Wav dataset, we collect a total of about 120 hours of talking face videos across 5 speakers. The speakers are from various online lecture series and chess analysis videos. We choose English as the sole language of the dataset. With about 20 hours of natural speech per speaker and vocabulary sizes over 5000 words[3] for each of them, our dataset is significantly more unconstrained and realistic than GRID [7] or TIMIT [13] datasets. It is thus ideal for learning and evaluating accurate person-specific models for the lip to speech task. Table 1 compares the features of our Lip2Wav dataset with other standard single-speaker lip-reading datasets. Note that a word is included in the vocabulary calculation for Table 1 only if its frequency in the dataset is at least five.

Dataset	Num. speakers	Total #hours	#hours per speaker	Vocab per speaker	Natural setting?
GRID [7]	34	28	0.8	56	×
TIMIT [13]	3	1.5	0.5	82	×
Lip2Wav (Ours)	5	**120**	**≈ 20**	**≈ 5K**	✓

Table 1. The Lip2Wav dataset is the first large-scale dataset tailored towards acting as a reliable benchmark for single-speaker lip to speech synthesis.

4. Lip to Speech Synthesis in the Wild

Given a sequence of face images $I = (I_1, I_2, \ldots, I_T)$ with lip motion, our goal is to generate the corresponding speech segment $S = (S_1, S_2, \ldots, S_{T'})$. To obtain natural speech in unconstrained settings, we make numerous key design choices in our Lip2Wav architecture. Below, we highlight and discuss how they are different from previous methods for lip to speech synthesis.

4.1. Problem Formulation

Prior works in lip to speech regard their speech representation as a 2D-image [10,36] in the case of melspectrograms or as a single feature vector [10] in the case of LPC features.

[3] approximate; texts obtained using Google ASR API

In both these cases, they use a 2D-CNN to decode these speech representations. By doing so, they violate the ordering in which they model the sequential speech data, i.e. the future time steps influence the prediction of the current time step. In contrast, we formulate this problem in the standard sequence-to-sequence learning paradigm [31]. Concretely, each output speech time-step S_k is modelled as a conditional distribution of the previous speech time-steps $S_{<k}$ and the input face image sequence $I = (I_1, I_2, \ldots, I_T)$. The probability distribution of each output speech time-step is given by:

$$P(S|I) = \Pi_k(S_k|S_{<k}, I) \quad (1)$$

Lip2Wav, as shown in Figure 3 consists of two modules: (i) Spatio-temporal face encoder (ii) Attention-based speech decoder. The modules are trained jointly in an end-to-end fashion. The sequence-to-sequence approach enables the model to learn an implicit speech-level language model that helps it to disambiguate homophenes.

4.2. Speech Representation

There are multiple output representations from which we can recover intelligible speech, but each of them have their trade-offs. The LPC features are low-dimensional and easier to generate, however, they result in robotic, artificial sounding speech. At the other extreme [36], one can generate raw waveforms but the high dimensionality of the output (16000 samples per second) makes the network training process computationally inefficient. We take inspiration from previous text-to-speech works [27, 30] and generate melspectrograms conditioned on lip movements. We sample the raw audio at 16kHz. The window-size, hop-size and mel dimension are $800, 200$, and 80 respectively.

4.3. Spatio-temporal Face Encoder

Our visual input is a short video sequence of face images. The model must learn to extract and process the fine-grained sequence of lip movements. 3D convolutional neural networks have been shown to be effective [18, 33, 36] in multiple tasks involving spatio-temporal video data. In this work, we try to encode the spatio-temporal information of the lip movements using a stack of 3D convolutions (Figure 3). The input to our network is a sequence of facial images of the dimension $T \times H \times W \times 3$, where T is the number of time-steps (frames) in the input video sequence, H, W correspond to the spatial dimensions of the face image. We gradually down-sample the spatial extent of the feature maps and preserve the temporal dimension T. We also employ residual skip connections [14] and batch normalization [16] throughout the network. The encoder outputs a single D-dimensional vector for each of the T input facial images to get a set of spatio-temporal features $T \times D$ to be passed to the speech decoder. Each time-step of the

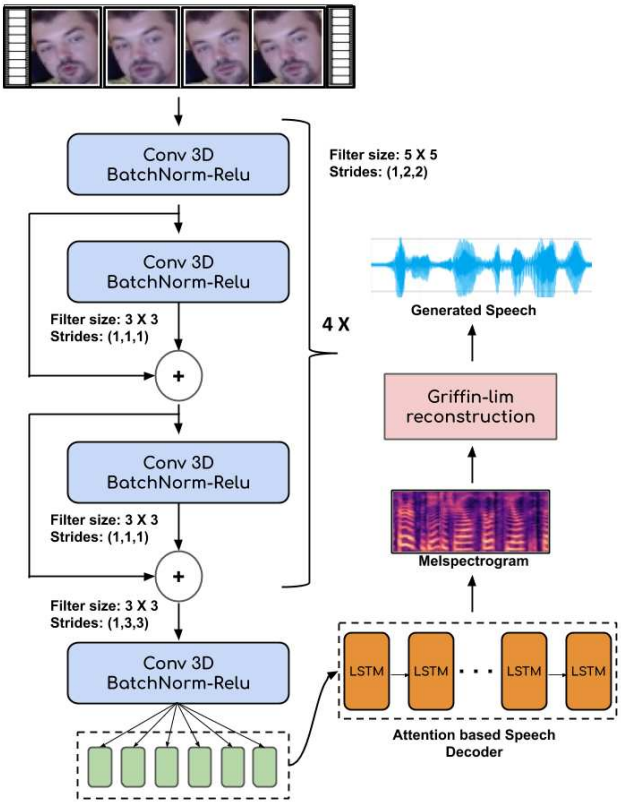

Figure 3. Lip2Wav model for lip to speech synthesis. The spatio-temporal encoder is a stack of 3D convolutions to extract the sequence of lip movements. This is followed by a decoder adapted from [30] for high-quality speech generation. The decoder is conditioned on the face image features from the encoder and generates the melspectrogram in an auto-regressive fashion.

embedding generated from the encoder also contains information about the future lip movements and hence helps in the subsequent generation.

4.4. Attention-based Speech Decoder

To achieve high-quality speech generation, we exploit the recent breakthroughs [27, 30] in text-to-speech generation. We adapt the Tacotron 2 [30] decoder which has been used to generate melspectrograms conditioned on text inputs. For our work, we condition the decoder on the encoded face embeddings from the previous module. We refer the reader to the Tacotron 2 [30] paper for more details about the decoder. The encoder and decoder are trained end-to-end by minimizing the L1 reconstruction loss between the generated and the ground-truth melspectrogram.

4.5. Gradual Teacher Forcing Decay

In the initial stages of training, up to $\approx 30K$ iterations, we employ teacher forcing similar to the text-to-speech counterpart. We hypothesize that this enables the decoder to

learn an implicit speech-level language model to help in disambiguating homophenes. Similar observations are made in lip to text works [2] which employ a transformer-based sequence-to-sequence model. Over the course of the training, we gradually decay the teacher forcing to enforce the model to attend to the lip region and to prevent the implicit language model from over-fitting to the train set vocabulary. We examine the effect of this decay in sub-section 7.3.

4.6. Context Window Size

The size of the visual context window for inferring the current speech time-step helps the model to disambiguate homophenes [10]. We employ about $6\times$ larger context size than prior works and show in sub-section 7.1 that this design choice results in significantly more accurate speech.

5. Benchmark Datasets and Training Details

5.1. Datasets

The primary focus of our work is on single-speaker lip to speech synthesis in unconstrained, large vocabulary settings. For the sake of comparison with previous works, we also train the Lip2Wav model on the GRID corpus [7] and the TCD-TIMIT lip speaker corpus [13]. Next, we train on all five speakers of our newly collected speaker-specific Lip2Wav dataset. Unless specified, all the datasets are divided into 90-5-5% train, validation and unseen test splits. In the Lip2Wav dataset, we create these splits using different videos ensuring that no part of the same video is used for both training and testing. The train and test splits are also released for fair comparison in future works.

5.2. Training Methodology and Hyper-parameters

We prepare a training input example by randomly sampling a contiguous sequence of 3 seconds which corresponds to $T = 75$ or $T = 90$ depending on the frame rate (FPS) of the video. The effect of various context window sizes is studied in Section 7.1. We detect and crop the face from the video frames using the S^3FD face detector [40]. The face crops are resized to 48×48. The melspectrogram representation of the audio corresponding to the chosen short video segment is used as the desired ground-truth for training. For training on small datasets like GRID and TIMIT, we halve the hidden dimension to prevent over-fitting. We set the training batch size to 32 and train until the mel reconstruction loss plateaus for at least $30K$ iterations. In our experiments for unconstrained single-speaker, convergence was achieved in about $200K$ iterations. The optimizer used is Adam [21] with an initial learning rate of 10^{-3}. The model with the best performance on the validation set is chosen for testing and evaluation. More details, specifically a few minor speaker-specific hyper-parameter changes can be found in the publicly released code[2].

5.3. Speech Generation at Test Time

During inference, we provide only the sequence of lip movements and generate the speech in an auto-regressive fashion. Note that we can generate speech for any length of lip sequences. We simply take consecutive T second windows and generate the speech for each of them independently and concatenate them together. We also maintain a small overlap across the sliding windows to adjust for boundary effects. We obtain the waveform from the generated melspectrogram using the standard Griffin-Lim algorithm [12]. We observed that neural vocoders [34] perform poorly in our case as our generated melspectrograms are significantly less accurate than state-of-the-art TTS systems. Finally, the ability to generate speech for lip sequences of any length is worth highlighting as the performance of the current lip-to-text works trained at sentence-level deteriorates sharply for long sentences that barely last over just 4 - 5 seconds [2].

6. Experiments and Results

We obtain results from our Lip2Wav model on all the test splits as described above. For comparing related work, we use the open-source implementations provided by the authors if available or re-implement a version ourselves. We compare our models with the previous lip to speech works using three standard speech quality metrics: Short-Time Objective Intelligibility (STOI) [32] and Extended Short-Time Objective Intelligibility (ESTOI) [17] for estimating the intelligibility and Perceptual Evaluation of Speech Quality (PESQ) [29] to measure the quality. Using an out-of-the-box ASR system[4], we obtain textual transcripts for our generated speech and evaluate our speech results using word error rates (WER) for the GRID [7] and TCD-TIMIT lip speaker corpus [13]. We, however do not compute WER for the proposed Lip2Wav corpus due to the lack of text transcripts. We also perform human evaluation and report the mean opinion scores (MOS) for the proposed Lip2Wav model and the competing methods. Next, we also perform extensive ablation studies for our approach and report our observations. Finally, as we achieve superior results compared to previous works in single-speaker settings, we end the experimental section by also reporting baseline results for word-level multi-speaker lip to speech generation using the LRW [6] dataset and highlight its challenges as well.

6.1. Lip to Speech in Constrained Settings

We start by evaluating our approach against previous lip to speech works in constrained datasets, namely the GRID [7] corpus and TCD-TIMIT lip speaker corpus [13]. For the GRID dataset, we report the mean test scores for 4 speakers which are also reported in the previous works.

[4]Google Speech-to-Text API

Tables 2 and 3 summarize the results for GRID and TIMIT datasets respectively.

Method	STOI	ESTOI	PESQ	WER
Vid2Speech [10]	0.491	0.335	1.734	44.92%
Lip2AudSpec [4]	0.513	0.352	1.673	32.51%
GAN-based [36]	0.564	0.361	1.684	26.64%
Ephrat et al. [9]	0.659	0.376	**1.825**	27.83%
Lip2Wav (ours)	**0.731**	**0.535**	1.772	**14.08%**

Table 2. Objective speech quality, intelligibility and WER scores for the GRID dataset unseen test split.

Method	STOI	ESTOI	PESQ	WER
Vid2Speech [10]	0.451	0.298	1.136	75.52%
Lip2AudSpec [4]	0.450	0.316	1.254	61.86%
GAN-based [36]	0.511	0.321	1.218	49.13%
Ephrat et al. [9]	0.487	0.310	1.231	53.52%
Lip2Wav (ours)	**0.558**	**0.365**	**1.350**	**31.26%**

Table 3. Objective speech quality, intelligibility and WER scores for the TCD-TIMIT dataset unseen test split.

As we can see, our approach outperforms competing methods across all objective metrics by a significant margin. The difference is particularly visible in the TIMIT [13] dataset, where the test set contains a lot of novel words unseen during training. This shows that our model learns to capture correlations across short phoneme sequences and pronounces new words better than previous methods.

6.2. Lip to Speech in Unconstrained Settings

We now move on to evaluating our approach in unconstrained datasets that contain a lot of head movements and much broader vocabularies. They also contain a significant amount of silences or pauses between words and sentences. It is here that we see a more vivid difference in our approach compared to previous approaches. We train our model independently on all 5 speakers of our newly collected Lip2Wav dataset. The training details are mentioned in the sub-section 5.2. For comparison with previous works, we choose the best performing models [9, 36] on the TIMIT dataset based on STOI scores and report their performance after training on our Lip2Wav dataset. We compute the same metrics for speech intelligibility and quality that are used in Table 3. The scores for all five speakers for our method and the two competing methods across all three metrics are reported in Table 4.

Our approach produces much more intelligible and natural speech across different speakers and vocabulary sizes. Notably, our model has more accurate pronunciation, as can be seen in the increased STOI and ESTOI scores compared to the previous works.

Method	Speaker	STOI	ESTOI	PESQ
GAN-based [36]	*Chemistry Lectures*	0.192	0.132	1.057
Ephrat et al. [9]		0.165	0.087	1.056
Lip2Wav (ours)		**0.416**	**0.284**	**1.300**
GAN-based [36]	*Chess Analysis*	0.195	0.104	1.165
Ephrat et al. [9]		0.184	0.098	1.139
Lip2Wav (ours)		**0.418**	**0.290**	**1.400**
GAN-based [36]	*Deep Learning*	0.144	0.070	1.121
Ephrat et al. [9]		0.112	0.043	1.095
Lip2Wav (ours)		**0.282**	**0.183**	**1.671**
GAN-based [36]	*Hardware Security*	0.251	0.110	1.035
Ephrat et al. [9]		0.192	0.064	1.043
Lip2Wav (ours)		**0.446**	**0.311**	**1.290**
GAN-based [36]	*Ethical hacking*	0.171	0.089	1.079
Ephrat et al. [9]		0.143	0.064	1.065
Lip2Wav (ours)		**0.369**	**0.220**	**1.367**

Table 4. In unconstrained single-speaker settings, our Lip2Wav model achieves almost $4\times$ more intelligible speech than the previous methods.

6.3. Human Evaluation

In addition to speech quality and intelligibility metrics, it is important to manually evaluate the speech as these metrics are not perfect [9] measures.

6.3.1 Objective Human Evaluation

In this study, we ask the human participants to manually identify and report (A) the percentage of mispronunciations, (B) the percentage of word skips and (C) the percentage of mispronunciations that are homophenes. Word skips denotes the number of words that are either completely unintelligible due to noise or slurry speech. We choose 10 predictions from the unseen test split of each speaker in our Lip2Wav dataset to get a total of 50 files. We report the mean numbers of (A), (B), and (C) in Table 5.

Model	(A)	(B)	(C)
GAN-based [36]	36.6%	24.3%	63.8%
Ephrat et al [9]	43.3%	27.5%	60.7%
Lip2Wav (ours)	**21.5%**	**8.6%**	**49.8%**

Table 5. Objective Human evaluation results. The participants manually identified the percentage of (A) Mispronunciations, (B) Word skips and (C) Homophene-based errors in the test samples.

Our approach makes far fewer mispronunciations than the current state-of-the-art method. It also skips words $3\times$ lesser, however, the key point to note is that the issue of homophenes is still a dominant cause for errors in all cases indicating there is still scope for improvement in this area.

6.3.2 Subjective Human Evaluation

We ask 15 participants to rate the different approaches for unconstrained lip to speech synthesis on a scale of $1-5$ for each of the following criteria: (i) *Intelligibility* and (ii) *Naturalness* of the generated speech. Using 10 samples of generated speech for each of the 5 speakers from our Lip2Wav dataset, we compare the following approaches: (i) Our Lip2Wav model (ii) Current state-of-the-art lip to speech models [9, 36] (iii) Manually transcribed text followed by a multi-speaker TTS [19, 30] to show that even with the most accurate text, lip to speech is not a concatenation of lip-to-text and text-to-speech. And finally, (iv) Human speech is also added for reference. In all the cases, we overlay the speech on the face video before showing it to the rater. The mean scores are reported in Table 6.

Approach	Intelligibility	Naturalness
GAN-based [36]	1.56	1.71
Ephrat et al. [9]	1.34	1.67
Lip2Wav (ours)	**3.04**	**3.63**
MTT + TTS [30]	3.86	3.15
Actual Human Speech	4.82	4.95

Table 6. Mean human evaluation scores based on speech quality and intelligibility for various approaches for lip to speech. MTT denotes "manually-transcribed text". The penultimate row simulates the best possible case of automatic lip to text followed by a state of the art text-to-speech system. The drop in naturalness score in this case illustrates the loss in speech style and prosody.

In line with the previous evaluations, we can see that our approach produces significantly higher quality and legible speech compared to the previous state-of-the-art [36]. It is also evident that generating the speech from the text that is read from lip movements (lip to text), cannot achieve the desired prosody and naturalness even if the text is fully accurate. Further, this method will also cause the lips and audio to be out of sync. Thus, our approach is currently the best method to produce natural speech from lip movements.

6.4. Multi-speaker Word-level Lip to Speech

Given the superior performance of our Lip2Wav approach on single-speaker lip to speech, we also obtain baseline results on the highly challenging problem of multi-speaker lip to speech synthesis for random identities. Note that the focus of the work is still primarily on single-speaker lip to speech. We adapt the approach presented in [19] and feed a speaker embedding as input to our model. We report our baseline results on the LRW [6] dataset intended for word-level lip-reading, i.e. it is used to measure the performance of recognizing a single word in a given short phrase of speech. We do not demonstrate on the LRS2 dataset [5] as its clean train set contains just 29 hours of data, which is quite small for multi-speaker speech generation. For instance, multi-speaker text-to-speech generation datasets [39] containing a similar number of speakers contain several hundreds of hours of speech data.

In Table 7, we report the speech quality and intelligibility metrics achieved by our multi-speaker Lip2Wav model on the LRW test split. As none of the previous works in lip to speech tackle the multi-speaker case, we do not make any comparisons with them. We also report the WER by getting the text using the Google ASR API. For comparison, we also report the WER of the baseline lip to text work on LRW [6]. Note that the speech metric scores shown in Table 7 for word-level lip to speech cannot be directly compared with the single-speaker case which contains word sequences of various lengths along with pauses and silences.

Method	STOI	ESTOI	PESQ	WER
Lip2Wav (Ours)	0.543	0.344	1.197	**34.2%**
Chung et al. [6]	NA	NA	NA	38.8%

Table 7. Objective speech quality and intelligibility scores on the LRW dataset. WER is also calculated after using an ASR on the generated speech. Our model outperforms the baseline method proposed in [6], without any text-level supervision. The speech metrics are not applicable for [6] as it is a lip to text work.

We end our experimental section here. Apart from showing significant increases in performance from previous lip to speech works, we also achieve word-level multi-speaker lip to speech synthesis. In the next section, we conduct ablation studies on our model.

7. Ablation Studies

In this section, we probe different aspects of our Lip2Wav approach. All results in this section are calculated using the unseen test predictions on the "Hardware Security" speaker of our Lip2Wav dataset.

7.1. Larger context window helps in disambiguation

As stated before, the lip to speech task is highly ambiguous to be inferred solely from lip movements. One of the ways to combat this, is to provide reasonably large context information to the model to disambiguate a given viseme. Previous works, however, use only about $0.3-0.5$ seconds of context. In this work, we use close to $6\times$ this number and provide a context of 3 seconds. This helps the model to disambiguate by learning co-occurrences of phonemes and words and the resulting improvement is evident in Table 8.

7.2. Model is Highly Attentive to the Mouth

We plot the activations of the penultimate layer of the spatio-temporal face encoder in Figure 4 to show that our encoder is highly attentive towards the mouth region of the speaker. The attention alignment curve in Figure 5 shows

Context Window size	STOI	ESTOI	PESQ
0.5 seconds	0.264	0.193	1.062
1.5 seconds	0.321	0.226	1.080
3 seconds	**0.446**	**0.311**	**1.290**

Table 8. Larger context information consistently results in more accurate speech generation. We limit the window size to 3 seconds due to memory constraints.

that the decoder conditions on the appropriate video frame's lips while generating the corresponding speech.

Figure 4. We plot the activations of the penultimate layer of the face encoder and the attention alignment from the decoder. We see that the face encoder is highly attentive towards the mouth region.

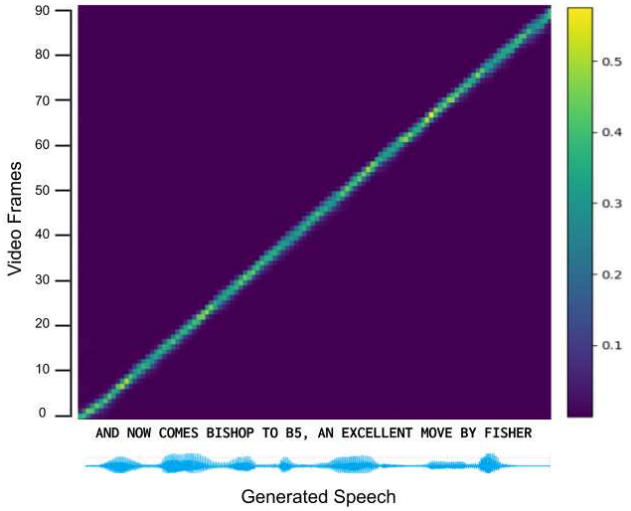

Figure 5. The decoder alignment curve illustrates that the model is generating speech by strongly conditioning on the corresponding lip movements.

7.3. Teacher Forcing vs Non-Teacher Forcing

To accelerate the training of a sequence-to-sequence architecture, typically, the previous time step's ground-truth (instead of the generated output) is given as input to the current time-step. While this is highly beneficial in the initial stages of training, we observed that gradually decaying the teacher forcing from $\approx 30K$ iterations significantly improves results and prevents over-fitting to the train vocabulary. A similar improvement is also observed in lip to text works [2]. In Table 9, we show the significant improvement in test scores by gradually decaying teacher forcing.

Teacher-forcing	STOI	ESTOI	PESQ
Always forced	0.221	0.162	1.141
Gradual decay	**0.446**	**0.311**	**1.290**

Table 9. Gradually decaying the teacher forcing enables the model to generalize to unseen vocabulary by forcing it to look at the visual input and not just predict from the previously uttered speech.

7.4. Effect of Different Visual Encoders

While using a 3D-CNN worked best in our experiments to capture both the spatial and temporal information in unconstrained settings, we also report in Table 10 the effect of using different kinds of encoders. We replace the encoder module while keeping the speech decoder module intact. We see that the best performance is obtained with a 3D-CNN encoder.

Encoder	STOI	ESTOI	PESQ
2D-CNN	0.291	0.211	1.112
2D-CNN + 1D-CNN	0.298	0.223	1.170
3D-CNN (ours)	**0.446**	**0.311**	**1.290**

Table 10. Our Lip2Wav model employs a 3D-CNN encoder to capture the spatio-temporal visual information and is the superior choice over the other alternatives.

8. Conclusion

In this work, we investigated the problem of synthesizing speech based on lip movements. We specifically solved the problem by focusing on individual speakers. We did this in a data-driven learning approach by creating a large-scale benchmark dataset for unconstrained, large vocabulary single-speaker lip to speech synthesis. We formulate the task at hand as a sequence-to-sequence problem, and show that by doing so, we achieve significantly more accurate and natural speech than previous methods. We evaluate our model with extensive quantitative metrics and human studies. All the code and data for our work has been made publicly available[2]. Our work opens up several new directions. One of them would be to examine related works in this space such as lip to text generation from a speaker-specific perspective. Similarly, explicitly addressing the dominant issue of homophenes can yield more accurate speech. Generalizing to vocabulary outside the typical domain of the speaker can be another fruitful venture. We believe that exploring some of the above problems in a data-driven fashion could lead to further useful insights in this space.

References

[1] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. The conversation: Deep audio-visual speech enhancement. *arXiv preprint arXiv:1804.04121*, 2018.

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

[3] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. 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. Lip2audspec: Speech reconstruction from silent lip movements video. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2516–2520, 2017.

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

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

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

[8] David A Ebert and Paul S Heckerling. Communication with deaf patients: knowledge, beliefs, and practices of physicians. *Jama*, 273(3):227–229, 1995.

[9] Ariel Ephrat, Tavi Halperin, and Shmuel Peleg. Improved speech reconstruction from silent video. *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 455–462, 2017.

[10] Ariel Ephrat and Shmuel Peleg. Vid2speech: speech reconstruction from silent video. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5095–5099. IEEE, 2017.

[11] Andrew Gibiansky, Sercan Arik, Gregory Diamos, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou. Deep voice 2: Multi-speaker neural text-to-speech. In *Advances in neural information processing systems*, pages 2962–2970, 2017.

[12] Daniel W. Griffin and Jae S. Lim. Signal estimation from modified short-time fourier transform. In *ICASSP*, 1983.

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

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[15] Lisa I Iezzoni, Bonnie L O'Day, Mary Killeen, and Heather Harker. Communicating about health care: observations from persons who are deaf or hard of hearing. *Annals of Internal Medicine*, 140(5):356–362, 2004.

[16] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 448–456, 2015.

[17] Jesper Jensen and Cees Taal. An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24:1–1, 11 2016.

[18] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2012.

[19] Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu, et al. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In *Advances in neural information processing systems*, pages 4480–4490, 2018.

[20] Christopher T Kello and David C Plaut. A neural network model of the articulatory-acoustic forward mapping trained on recordings of articulatory parameters. *The Journal of the Acoustical Society of America*, 116(4):2354–2364, 2004.

[21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

[23] Thomas Le Cornu and Ben Milner. Generating intelligible audio speech from visual speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(9):1751–1761, 2017.

[24] David J Lewkowicz and Amy M Hansen-Tift. Infants deploy selective attention to the mouth of a talking face when learning speech. *Proceedings of the National Academy of Sciences*, 109(5):1431–1436, 2012.

[25] Christine Chong-hee Lieu, Georgia Robins Sadler, Judith T Fullerton, and Paulette Deyo Stohlmann. Communication strategies for nurses interacting with patients who are deaf. *Dermatology Nursing*, 19(6):541, 2007.

[26] Helen Margellos-Anast, Teri Hedding, Toby Perlman, Linda Miller, Raymond Rodgers, Lisa Kivland, Dorothea DeGutis, Barbara Giloth, and Steven Whitman. Developing a standardized comprehensive health survey for use with deaf adults. *American Annals of the Deaf*, 150(4):388–396, 2005.

[27] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. Deep voice 3: Scaling text-to-speech with convolutional sequence learning. *arXiv preprint arXiv:1710.07654*, 2017.

[28] Leyuan Qu, Cornelius Weber, and Stefan Wermter. Lipsound: Neural mel-spectrogram reconstruction for lip reading. *Proc. Interspeech 2019*, pages 2768–2772, 2019.

[29] Antony Rix, John Beerends, Michael Hollier, and Andries Hekstra. Perceptual evaluation of speech quality (pesq): A

new method for speech quality assessment of telephone networks and codecs. volume 2, pages 749–752 vol.2, 02 2001.

[30] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783. IEEE, 2018.

[31] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[32] Cees Taal, Richard Hendriks, R. Heusdens, and Jesper Jensen. A short-time objective intelligibility measure for time-frequency weighted noisy speech. pages 4214 – 4217, 04 2010.

[33] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[34] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. In *SSW*, 2016.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[36] Konstantinos Vougioukas, Pingchuan Ma, Stavros Petridis, and Maja Pantic. Video-driven speech reconstruction using generative adversarial networks. *arXiv preprint arXiv:1906.06301*, 2019.

[37] Michael Wand, Jan Koutník, and Jürgen Schmidhuber. Lipreading with long short-term memory. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6115–6119. IEEE, 2016.

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

[39] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. Libritts: A corpus derived from librispeech for text-to-speech. *arXiv preprint arXiv:1904.02882*, 2019.

[40] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z Li. S3fd: Single shot scale-invariant face detector. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 192–201, 2017.

[41] Hang Zhou, Ziwei Liu, Xudong Xu, Ping Luo, and Xiaogang Wang. Vision-infused deep audio inpainting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 283–292, 2019.