Learning Speaker-specific Lip-to-Speech Generation

∗ Munender Varshney‡ , ∗ Ravindra Yadav†, Vinay P. Namboodiri § and Rajesh M Hegde†
† Electrical department, Indian institute of Technology Kanpur, India
‡ Computer Science and Engineering department, Indian institute of Technology Kanpur, India
§University of Bath, UK
Email: munenderv@cse.iitk.ac.in, ravin@iitk.ac.in, vpn22@bath.ac.uk, rhegde@iitk.ac.in

Abstract—Understanding the lip movement and inferring the speech from it is notoriously difficult for the common person. The task of accurate lip-reading gets help from various cues of the speaker and its contextual or environmental setting. Every speaker has a different accent and speaking style, which can be inferred from their visual and speech features. This work aims to understand the correlation/mapping between speech and the sequence of lip movement of individual speakers in an unconstrained and large vocabulary. We model the frame sequence as a distribution of features from the transformer in an auto-encoder setting and learn the embeddings jointly that exploits temporal properties of both audio and video. We learn temporal synchronization using deep metric learning, which guides the decoder to generate speech in sync with input lip movements. The predictive posterior thus gives us the generated speech in a word (bark) then in a similar way like park [1]–[3]. Similarly, there are a huge amount of words in the English language that cannot differentiate based on lip-reading alone [4], [5]. Hence, even professional lip readers depend on multiple information streams such as the topic on which person is speaking, familiarity with the person, linguistic knowledge, head gestures, and facial expression. The lip-reading task is very perplexing, which is supported by the fact that a human can adequately understand speech across various accents and in a noisy environment. However, at the same time, they perform comparatively poorly on lip-reading tasks [6], [7].

Nevertheless, lip-reading has a variety of practical applications in speech transcription where audio is either missing or noisy. Some of the applications are long-range listening in surveillance video, efficient speech recovery in a noisy environment [8], re-dubbing and transcribing the archival silent films, and a silent speech control system for resolving simultaneous speech from multi-talker [9]. The “voice inpainting” [10] generate speech from lip movement to replace a corrupted speech segment. Lip-reading also has crucial applications in helping speech-impaired people, such as hearing aids and generation of voice for people who communicate using lip movements only as they are unable to generate voice (aphonia). The advent of deep neural network models [11]–[15] and the availability of huge labeled datasets have helped in achieving a significant improvement in the task of the lip to text generation to solve a great variety of applications [6], [16]. These datasets consist of a large vocabulary in an unconstrained environment with thousands of speakers. However, this conventional task needs data to be annotated by the human [17](i.e., transcribe speech into text format). On the other hand, the lip to speech generation task does not require any annotations; hence it has drawn considerable attention as an alternative form of lip reading. Fundamentally, generating speech from a silent video of lip movement can be observed as a function or mapping from visemes to their corresponding phonemes. The mapping is learned by “observing” the lip movement of the speaker for a prolonged period [4] while learning from a short clips of lip movement is challenging due to homophones [18]. Therefore, learning viseme-to-phoneme mapping only need long talking video clip of people without annotation to remove ambiguity using added visual context.

Inspired by various studies [4], [19], professional lip readers and individuals can easily do lip-reading for people with frequent contact, i.e., an infant learns to speak by observing the lip movement of people around. Our focus in this work is the generation of speech from lip movements and contextual cues in the video of an individual in their speaking style. The generation is performed by just observing a particular person’s speech patterns for a prolonged time period, not for any random speaker in the wild. Given the recent success of transformer models [20] in natural language processing and its applications [21]–[26], these models are applied in various problem domains of other modalities such as audio, text,
image and video. In recent times, transformers have exhibited better representational capability as compared to CNNs, thus various design formulation of the transformers with variational auto-encoder models have been applied in various multi-modal domain applications such as music generation [27], story generation [28] and sentiment analysis [29]. We use the transformer model to learn latent distributions from both input modalities jointly and the output embeddings are the sample drawn from these latent space. We used the deep metric learning-based approach proposed in [30] to learn the temporal synchronization of lip movement and speech. We have tested our model on GRID and lip2wav chemistry lecture datasets. It outperforms all the previous state-of-the-art across various evaluation metrics on Lip2Wav Chemistry dataset, having a large vocabulary in an unconstrained setting while marginally outperforms on the GRID dataset.

Section II discusses the recent developments and describes our approach and model architectures in the next section. The training approach and speech prediction are described in Sections V and VI. Finally, the detailed experimental evaluation is discussed in the next section and we conclude the work in Section VIII.

II. Previous Work

There is a vast history of work on lip reading; a nice detailed review of the work in the non-deep learning era is covered in the survey paper [31]. With advent of deep learning in the current era [12], [14] and the development of large-scale datasets [11] for lip-reading such as LRW [6], LRS2 [1] and LRS3 [32], various work have used CNN to do recognition of smallest distinguishable units of sound from still images i.e., phonemes and visemes (visual equivalent of phonemes) [1], [33], [34]. The researchers have proposed various spatiotemporal end-to-end trainable CNN models such as ResNet and BiLSTM for word-level learning [33], [35]. Next to this, the sentence level learning is performed using methodology of modelling automatic speech recognition (ASR) system using sequence-to-sequence models with LSTM or Connectionist Temporal Classification (CTC) [1], [36] i.e., Assael et al. proposed LipNet [7], an end-to-end trainable model which uses CNN and LSTM-based architecture to map a variable-length sequence of frames to text using an auxiliary CTC loss for learning. The model achieves state-of-the-art performance in the speaker-independent setting of constrained grammar with vocabulary size (51 words) in GRID dataset [37]. Other work in this direction uses the recently proposed powerful Transformer-based architectures [38] i.e., Conformer [39], a hybrid architecture or its variants with different convolution blocks [40]. A few papers have explored the idea of knowledge distillation from ASR models, trained on the larger datasets to indirectly train the lip-reading model [41], [42].

The ASR problem needs manual annotations in the form of text (transcription), while generation of speech, a closely linked problem, can be performed directly from lip movement. Preliminary research exploits CNN models to predict auditory features corresponding to a short video clip and replicate the same over a complete silent video using a sliding window to produce entire speech feature [2], [43]. For example, Vid2Speech [2] and Lipper [44] are trained end-to-end to predict LPC (Linear Predictive Coding) features from an input clip of $K$ frame using a 2D-CNN model and then these features are converted into speech. However, these features do not have significant speech information that leads to the generation of robotic speech and the model is inadequate to generate real-world talking face videos. The follow-up paper uses mel-spectrogram, a high-dimensional feature rather than LPC features, and the model uses optical flows with raw frames to enforce conditioning on lip motion. Akbari et al. [45] designed a lip-reading network having seven-layer 3D-CNN with LSTM to capture temporal information and use the generated feature to reconstruct auditory spectrogram from the decoder of the deep auto-encoder network. Researchers have also explored various generative approaches [46], [47], such as Vougioukas et al. [48] uses the 1D GAN model to synthesize raw waveform from a given video. Similarly, Michelsanti et al. [49] uses a deep auto-encoder architecture where the encoder with a recursive module maps video frames of the speaker to vocoder features. There are five decoders for multitasking and the various outputs are used to reconstruct speech from the vocoder.

Different from these work, Prajwal et al. [50] gave a different perspective of mapping speech to lip sequence for an individual by adhering to their speaking style as well. They used a Seq2Seq architecture and a stack of 3D CNN network trained on a large vocabulary in an unconstrained setting. Yadav et al. [51] approaches speaker-specific generation using variational auto-encoder as a stochastic model and a recent work uses audiovisual analogy [18]. Our work addresses the same problem of speech generation from the lip sequence of a specific speaker.

From the trending success of Transformer models [20] in natural language processing domain on various research problems and it’s applications [21]–[26], researchers have used these to solve the problem in other domains, such as image, audio, and video. Various transformers are proposed to handle a set of modalities such as video with text, image with text, and image with depth [52]. These are famous as Multi-modal transformers [53], [54]. These transformers can be used as a building block in a variational auto-encoder setting. This in turn help in generative modelling with multiple input modalities such as text, audio, image and video [27], [28]. For example, Jiang et al. designed a hierarchical model which unifies transformers for understanding long-term dependencies using their powerful attention mechanism and VAE to learn a disentangled latent space for music generation [27]. On the same line of thought researchers has unified VAE with variant of transformers for various other applications such as story generation [28], response generation [55], sentiment analysis [29], and 3D human pose generation [56]. Recently, audio and visual modalities have been used jointly to improve various tasks such as zero-shot learning [57], depth estimation [58] etc.
III. MODEL ARCHITECTURE

In this work, we have proposed the framework to integrate the latent variable model [59] and the transformer [20] to leverage the power of the generative model with the self-attention mechanism. Our model uses multi-modal data as input streams; one mode is a video stream, and the other is audio. The first input modality is video, i.e., a sequence of frames consisting of lip movement, which can be represented as \( \{F_1, F_2, F_3, \ldots, F_t\} \) where \( t \) represents the number of frames and their features are used as input to the video-transformer. At the same time, the other modality, speech, can be represented as \( \{S_1, S_2, S_3, \ldots, S_t\} \) in Mel Frequency Cepstral Coefficients (MFCC) feature space [60] and given as input to the audio transformer. The initial output representation of both encoders is specific to modality. Our hierarchical framework uses transformer models as an abstraction to the variational autoencoder Framework. The output features of both modalities from transformers are used to learn a joint cross-modal relationship between these embedding of the video and the audio sequence. The joint cross-modal relationship also exploits the temporal synchroniztion using deep metric learning [30] and, at the same time, draws distributions of both modalities closer corresponding to the given timestamp.

Our main objective is to learn speech prediction from the given sequence of silent video frames \( \{F_1, F_2, F_3, \ldots, F_t\} \). We designed the framework to learn a decoder implicitly in the VAE framework using transformers as a building block. The speech prediction is performed using learned decoder (another transformer module) in the form of a sequence of audio window represented as \( \{S_1, S_2, S_3, \ldots, S_t\} \). The generation of speech is governed by the transformer module (deep generative network ) \( p_\theta(S|z) \), parameterized by \( \theta \) and this decoder is defined by an auto-regressive form [61]

\[
p_\theta(S|z) = \prod_{t=1}^{T} p_\theta(S_t|S_{<t}, z) \tag{1}
\]

which depends on previously generated \( S_{t-1}'s \) to generate the next output MFCC feature [60] element \( S_t \). Thus in the proposed model, the decoder is a transformer due to its capability of effectively modeling the temporal relationships and the VAE framework [59] enables the use latent space of \( z \) to capture the distribution of the speech modality which helps generation. The proposed detailed models is shown in the Fig 1. The VAE framework formed has a learnable distribution \( p_\phi(z|F) \) from the frame sequence \( \{F_1, F_2, F_3, \ldots, F_t\} \) consist of lip movement. Training VAE need to maximize the marginal log-likelihood [59] of audio data \( S \) for distribution \( D \), which is given by

\[
\mathbb{E}_{S \sim D}[\log p_{\phi_1, \phi_2, \theta}(S)] \tag{2}
\]

However, optimization of marginal likelihood is computationally infeasible. Consequently, we approximate the true posterior distribution \( \log p_{\phi_1}(z|S) \propto \log p_\theta(S|z)p(z) \) by introducing an \( \phi_1 \)-parameterized encoder which serve as an...
approximate inference model with distribution \( \log q_{\phi_1}(z|S) \). The variational inference is used for training the VAE to yield the evidence lower bound (ELBO):

\[
\mathbb{E}_{S \sim \mathcal{D}} \log p(\phi_1, \phi_2, \theta) \geq \mathbb{E}_{S \sim \mathcal{D}} \left[ \mathbb{E}_{z \sim q_{\phi_1}(z|S)} \log p_{\theta}(S|z) \right] - \mathbb{E}_{S \sim \mathcal{D}} \left[ KL \ q_{\phi_1}(z|S) || p_{\phi_2}(z|F) \right] \\
\geq \mathcal{L}_{REC} - \mathcal{L}_{KL}
\tag{3}
\]

The first term in equation 3 is reconstruction loss and the second term is the KL Divergence between posteriors and the priors.

A. Video representation

In the NLP transformer applied directly on the word tokens [20], while in the proposed model, we have frame sequence \( \{F_1, F_2, F_3, ..., F_t\} \) therefore, the transformer operates on a sequence of embedding vectors obtained from the frame encoder network. The transformer is designed using a stack of \( L \) layers of multi-head attention and a point-wise, fully connected feed-forward network corresponding to both encoder and decoder. The multi-head attention module has multiple self-attention networks working parallel to capture the temporal relationship and contextual clues in any given data sequence. In principle, a transformer maps the input into a query and key-value pairs. The corresponding attention is defined as a mapping between a query and the set of key-value pairs to give an output vector. The output vector is the weighted average of computed values and the corresponding weights is calculated using a compatibility function that operates between the query and the corresponding key, as shown below,

\[
Attention(Q, K, V) = \text{softmax} \left( \frac{QK^T}{\sqrt{d_k}} \right) V
\tag{4}
\]

Where Q represents a matrix consisting of a set of queries while variables K and V represent key and values, respectively, packed into a single matrix for all the inputs. These queries and keys of input have a dimension \( d_k \), while values have a different dimension \( d_v \). Thus, attention is computed by the dot-product of the query and the key vector. However, instead of computing attention only once, the transformer benefits by linearly projecting the queries, keys, and values multiple times in parallel (called Multi-headed attention) and then projects the combined attention to yield final values. For a more exhaustive background description of Transformer models, we request readers to refer to [20], [62]. Multi-head attention helps the model to utilize various representation sub-spaces at different positions to attend to their information jointly.

Thus, the embedding vector sequence is given as input to the transformer. Instead of producing a single vector, the transformer model output is an isotropic Gaussian distribution of the latent space with parameters mean and variance as \( (\mu_i^s, \sigma_i^s) \) for each \( z^th \) time step. The output distribution of the model behaves as prior distribution \( p_{\phi_2}(z|F) \) for learning of posterior distribution jointly. We use these transformer models to learn the interaction between the samples obtained from these Gaussian distributions.

B. Audio representation

First, we preprocess the given video to take out the audio waveform and calculate the MFCC features from these waveforms. The sequence of MFCC features thus obtained, can be represented as \( \{S_1, S_2, S_3, ..., S_t\} \) and given as input to the transformer model instead of the raw audio. The multi-head attention mechanism of the transformer learns the temporal correlation between the sequence of MFCC features. In the given framework, the transformer learns a normal distribution as output for each timestamp, not just a single output vector. The parameters for learned distribution are mean and variance \( (\mu_i^s, \sigma_i^s) \). The learned Gaussian distribution represents the variational posterior of the VAE framework. Since the learnable prior is also Gaussian, the KL divergence is analytically solvable. At training time, a gradient is passed through Gaussian sampling using the re-parameterization trick [59] to get the trained VAE model.

IV. Training

Training a network framework is equivalent to minimizing the loss function in equation 3. Since, maximizing the marginal log-likelihood is intractable, so we maximize its evidence lower bound ELBO [59] which is equivalent to minimizing the reconstruction loss and the KL divergence between posterior latent distribution \( q_{\phi_1}(z|S) \) and the prior distribution \( p_{\phi_2}(z|F) \). Thus the distributions of both modalities come closer to each other for each timestamp, leading to a high correlation between the sequence of lip movement and its corresponding speech in the time domain. However, Learning the synchronization between speech and lip-movement sequence will also need the distributions of varying timestamps to be farther distant to help better speech generation. So, we can exploit deep metric learning to reduce the distance between the embeddings of the same timestamp (in synchronization) and pull apart the embeddings of different timestamps in cross modalities.

Fig. 2. The blue marked face is the anchor and it forms a positive pair with MFCC feature segment coloured in blue while anchor form a negative pair with MFCC feature in red colour.
A. Deep metric learning

The focus of deep metric learning is to learn closer embedding for audio and video corresponding to same timestamp and a distant embedding for different timestamp. The audio-frame pair for same timestamp is termed as positive pair while for different timestamp is termed as negative pair and our objective is to preserve that implicit structure. Concretely, we select a frame $F_i$ at random from the video, called anchor and select two audio feature sample, first is in-sync audio $S_i$ (positive pair) and other is out-of-sync audio sample $S_j$ (negative pair) from the given video i.e., anyone expect $S_i$. Finally, our deep metric learning approach uses a structured loss to pull the distance between the positive pair i.e. $F_i$ and $S_i$, and at the same time pushing the negative set $S_j | j \neq i$ away from the positive pair using the following learning objective function written below,

$$
\mathcal{L}_{\text{met}} = \frac{1}{2N} \sum_{n=1}^{N} \max(0, \mathcal{J}_n)^2 \tag{5}
$$

$$
\mathcal{J}_n = \mathcal{D}(F_i, S_i) + \log(e^{\lambda - \mathcal{D}(F_i, S_i)} + e^{\lambda - \mathcal{D}(S_i, S_j)}) \tag{6}
$$

where $\mathcal{D}(F_i, S_j)$ represents the Euclidean distance between the $i$th frame feature of video $F$ and $j$th MFCC feature of speech $S$. The hyper-parameter $\lambda$ controls the distance margin between positive pair and negative set and $\mathcal{J}_n$ defines the similarity between positive pairs as well as dissimilarities of negative sets with the positive pairs. Therefore minimizing $\mathcal{J}_n$ helps in enlarging distances of out-of-sync feature elements while minimizing the distance of in-sync feature elements.

B. Joint learning

Our model has integrate VAE with transformer and used metric learning for synchronization. Thereof, we jointly learn modality-independent feature representation and synchronised feature embedding by minimizing the joint loss defined in equation 3 and 5. The joint loss can be given as:

$$
\mathcal{L}_{\text{joint}} = \mathcal{L}_{\text{REC}} + \mathcal{L}_{KL} + \lambda \mathcal{L}_{\text{met}} \tag{7}
$$

Where $\lambda$ is the hyper-parameter corresponding to metric learning loss, we choose the best value of $\lambda$ using the validation data.

V. SPEECH PREDICTION

At test time, We need to predict the speech from the given stream of frame sequence $F = f_1, f_2, ..., f_N$ at hand only since we do not have the audio. Thus, at every timestamp $t$, transformer encoder is used to obtain the posterior distribution $q(z|f_t)$. The sample, thus obtained from the posterior distribution, is then passed into the audio decoder, a transformer network that operates on the latent space learned by the posterior distribution. Thus, the decoder transformer uses the modality-independent latent representation to generate the MFCC feature embedding in an auto-regressive manner. We maintain a small overlap across the sliding windows which reduces the boundary effect and the final waveform is obtained using the standard Griffin-Lim algorithm from MFCC features.

VI. EXPERIMENTAL RESULTS

We have evaluated our proposed model for the speech prediction task on the two large scale publicly available datasets, GRID [37] and Lip2Wav [50]. The GRID dataset has a limited vocabulary, where each speaker is speaking only six words chosen from the fixed dictionary. Therefore, to evaluate the performance on a dataset of much larger vocabulary size, we trained the proposed model on the Chemistry dataset introduced in Lip2wav [50]. The dataset consists of nearly 20 hours of video recording of online teaching; it has a vocabulary size of nearly 5000 words, which is much larger than the GRID dataset.

A. Quantitative results

For quantitative comparison, we used various metrics such as Perceptual evaluation of speech quality (PESQ) [63], Extended short-time objective intelligibility (ESTOI) [64] and Standard Short-Time Objective Intelligibility (STOI) [65]. The STOI and ESTOI metrics measure the intelligibility of generated speech, while PESQ measures the speech quality by comparing it with the ground truth speech signal.

We report the values of these metrics for the Grid dataset in Table I. We observe that models that use the MFCC feature instead of the raw audio or LSP features perform comparatively better. We also observe that the proposed approach outperforms or is at par with the recent Lip2wav [50] and [51] models on the given metrics. We have shared the generated samples on the project page link given below; we request the reader to listen to the generated results at their discretion.

We report the results on the lip2wav chemistry lecture dataset in Table II. These results show that the proposed framework outperforms the state-of-the-art Lip2wav approaches in both STOI and ESTOI metrics significantly (0.416 vs. 0.490, 0.284 vs. 0.316) and performs marginally lower in the PESQ metric (1.300 vs. 1.233).

B. Qualitative results

We present the speech waveforms and the spectrogram generated by the Lip2wav model [50] and the proposed model in Figure 3. These waveforms and the spectrogram indicate

\*Project page: https://sites.google.com/view/lip-to-speech/home
that the speech signal generated by the Lip2wav model is overly smoothed. Thus, the generated speech sounds robotic. On the other hand, the waveform and spectrograms generated by the proposed model match closely with the ground truth speech signal.

C. User study

We have done a subjective evaluation by conducting a user study. We asked ten subjects to rate the samples generated by the state-of-the-art models and our proposed framework model for the GRID dataset. Note that, in the GRID dataset, we have a fixed dictionary and each speaker speaks only six words taken from it. For each of the speech prediction models ( [50], [51], and ours), we showed randomly chosen 30 different prediction samples from a pool of 10 subjects. Thus, the samples shown to one subject may differ from those shown to another subject to counter the priming effect.

For each predicted speech sample, we asked the subjects to rate the speech quality based on how realistic (non-robotic) the generated speech sounds, i.e., the similarity of the generated speech to the ground truth for evaluating the speaker speaking style and correct pronunciation of words in sentences. All scores were represented on a scale of 1 to 5, where a higher score means better in for all three criteria and the results of corresponding studies are shown in Table III. We observe the predictions of our model are rated higher in all three criteria, which shows our results are more realistic and accurately understood (transcribed) by the listener.

VII. CONCLUSION

In this work, we look into synthesizing speech based on a sequence of lip movements from the perspective of individual speakers. We proposed a framework to unify the transformer model with VAE to learn a joint latent distribution space for smooth speech generation. We exploit the deep metric learning-based approach to learn the temporal synchronization of lip movement and speech. Our model outperforms state-of-the-art approaches on large-scale benchmark GRID and Lip2wav Chemistry datasets for the unconstrained, large vocabulary of single-speaker, which shows the effectiveness of our approach. We also observe from qualitative results and user studies that our approach produces more realistic speech than prior approaches.
