COMPUTER VISION LIP READING (CV)

S. SUMANTH[1], K. JYOSTHANA[2], J.KARTHIK REDDY[2], G.GEETHA[2]

[1, 2] Department of Information Technology, Velagapudi Ramakrishna Siddhartha Engineering College, Chalasani Nagar, Kanuru, Vijayawada, Andhra Pradesh 520007

Corresponding Author: sumanthsomireddy@gmail.com ,Tel: +91-8125333377

***ABSTRACT*:** In this proposed work, speaking of which, the prosody and content of speech can be taken with the movement of our lips. In this work, we explore the function of combining lip and speech, that is, learning to produce speech that is only given by the movement of the speaker's lips where we focus on an accurate reading of how to make a cup of speech for multiple speakers in optional, large word settings. It captures the identity of the speaker's voice by its facial features, i.e., age, gender, nationality and its shape and lip movement to produce a noticeable expression of the speaker's identity. In this case, we present a novel approach with key design options to achieve the correct composition of the lips in unrestricted situations. We also do various tests and comprehensive tests using quantity, quality metrics and personal testing.

***KEYWORDS*:** Lip Reading in the Wild, Speech synthesis, Multi-speaker speech generation.

I.INTRODUCTION

Lip reading is a way of expertise speech by way of visualizing the motion of the lips. Even though lip-studying is broadly utilized by the deaf and hearing impaired, many human beings have ordinary listening to the potential to look through the mouth[1]. Additionally, expertise in the signs and symptoms of language in lip reading can improve the clarity of verbal exchange in regions with noisy backgrounds. every other feasible manner to read lips is to provide textual content this is constant with the motion of the lips. Although this approach works, human beings additionally explicit feelings through transferring their lost lips in a written illustration. also, they can not constitute a non-linguistic vocal repertoire[2]. In addition, some conditions of use including voice portray, speech recovery in heritage sound, and sound production of folks that are not able to produce said sounds (aphonia) do no longer arise.

While Lip-to-text-to-speech can also clear up positive troubles, intermediate textual content know-how dispels the prosody, pace, and verbal alerts of lips that are critical in fixing many of the issues referred to above. therefore, both Lip-to-textual content, From phrase-of-mouth speech is in a critical situation and calls for direct oral communique. in addition, text-primarily based systems require written statistics sets to be taught, that's difficult to obtain because they require a manual annotation. And the schooling of lip-talking techniques can be completed in a managed manner as a maximum of the video comes paired with the accompanying sound. This type of trouble is a first-rate synthetic intelligence/device gaining knowledge of challenge since it pushes a device's potential to symbolize something instinctive and innate in human beings. moreover, there's an extensive amount of information that can be utilized to train a gadget to achieve this. This undertaking will awareness of a fundamental and de-scoped version of the CV lip-analyzing hassle, with records together with simple face-forward video frames of people saying unmarried English phrases.

We eliminate many of the complexities described above in this manner. Our raw input data will be JPEG images of these video frames, and the model's output will be a classification. One intriguing challenge I attempted to address in this project was the issue of training the massive amount of data and seeing how well a neural network could perform classification with only a few coordinate points representing the general shape of the mouth.

XIII. BACKGROUND

This section provides the necessary background information to understand the project better. It identifies LSTM and Griffin Lim algorithm from the prominent research.

2.1 **Griffin Lim Algorithm**

In recent years, spectrogram phase recovery has received a lot of attention in acoustical applications. On the one hand, recent research has shown that recovering the phase spectrogram can improve the quality of speech signals whose amplitude spectrograms have been improved by some processing. Recent statistical speech synthesis methods, on the other hand, attempt to synthesize an amplitude spectrogram without phase that is accompanied by lip movement to produce speech. The Griffin–Lim algorithm (GLA) has been used as the standard tool for phase recovery in the absence of phase information. GLA is made up of two projections that seek the consistent spectrogram with the given amplitude, where a spectrogram is said to be consistent when its bins retain the neighbourhood relation due to the overlapping window in the time domain[3]. Although GLA has been successfully applied, its slow convergence may necessitate a significant number of iterations before the reconstruction quality is satisfactory.

As a result, some GLA modifications for acceleration have been proposed. The importance of accelerating GLA is not only in terms of computation but also in terms of performance improvement. For example, synthesised sound with the fast GLA (FGLA) is frequently superior to that with GLA. This should be due to the stronger preference for a global minimum, which allows such acceleration to avoid poor local minima. Meanwhile, many applications have demonstrated the ability of the alternating direction method of multipliers (ADMM) to avoid poor local minima. This suggests that by using ADMM, it may be possible to improve the performance of GLA.

2.2 **Long Short Term Memory (LSTM)**

Artificial Neural Networks (ANN) are loosely modelled after biological learning systems and are inspired by them. Biological learning systems are intricate networks of neurons. Neurons are basic units that accept a vector of real-valued inputs and produce a single real-valued output. Feed-forward neural networks are the most common type of standard neural network. There are three layers of neurons in this model: one input layer, one output layer, and at least one intermediate hidden layer. Feed-forward neural networks are only capable of performing static classification tasks. As a result, they can only provide a static mapping between input and output. A dynamic classifier is required to model time prediction tasks[4]. Because of a vanishing gradient problem that arises when working with longer data sequences, standard Recurrent Neural Networks (RNNs) suffer from short-term memory. Fortunately, more advanced RNNs can keep important information from earlier parts of the sequence and carry it forward. Long Short-Term Memory (LSTM) and Gated Recurrent Units are the two most well-known types (GRU).

XIV. EXISTING METHODS

Many solutions to this problem have been proposed, and they can be divided into two categories. MLP is used to extract features in the first phase, and the LSTM layers are similar to the temporal sequence dynamics. The 3D Conv layer is applied in the second stage, followed by the standard conv layers applied with LSTM[5,6]. These methods, however, do not directly affect the sound effects and rely on MFCCs as a feature factor, and are the focus method for oral plants and MFCCs[7,8]. The method is samples of the target speaker's voice that are then embedded in the content to create the content.

XV. DATASET

This segment describes the Oxford-BBC Lip analysis within the Wild (LRW) dataset. The LRW dataset[9] is a huge dataset that is publicly available (upon request) for non-commercial and educational studies. The dataset is made up of 1.16-second video clip segments (approximately 29 frames). It contains hundreds of speaker videos from BBC programmes, normally handiest the part of the speech in which the speaker says the phrase. The dataset consists of metadata that can be used to determine the beginning and end frames for every word length within the video. A big variety of audio systems, mixed with variants in head poses, viewpoints (frontal to profile), and light conditions, make the dataset difficult and necessitates preprocessing to healthy the schooling model. This dataset contains 500 awesome word instances, each with as many as one thousand utterances spoken using an extraordinary speaker. This dataset was created using a multistage pipeline with the purpose of audio-visible speech recognition.

Speech is a big-scale audio-visible dataset that consists of the most effective speech clips without a history noise. The segments range in length from 3 to ten seconds, and every clip features an unmarried speaking man or woman with the simplest seen face in the video and audible sound inside the soundtrack. The dataset consists of approximately 4700 hours of video segments with about 150,000 distinct speakers from an extensive range of people, languages, and face poses.

XVI. PROPOSED METHODOLOGY

The proposed model's block diagram is shown in fig.1. The video frame sequence and a single frame from the sequence are fed into the network. The above method is based on encoder-decoder architecture, where we encode the lip movements along with the face features and decode the Mel spectrogram of the generated speech. To train the speaker encoder on AVSpeech, for every video sequence, a random frame is taken and we crop the face and its corresponding audio window, where the window length is chosen randomly from 1∼to 3 seconds. We train the network with a batch size of 64 and train until the contrastive criterion does not improve. The network is trained for 300 epochs on LRW, with early stopping when validation accuracy does not improve for 10 epochs. We also augment the train data with random horizontal flips for the lip sequence. During inference, the video frame sequence and a single frame from the sequence are fed into the network. Following the generation of the melspectogram, the Griffin-Lim Algorithm is used to reconstruct the raw audio waveform. To reconstruct long sequences, we divided the video into 2-second chunks and generated a melspectogram for each chunk. To obtain the raw audio waveform, we concatenate these chunks and use Griffin-Lim. This method worked well for us because the melspectogram inversion implicitly overlaps across melspecframes to reconstruct the raw audio waveform, preventing abrupt discontinuities between adjacent audio chunks.

XVII. SYSTEM ARCHITECTURE

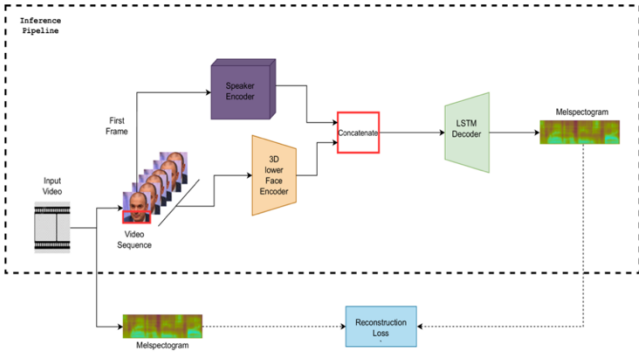

Figure1. Block diagram

XVIII. OVERVIEW

Given an input video of a talking face, our purpose is to synthesise the talking face's speech with the speaker's identification estimated through their face attributes, such as gender, age, and ethnicity. The video consists of a chain of photo frames, F = (F1,F2,F3,.....,FN), while the synthesised speech incorporates melspectogram speech frames, S = (S1,S2,S3,.....,ST). This will be expressed as series-to-sequence gaining knowledge of task[10], wherein the correlations between photo frames and speech frames are discovered. these correspondences capture each speaker's identity and the content material of the speech. After studying the correspondences, speech frames can be car-regressively synthesised for photo frames. It needs to be stated that the variety of photo frames, N, and the number of speech frames, T, do not ought to be identical. especially, each output speech body, St, is represented as a conditional distribution of the previous speech and photograph frames, F.

The first body of an enter frame sequence F = (F1, F2, F3,....., FN) is handed to the speaker encoder to acquire the "speaker embeddings." Similarly, the lip ROIs are cropped and handed to the 3-D decrease face encoder for everybody Fn to obtain the face functions. The speaker embeddings are tiled and concatenated to the facial functions in conjunction with body collection N, which we discuss as "visual capabilities". these visible functions are fed into the LSTM decoder, which generates the speech melspectogram in an auto-regressive manner. As shown in figure 2, we use the melspectogram from the input video throughout education to minimise the imply square blunders of the generated melspectogram.

Figure 2 Video as a sequence of image frames

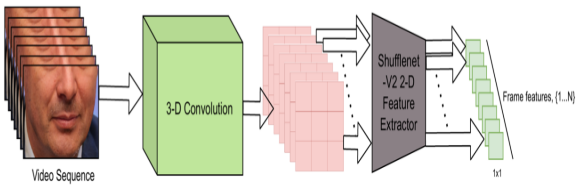

Figure3. The video feature extractor encodes the frame sequence by applying spatiotemporal convolutions.

XIX. VIDEO FEATURE EXTRACTOR

Figure 3 shows how the Video feature extractor is used to extract semantic and temporal lip features from a video frame sequence. We apply a 3-D convolution block to F = (F1, F2, F3,....., FN) to induce temporal awareness of the frame features, and then apply a 2-D feature extractor, ShuffleNet V2[21], to extract the semantic features. We keep the temporal length, N, and reduce the spatial dimensions to one by applying global average pooling to the extracted frame features.

XX. SPEAKER ENCODER

The speaker identity module, also known as the speaker encoder, is used to learn the relationships between a speaker's face and voice. The network takes a text and samples the audio of a speaker to generate the given text in the given voice. We know from [4, 11] that humans can "put a face to a voice," and [12] attempts to generate a face from a given voice sample. In this work, the prime attempt is to generate a voice to face. And we accomplish this by creating encodings that are unique to each speaker. While [13,14,15] use encodings to map speaker identity, they condition them with the speaker's voice. As the face recognition encoder, we employ an Inception-ResNet[16] that has been pre-trained on CASIA-WebFace[17]. We freeze the first three-network blocks and fine-tune the reset, which is two blocks. This network accepts a face and generates a face encoding. Our goal is to learn the correspondence of the face to the speaker's voice, so we use another network, the speech encoder[18]. The speech encoder receives a melspectogram and generates a speaker encoding that corresponds to the identity of the speaker. Moreover, the speech encoder remains frozen throughout the training. With only a few outliners, the network correctly distinguishes between male and female voices, and different age groups have different amounts of energy in their voice. The voices of both genders do not deviate much from the mean voice. It is also noticeable that the voices produced by young students sound like adults. Also, for ethnicities other than white, the generated audio sound distorted rather than producing a proper accent.

XXI. MELSPECTOGRAM DECODER

According to the frame sequence length N, the speaker embeddings are tiled and concatenated to the frame features. These "visual features" are decoded to melspectogram frames (mil spec-frame) in an auto-regressive manner by the melspectogram decoder, as shown in Figure 4. Because the frame sequence length N and the number of time step T do not have to be the same, we first encode the visual features into a latent representation using the hidden and cell state of the 2-layer BiLSTM. To decode, we use a 4-layer LSTM[19] initialised with the BiLSTM's hidden and cell state. The initial melspec-frame input to the decoder can be trained and optimised alongside the network. Furthermore, the regression of each time step is accomplished by sending the previous melspec-frame to a bottleneck network with high dropouts known as "Prenet," which reduces ground-truth bleeding during teacher forcing. As we restrict the information flow from the previous time-step, it also forces the LSTM not to ignore the attended queries. To determine when to stop the regression, we use a linear projection of each time-step'stimestep'stimestep's hidden state concatenated with the latent representation's hidden state.

The sigmoid function, which serves as the gating value, is used to activate the linear projection. During inference, we stop if the gating value exceeds a certain fixed threshold, which is set to 0.5 by default. Because the condensed latent encoding of the frame sequences cannot fully represent the temporal semantic flow, we use the localised attention mechanism to improve the contextual information from the frame sequences. The melspec decoder encodes the frame sequence into a condensed latent representation and decoded the melspecmil spec-frame auto-regressively. An attention mechanism is used to improve the temporal semantic flow.

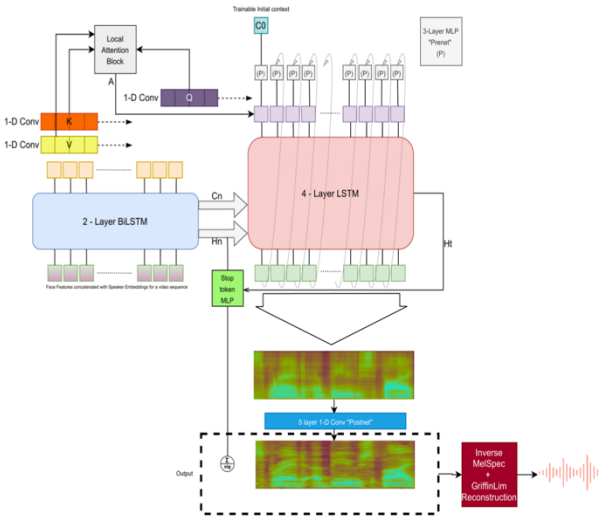

Figure4. Workflow of Melspectogram Decoder

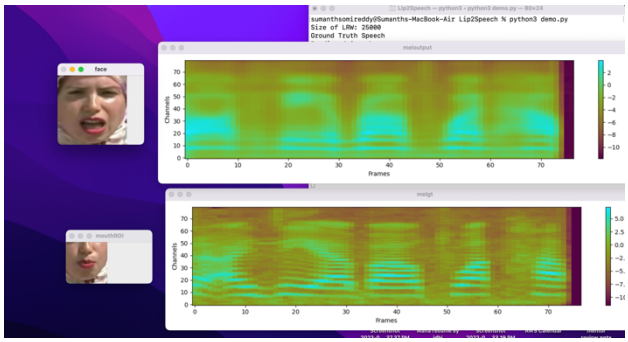

Figure5. Faces and the melspectograms generated with our network.

XXII. IMPLEMENTATION

We evaluate our model using the LRW dataset. We selected 153 videos from the entire dataset with people of various ages, genders, ethnicities, and accents. Figure 5 depicts sample faces and the corresponding generated melspectograms. PyTorch was used to carry out all experiments. With an initial learning rate of 0.001, all models are trained.We created a model prototyping framework in which any change to the model generates appropriate model/validation/training logs automatically. This is heavily used to record our experiments.

```
Size of LRW: 25000
Ground Truth Speech
Predicted Speech
Ground Truth Speech
Predicted Speech
Ground Truth Speech
Predicted Speech
Ground Truth Speech
Predicted Speech
Ground Truth Speech
Predicted Speech
Ground Truth Speech
Predicted Speech
Ground Truth Speech
Predicted Speech
Ground Truth Speech
Predicted Speech
Ground Truth Speech
Predicted Speech
Ground Truth Speech
Predicted Speech
Ground Truth Speech
Predicted Speech
```

Figure6. List of original and predicted speech of each sample

XXIII. RESULTS AND DISCUSSION

The corresponding code has been run on Apple MacBook Air, 1.1 GHz Dual-Core Intel Core i3. From fig 6, at first, the ground speech from the original video is played. Correspondingly, the melspectogram of ground speech is displayed in fig 7. In the model computation, wherein the melspectogram of the original speech is sent as input. The predicted speech is detected by the model. On speech output, the lip movement in the feature is also extracted and visualized. Correspondingly, the predicted speech melspectogram is displayed in fig 8.

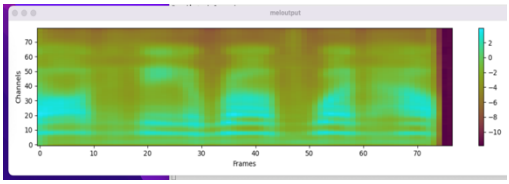

Figure7. Generated Melspectogram of Ground speech

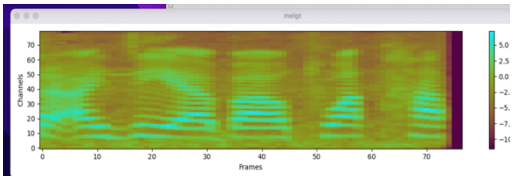

Figure8. Generated Melespectogram of Predicted Speech

XXIV. OBSERVATIONS

Our model has a higher overall quality, the speaker audio embeddings produce voices with a higher correlation. It is clear that the generated audio, regardless of the embeddings, is heavily skewed toward middle-aged white people. With only a few outliners, the network correctly distinguishes between male and female voices, and different age groups have different amounts of energy in their voice. The voices of both genders do not deviate much from the mean voice. It is also noticeable that the voices produced by young students sound like adults. Also, for ethnicities other than white, the generated audio sound distorted rather than producing a proper accent. The generated voice is lacking in variation, based on the samples created for the MOS with our speaker face embeddings.

XXV. CONCLUSION

We proposed a method to synthesize speech from lip movements in this work. Specifically, we focused on generating a different vocal identity for individual speakers based on their age, gender, and ethnicity. As our method uses speaker embedding to incorporate speaker-specific information in the speech synthesis step, which allows the model to be used for all speakers, even ones that were previously unseen during training.

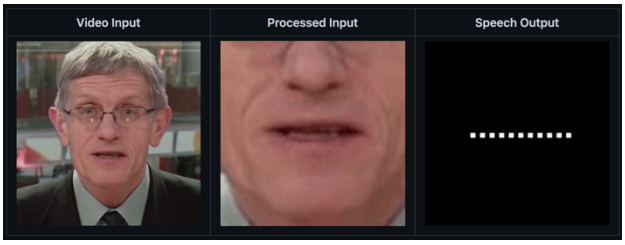

Figure9. The overall output of the work

REFERENCES

[1] Woodhouse, L., Hickson, L., Dodd, B.: Review of visual speech percep- tion by hearing and hearing-impaired people: clinical implications. Inter- national Journal of Language & Communication Disorders 44(3), 253–270 (Jan 2009). https://doi.org/10.1080/13682820802090281

[2] Anikin, A., B ̊a ̊ath, R., Persson, T.: Human non-linguistic vocal repertoire: Call types and their meaning. Journal of Nonverbal Behavior 42(1), 53–80 (Sep 2017). https://doi.org/10.1007/s10919-017-0267-y, https://doi.org/10. 1007/s10919-017-0267-y

[3] Griffin, D., Lim, J.: Signal estimation from modified short-time fourier transform. In: ICASSP '83. IEEE International Conference on Acous- tics, Speech, and Signal Processing. vol. 8, pp. 804-807 (1983). https://doi.org/10.1109/ICASSP.1983.1172092

[4] Kamachi, M., Hill, H., Lander, K., E.: ‘putting the face to the voice’

[5] Petridis, S., Li, Z., Pantic, M.: End-to-end visual speech recognition with lstms. In: 2017 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2017 - Proceedings. pp. 2592–2596. IEEE, United States (Jun 2017). https://doi.org/10.1109/ICASSP.2017.7952625, http://www.ieee-icassp2017.org/, 42nd IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2017, ICASSP ; Confer- ence date: 05-03-2017 Through 09-03-2017

[6] Wand, M., Koutn ́ık, J., Schmidhuber, J.: Lipreading with long short-term memory (2016)

[7] Chung, J.S., Senior, A., Vinyals, O., Zisserman, A.: Lip reading sentences in the wild. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Jul 2017). https://doi.org/10.1109/cvpr.2017.367, http://dx.doi.org/10.1109/CVPR.2017.367

[8] Xia, L., Chen, G., Xu, X., Cui, J., Gao, Y.: Audiovisual speech recognition: A review and forecast. International Journal of Advanced Robotic Systems 17(6), 1729881420976082 (2020). https://doi.org/10.1177/1729881420976082, https://doi.org/10.1177/ 1729881420976082

[9] Chung, J.S., Zisserman, A.: Lip reading in the wild. In: Asian Conference on Computer Vision (2016)

[10] Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks (2014)

[11] Nagrani, A., Albanie, S., Zisserman, A.: Seeing voices and hearing faces:Cross-modal biometric matching (2018)

[12] Oh, T.H., Dekel, T., Kim, C., Mosseri, I., Freeman, W.T., Rubinstein, M.,Matusik, W.: Speech2face: Learning the face behind a voice (2019)

[13] Gibiansky, A., Arik, S.O¨., Diamos, G., Miller, J., Peng, K., Ping, W., Raiman, J., Zhou, Y.: Deep voice 2: Multi-speaker neural text-to-speech. In: NIPS (2017)

[14] Nachmani, E., Polyak, A., Taigman, Y., Wolf, L.: Fitting new speakers
based on a short untranscribed sample. ArXiv abs/1802.06984 (2018)

[15] Arik, S.O¨., Chen, J., Peng, K., Ping, W., Zhou, Y.: Neural voice cloning with a few samples. In: NeurIPS (2018)

[16] Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.: Inception-v4, inception- resnet and the impact of residual connections on learning (2016)

[17] Yi, D., Lei, Z., Liao, S., Li, S.Z.: Learning face representation from scratch (2014)
[18] Jia, Y., Zhang, Y., Weiss, R.J., Wang, Q., Shen, J., Ren, F., Chen, Z.,
Nguyen, P., Pang, R., Moreno, I.L., Wu, Y.: Transfer learning from speaker verification to multispeaker text-to-speech synthesis (2019)

[19] Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation 9(8), 1735–1780 (1997)