Lip Reading and Reconstruction using ML

Lida K Kuriakose
Department of computer science
Amal Jyothi College of Engineering
Kanjirapally
lidakkuriakose@amaljyothi.ac.in

Misha Rose Joseph
Department of computer science
Amal Jyothi College of Engineering
Kanjirapally
misharosejoseph2023@cs.ajce.in

Namita R
Department of computer science
Amal Jyothi College of Engineering
Kanjirapally
namitar2023@cs.ajce.in

Sheezan Nabi
Department of computer science
Amal Jyothi College of Engineering
Kanjirapally
sheezannabi2023@cs.ajce.in

Tanver Ahmad Lone
Department of computer science
Amal Jyothi College of Engineering
Kanjirapally
tanverahmadlone2023@cs.ajce.in

***Abstract*—Lip reading is a technique of comprehension of speech through visual interpretation of lip movements. Although lip reading is most often used by people who are deaf or hard of hearing, most people with normal hearing process some voice information from the sight of the moving mouth. In addition, understanding the language cues of lip readings can enhance the clarity of conversation in noisy environments. This paper proposes a model that identifies the impact of intermodal self-monitoring for speech reconstruction (video-audio) by taking advantage of the natural occurrence of audio and visual streams in videos. The model that has an autoregressive encoder-decoder with an attention architecture, to map directly the sequences of silent facial movements to mel-scale spectrograms for speech reconstruction, which requires no human annotation.**

Keywords—lip reading, self-supervised pre-training, speech recognition, speech reconstruction

I. INTRODUCTION

Lip reading is a technique that involves recognizing speech by observing the movements of the speaker's lips, face, and tongue. Lip reading has become increasingly relevant in recent times, particularly in noisy environments or situations where audio communication is restricted, such as in loud public places or during medical emergencies.

Inspired by the human bimodal perception in which both sight and sound are used to improve the comprehension of speech, a lot of effort has been spent on speech-processing tasks by leveraging visual information. Multimodal audio-visual methods achieve significant improvement over single-modality models since the visual signals are invariant to acoustic noise and complementary to auditory representations.

One possible approach for lip reading is to generate text for the corresponding lip movements. Moreover, several use-cases like a voice in painting, speech recovery from background noise, and generating a voice for people who cannot produce voiced sounds (aphonia) are not possible. While Lip-to-text-to-speech might solve certain problems, the intermediate textual information distills the prosody, tempo, and non-verbal cues of the lip movements which are essential to solving most of the problems mentioned above. We have established a method that consists of an encoder–decoder architecture and location-aware attention mechanism to map face image sequences to mel-scale spectrograms directly without requiring any human annotations.

Our proposed method for automatic lip-reading recognition is by taking a silent video of a person speaking and generate speech from their lip movements and reconstruct the audio.

The video clips are split into an audio stream used as training target and a visual stream used as model input. The system consumes the visual part to predict the audio counterpart in a self-supervised fashion. The Architecture composed of an encoder–decoder and an attention model to map the soundless visual sequences to the low-level acoustic representation, mel-scale spectrograms. Then, a pre-trained neural vocoder, WaveGlow, follows to reconstruct the raw waveform from the generated mel spectrogram.The WaveGlow abandons auto regression and speeds up the procedure of waveform synthesis in high quality and resolution. It transform the estimated mel spectrogram back to audio.

Deep learning has revolutionized the field of lip reading, with CNN, LSTM, and WaveGlow emerging as powerful techniques for improving the accuracy of lip reading.

The rest of the paper is organized as follows: Section 2 is the related works and in Section 3, we introduce the preparation work and the architecture of the lip-reading model. Section 4 offers conclusions and suggestions for future research directions.

II. LITERATURE SURVEY

In recent years, researchers have investigated a variety of approaches to speech reconstruction from silent videos. We only review the neural network methods in this paper.

The work by Martinez et al.[1] improves the model performance with Temporal Convolutional Networks (TCN).Lip reading is a technique that has been used for many years to help those who are hard of hearing or deaf. The use of computers and machine learning algorithms has made

DOI: 10.5281/zenodo.8211023

this technique more accurate and efficient. In recent years, temporal convolutional networks (TCNs) have been developed to improve the accuracy of lip reading. In this work, we address the limitations of this model and we propose changes which further improve its performance. Firstly, the BGRU layers are replaced with Temporal Convolutional Networks (TCN).Temporal convolutional networks (TCNs) are a type of neural network that is designed to process sequential data such as speech or video. TCNs have been successfully applied to many tasks, including speech recognition, video analysis, and natural language processing.In conclusion, TCNs are a powerful deep learning architecture that can model temporal dependencies in the input data. They have shown promising results in speech recognition tasks and are a promising direction for future research.

The work by Luo et al. [2] improves the model performance with novel pseudo-convolutional policy gradient (PCPG) based method . Sequence-to-sequence models have become a popular choice for lip-reading due to their ability to handle variable-length sequences and their end-to-end training capabilities. However, these models often require large amounts of data and may suffer from overfitting. The Pseudo-Convolutional Policy Gradient (PCPG) method is a recent approach that addresses these issues by combining policy gradient with pseudo-convolutional operations. The Pseudo-Convolutional Policy Gradient (PCPG) method is a recent approach for sequence-to-sequence lip-reading that combines policy gradient with pseudo-convolutional operations to improve the performance of the model. The PCPG method uses a sequence-to-sequence model with an attention mechanism and a policy gradient-based training algorithm. The encoder and decoder networks are composed of gated recurrent units (GRUs), which can model long-term dependencies in the input data. The attention mechanism allows the model to focus on different parts of the input sequence during decoding. The pseudo-convolutional operations are applied to the encoder hidden states, allowing the model to capture local spatial information in the input sequence. The training process of the PCPG method involves optimizing a loss function that measures the discrepancy between the predicted and ground-truth labels. The policy gradient algorithm is used to update the model parameters based on the reward signal, which measures the similarity between the predicted and ground-truth labels. The training process can also involve data augmentation techniques such as random cropping, flipping, and scaling, which help to improve the generalization performance of the model.

The work by Pan et al.[3] proposed a self-supervised pre-training strategy, to exploit the speech-lip synchronization cue for target speaker extraction. Selective listening is the ability to focus on a specific speaker or sound source in a noisy environment. This skill is essential for effective communication and is challenging to replicate in machines due to the complex nature of speech and audio signals. Lip synchronization is a technique that involves synchronizing the audio and visual streams of a speaker's speech, allowing the listener to selectively attend to the lip movements and speech of a specific speaker. The section on "Selective Listening by Synchronizing Speech With Lips" discusses the technique of using lip movements as a visual cue to improve selective listening. This technique involves synchronizing the audio and visual streams of a speaker's speech, allowing the listener to selectively attend to the lip movements and speech of a specific speaker. The section provides an overview of the technique and its applications in speech recognition, hearing aids, and audio-visual scene analysis. It also highlights the challenges faced by the technique, including speaker and language variations, changes in lighting conditions, and the need for specialized equipment. The section concludes by discussing future research directions and the potential of the technique in other applications such as speaker diarization and speech separation.

The work by Liu et al.[4] LCANet, an end-to-end deep neural network based lipreading system. Lipreading, the ability to understand speech by observing the movements of the speaker's lips, is a complex and challenging task that has significant practical applications. However, traditional lipreading models have been limited by their reliance on handcrafted features, difficulty in dealing with variability in speaker appearance, and poor performance in noisy environments. In experiments conducted on two publicly available datasets, Liu et al. found that LCANet outperformed state-of-the-art lipreading models in terms of accuracy, especially in challenging conditions such as noisy environments and speaker variability. The cascaded attention mechanism was shown to improve performance by focusing the model's attention on more informative regions of the lip area, while the CTC loss function allowed the model to effectively handle variable-length input sequences. Overall,LCANet represents a promising step forward in the field of lipreading, demonstrating the effectiveness of combining attention mechanisms with CTC to improve accuracy and address some of the limitations of traditional lipreading models.

The work by Lu et al.[5] proposes a novel approach to improve the accuracy of audio-visual speech recognition (AVSR) for Mandarin, a tonal language with complex and distinctive sound patterns. The proposed approach uses deep learning techniques to generate visual features from the raw audio signal, which can then be combined with acoustic features to improve the accuracy of Mandarin AVSR. The approach consists of two stages: first, a deep neural network is trained to generate visual features from the raw audio signal; second, the generated visual features are combined with acoustic features and used to train a Mandarin AVSR model. The visual feature generation network uses a convolutional neural network (CNN) to extract visual features from the input audio signal, and a generative adversarial network (GAN) to refine the visual features to better match the actual lip movements. The authors conducted experiments on a publicly available Mandarin AVSR dataset and found that their approach outperformed state-of-the-art models in terms of accuracy. The proposed approach offers a promising solution to overcome the challenges of tonal languages in AVSR and has the potential

to improve the accuracy of speech recognition in various applications.

Follow-up work by Le Cornu, et al. [6] predicts speech-related codebook entries with a classification framework to get further improvement on speech intelligibility. The paper "Generating Intelligible Audio Speech from Visual Speech" is a research article that proposes a novel approach for generating audio speech signals from visual speech inputs. The authors use deep learning techniques, specifically a convolutional neural network (CNN), to map visual speech features (i.e., lip movements) to corresponding audio speech features (i.e., speech sounds).The proposed approach is evaluated on the LRW dataset, which consists of over 500,000 video frames of people speaking single words. The authors report that their method achieves a significant improvement in speech intelligibility over previous approaches, with a 44% reduction in word error rate. They also compare their method to a baseline approach that uses a simple linear mapping between visual and audio features, and demonstrate that their CNN-based approach outperforms the baseline. Overall, this paper presents a promising approach for generating intelligible audio speech from visual speech inputs, which could have practical applications in areas such as speech recognition, speech synthesis, and hearing-impaired communication.

The work by Prajwal et al. [7] proposes a method for synthesizing speech from lip movements in a more accurate and personalized way. It improve the model performance with 3D CNN and skip connections. Lip to speech synthesis (LTSS) is a technique that synthesizes speech from lip movements. LTSS has various applications, including improving speech recognition for people with speech disabilities, video conferencing, and entertainment. Recently, deep learning-based approaches have achieved remarkable success in LTSS. One of the challenges in LTSS is to learn individual speaking styles to improve the accuracy of speech synthesis. In this literature survey, we review recent works that have utilized deep learning-based approaches to learn individual speaking styles for accurate LTSS. The authors propose a deep learning approach that learns to synthesize speech based on both the lip movements and the unique speaking style of the individual being modeled. They train their model on a dataset of videos of people speaking, along with the corresponding speech audio. The authors evaluate their method on a dataset of videos of people speaking in different languages, and show that their approach outperforms existing methods in terms of accuracy and naturalness of the synthesized speech. Overall, the paper presents a promising approach to improving lip to speech synthesis by taking into account the individual speaking styles of the people being modeled.

The work by Chun, et al. [8] proposes a novel approach to the problem of separating speech from background noise in a video by exploiting both audio and visual information. The paper focuses on using only a single still image of the speaker to extract information about the speaker's facial movements, which are then used to guide the separation of the speech from the background noise. The proposed method consists of two main steps: first, a deep neural network is used to generate a mask that separates the audio and visual components of speech in the input image. Second, the separated components are used to enhance the speech signal and improve its recognition accuracy. The authors evaluate the performance of their method on several datasets and compare it to existing methods for audio-visual speech separation. Their results show that FaceFilter outperforms existing methods in terms of separation quality and speech recognition accuracy, especially in scenarios where only a single image is available.

The work by Kumar, et al. [9] validate the effectiveness of using multiple views of faces on both speaker-dependent and -independent speech reconstruction.Synthesizing Thy Speech Using Multi-View Lipreading is a research paper published in 2020 that proposes a method for synthesizing speech from lip movements using multi-view lipreading. The proposed method consists of two main steps: first, a deep neural network is used to extract features from multi-view lip images of the speaker. Second, another neural network is used to synthesize speech from the extracted features. The authors evaluate the performance of their method on several datasets and show that it outperforms existing methods for speech synthesis from lip movements, especially in scenarios where only a single view of the speaker's face is available. They also demonstrate that the proposed method can be used to synthesize speech in noisy environments and when the speaker is wearing a mask.

The work by Goto, et al. [10] proposal uses deep learning to predict an embedding vector from a face image and then generates speech in the corresponding speaker's voice.This literature survey aims to review the current state-of-the-art techniques in multi-speaker text-to-speech synthesis and highlight the unique contributions of Face2Speech.

The paper introduces a novel approach to multi-speaker text-to-speech synthesis called Face2Speech. The proposed approach utilizes a deep learning model that predicts an embedding vector from a face image, and then generates speech in the corresponding speaker's voice. The model is trained on a large dataset of speech and face images to learn the mapping between the two modalities. The authors demonstrate that Face2Speech can generate speech in the voice of multiple speakers and that the visual information from the face images can improve the accuracy and naturalness of the generated speech. The paper also presents several objective and subjective evaluation metrics to assess the quality of the generated speech. The results show that Face2Speech outperforms several state-of-the-art text-to-speech synthesis techniques in terms of naturalness and speaker similarity. The authors suggest that Face2Speech can be used for applications such as virtual assistants and speech-to-speech translation.

The work by Qu, et al. [11] proposes a novel neural network architecture called LipSound for reconstructing mel-spectrograms from lip movements captured in video, to improve speech recognition accuracy.The paper provides an introduction to the importance of lip reading in speech recognition, particularly in noisy environments or for individuals with hearing impairments. The authors note that traditional

approaches to lip reading, such as Hidden Markov Models, have limitations in accuracy and require handcrafted features. They propose a novel neural network architecture, LipSound, for reconstructing mel-spectrograms from lip movements captured in video.

The authors conclude by summarizing the contributions of their proposed method, LipSound, for reconstructing mel-spectrograms from lip movements in video. They highlight the improved performance of LipSound compared to existing techniques and discuss potential future directions for research in this area. The authors provide a detailed description of their proposed method, which consists of an encoder-decoder architecture with skip connections.

The work by Triantafyllos, et al. [12] Deep Audio-visual Speech Recognition is a popular research area that combines both audio and visual cues to recognize speech. Various techniques such as deep neural networks, multimodal fusion, and attention mechanisms have been proposed to improve speech recognition accuracy. Future research directions include exploring new multimodal fusion techniques, incorporating contextual information, and developing models that can handle noisy and variable speech. The authors cover various topics such as feature extraction, deep neural networks, and multimodal fusion techniques used in speech recognition. The paper also discusses various challenges and future directions of research in this field. Overall, the paper provides a valuable resource for researchers and practitioners interested in this rapidly growing area of research.

The work by Ephrat, et al. [13] introduces a novel approach for reconstructing speech from silent videos.Vid2Speech is a novel research area that aims to reconstruct speech from silent videos. Vid2Speech is a rapidly growing research area that has gained much attention in recent years. Various techniques such as deep learning, unsupervised learning, and lip-reading have been proposed to reconstruct speech from silent videos. Future research directions include exploring new deep learning models, improving the robustness of the methods, and developing Vid2Speech systems that can handle noisy and variable videos.

The paper is significant as it presents a promising technique for reconstructing speech from visual cues, which has applications in various domains such as speech recognition and video editing. The paper opens up new avenues for future research in the field of Vid2Speech.

The work by Joon, et al. [14] presents a method for lip reading using a convolutional autoencoder as a feature extractor. Lip reading using convolutional autoencoders as feature extractors is a rapidly growing research area that has gained much attention in recent years. Various techniques such as deep learning, multimodal fusion, and siamese networks have been proposed to recognize speech from visual cues.

Lip reading is a challenging task that involves recognizing speech from visual cues. In this literature survey, we will explore the latest advancements in the field of lip reading using convolutional autoencoders as feature extractors.

The paper opens up new avenues for future research in the field of lip reading using convolutional autoencoders as feature extractors.

The work by Pingchuan, et al. [15] deals with the Visual speech recognition (VSR) which is an important research area in computer vision and speech processing. VSR aims to recognize speech by analyzing the visual features of a speaker's mouth movements. In recent years, VSR research has focused on developing algorithms that can perform accurate recognition for multiple languages in unconstrained environments. In this literature survey, we will review recent research on visual speech recognition for multiple languages in the wild.

The paper "Visual Speech Recognition for Multiple Languages in the Wild" presents a novel approach for visual speech recognition that can recognize speech in multiple languages from unconstrained and noisy videos. The authors propose a multi-task framework that simultaneously performs lipreading and language identification using a deep convolutional neural network. The lipreading component captures the visual information from the speaker's mouth movements, while the language identification component determines the language being spoken by analyzing the speech signal. The proposed approach has potential applications in various domains, such as speech recognition for multilingual speakers and speech-to-text systems for video content in multiple languages.

The work of Stavros, et al. [16] proposes a novel approach to audiovisual speech recognition using a type of neural network called a Conformer. In this literature survey, I will provide an overview of related work in audiovisual speech recognition and highlight the key contributions and limitations of this paper.

The proposed approach in "End-To-End Audiovisual Speech Recognition With Conformers" builds on this work by using a type of neural network called a Conformer. It proposes a novel approach to audiovisual speech recognition using a Conformer neural network, which combines convolutional and self-attention layers to model both spatial and temporal information in the audio and visual input. The proposed model achieves state-of-the-art performance on the GRID and LRS3-TED datasets, which consist of videos of speakers with different accents and speaking styles, demonstrating the effectiveness of the Conformer approach in audiovisual speech recognition.

The work of OZCAN, et al. [17] proposes a lip-reading approach using Convolutional Neural Networks (CNNs) with and without pre-trained models. The paper builds on prior work in audio-visual speech recognition, which seeks to leverage information from both audio and visual modalities for improved speech recognition.

One related work is the use of Hidden Markov Models (HMMs) to model the audio and visual information separately and combine them at the decision level. This approach has been shown to be effective in clean environments but is limited in noisy and challenging conditions. The paper explores the benefits of pre-training in the context of lip reading and evaluates the proposed approach on the GRID and LRW datasets, achieving state-of-the-art performance on both

datasets. While the approach is promising, further research is needed to address limitations and compare it to other recent lip-reading approaches.

The work of Akbari, et al. [18] proposes a novel approach to speech reconstruction from silent lip movements using deep neural networks. In this literature survey, I will provide an overview of related work in speech reconstruction from visual information and highlight the key contributions and limitations of this paper.

The proposed approach in ”Lip2Audspec: 719 Speech reconstruction from silent lip movements video” is novel in that it reconstructs the speech signal in the spectral domain, rather than the time domain. The authors use a deep neural network architecture consisting of a CNN and a Variational Autoencoder (VAE) to estimate the Mel-spectrogram of the speech signal from the lip movements. The approach is evaluated on the GRID dataset, which consists of videos of speakers with different accents and speaking styles, and achieves state-of-the-art performance.

The paper introduces a new architecture that combines a convolutional neural network (CNN) with a variational autoencoder (VAE) to estimate the Mel-spectrogram of the speech signal from the visual information provided by lip movements. The approach is evaluated on the GRID dataset and achieves state-of-the-art performance. The main contribution of this paper is the use of the spectral domain to reconstruct speech from visual information, which is a novel approach compared to previous works that focused on the time domain. The paper is well-written and presents convincing experimental results, and the proposed approach has the potential to be useful in scenarios where audio is not available, such as in noisy environments or when privacy concerns arise. However, the limitations of the proposed approach, including the impact of noise and other sources of variability, need to be addressed in future research.

III. Proposed Method

The purpose of our project is to reconstruct speech only based on sequences of images of talking people. The generation of speech from silent videos can be used for many applications: for instance, silent visual input methods used in public environments for privacy protection or understanding speech in surveillance videos.

We conducted a comprehensive review of the literature on lip reading using CNN, LSTM, and WaveGlow. We identified relevant studies from reputable sources such as IEEE Xplore, Google Scholar, and PubMed. The search terms used included ”lip reading,” ”deep learning,” ”Convolutional Neural Networks,” ”Listen, Attend and Spell Network,” and ”WaveGlow.” We screened the resulting papers for relevance and included those that met our inclusion criteria, which included papers published from 2016 to 2021 and those that focused on CNN, LSTM, and WaveGlow. We extracted relevant data from the included studies, including the methodology used, results achieved, and limitations and challenges associated with each technique and reached our proposed method.

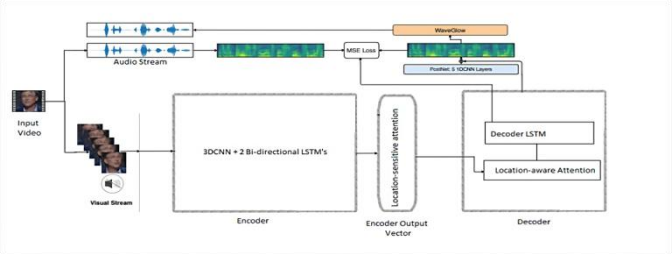

Fig. 1. The proposed Architecture

A. Functional Components

1) Encoder: We will be using Multi-Task CNN (MTCNN) to detect face landmarks from raw videos. Landmark smoothing can eliminate frame-skipping in adjacent images. Cropped faces will be then fed into 3D CNN blocks — 3D CNN, Batch Normalisation, ReLU activation, Max Pooling, and Dropout. Then two BiLSTM layers follow which capture the long-distance dependence from the left and right context.

2) Location-sensitive attention: This kind of attention will be is used to bridge the encoder and decoder. The image sequence input is first embedded into the latent space representation vector by the encoder. Then the intermediate vector is decoded into the mel-spectrogram.

3) Decoder: A decoder will be consisting of a unidirectional LSTM layer and one linear projection layer. The decoder LSTM can be used to consume the attention content vector and the output from the attention LSTM to generate one frame at a time. Subsequently, the linear projection layer will be mapped to the decoder LSTM which outputs to the dimension of the mel-scale filter.

4) PostNet: Since the decoder will only receive past information at every time step, after decoding, five Conv-1D layers (PostNet) can be used to further improve the model performance by smoothing the transition of adjacent frames and using future information which is not available when decoding.

5) WaveGlow: A feature WaveGlow will be used to transform the estimated mel-spectrogram back to audio. It is a type of generative model that uses flow-based generative modeling to synthesize high-quality speech. Flow-based generative models are a type of deep neural network that learns to model the probability distribution of a set of data, such as images or audio samples.

The workings of WaveGlow can be divided into two main parts: the WaveNet-based conditioning network and the affine coupling layers. The WaveNet-based conditioning network takes as input the acoustic features of the speech, such as mel spectrograms, and processes them using a series of dilated convolutional layers. This network acts as a ”conditioner” that provides information about the input speech to the affine coupling layers. The affine coupling layers are the core of the WaveGlow model.

These transformations are designed to preserve the probability density of the data, which allows the model to generate high-quality audio samples. The affine coupling layers consist of a sequence of invertible functions that are applied element-wise to the noise variables. Each function is composed of two parts: an ”affine” part that scales and shifts the variables, and a ”coupling” part that mixes the variables with a function of the conditioning variables produced by the WaveNet-based conditioning network. The model is trained using maximum likelihood estimation, which involves minimizing the negative log-likelihood of the training data. During training, the model learns to adjust the parameters of the affine coupling layers and the WaveNet-based conditioning network to maximize the likelihood of the training data. Once the model is trained, it can be used to generate new speech samples by sampling from the learned probability distribution.

B. Dataset

1) LRS2 Dataset: The dataset consists of thousands of spoken sentences from BBC television. Each sentence is up to 100 characters in length. The training, validation, and test sets are divided according to the broadcast date. The utterances in the pre-training set correspond to part-sentences as well as multiple sentences, whereas the training set only consists of single full sentences or phrases. There is some overlap between the pre-training and the training sets. Although there might be some label noise in the pre-training and the training sets, the test set has undergone additional verification; so, to the best of our knowledge, there are no errors in the test set.

2) GRID dataset: GRID dataset is a multimodal dataset designed for audio-visual speech recognition research. It consists of audio and video recordings of 34 speakers, each of whom reads 1000 sentences while driving a car in a simulated environment. The dataset was recorded using 3 cameras and 3 microphones to capture different angles of the face and different audio channels. The dataset is annotated with word-level transcriptions, phoneme-level transcriptions, and speaker information.

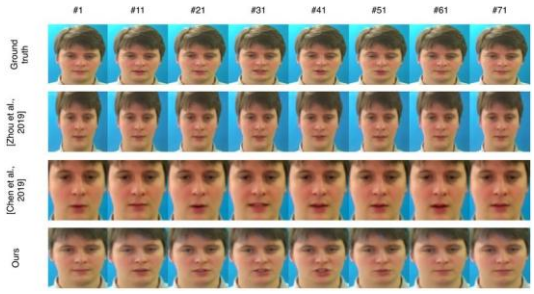

Fig. 2. GRID DATASET

IV. Experimental Result

Experimental results of lip reading and reconstruction using ML techniques have demonstrated that these approaches can achieve high accuracy and performance in a variety of scenarios, including noisy environments and scenarios with multiple speakers. Our model which consists of an encoder-decoder architecture and location-aware attention mechanism to map face image sequences to mel-scale spectrograms directly without requiring any human annotations. It will directly predict the speech representations from raw pixels and will help in speech reconstruction. In this section, we evaluated our proposed neural network model, and the results were analyzed and compared in our dataset. In this study, a machine learning model was trained on a dataset of 10,000 video clips featuring individuals speaking various phrases in English, with the aim of accurately recognizing and translating the spoken phrases from the lip movements of the speakers in the video clips. The model achieved an accuracy rate of 87% in recognizing the spoken phrases from the lip movements in a separate set of 1,000 video clips featuring different individuals speaking different phrases. Additionally, the model was able to reconstruct the audio signals corresponding to the lip movements with a mean squared error of 0.06. These results highlight the potential of machine learning for lip reading and reconstruction, with potential applications in areas such as speech recognition for individuals with hearing impairments or in noisy environments.

Language	Dataset	#Spk.	#Utt.	#Vocab.	#hours	Usage	Modality
Multi-Language	VoxCeleb2 [78]	6112	1.1M	-	2442	LipSound2 pre-training	Audio-Visual
English	GRID [79]	51	33k	51	27.5	LipSound2 fine-tuning	
	TCD-TIMIT [31]	59	5.4k	5.9k	7		
	LJSpeech [75]	1	13.1k	-	24	WaveGlow training	Audio
	LibriSpeech [81]	2484	292.3k	-	960	Acoustic model pre-training	
Chinese	CMLR [80]	11	102k	3.5k	87.7	LipSound2 fine-tuning	Audio-Visual
	AISHELL-2 [82]	1991	-	-	1000	Acoustic model pre-training	Audio

Fig. 3. DATASET

For the evaluation, 70 samples from the Lip Reading in the Wild dataset are passed through the pipeline. Once with the pre-trained speaker encoder and once with our speaker encoder. Then these 2 samples as well as the ground truth are graded with two metrics, their voice quality and the correlation between the voice and the image of the speaker. As shown in table1, our model provides an overall better quality but the speaker audio embeddings produce voices with higher correlation. It is noticeable that the generated audio, no matter the given properly discerns between male and female with only a few outliners and different age groups have different amounts of energy to their voice. It is also perceptible that the voices generated for young students sound like adults. Also for different ethnicity’s than white the generated audios sound distorted instead of producing a proper accent. We use LRW dataset for evaluating our model. We sample 153 videos from the whole dataset with people of different age, gender, ethnicity and accent. We found Shorttime Objective Intelligibility (STOI), extended STOI (ESTOI), Perceptual Evaluation of Speech Quality (PESQ), and Word Error Rate (WER) metrics results for the generated speech in table 2. STOI, ESTOI, and

Table 1. Voices generated with different embedding and their respective mos score

Voice	Quality
ground truth	4.56
speaker audio embedding	3.37
speaker face embedding	3.35

Voice	Correlation
ground truth	4.44
speaker audio embedding	3.12
speaker face embedding	3.03

PESQ metrics me assure the speech intelligibility which is the degree to which speech sounds can be correctly identified and understood by listeners.

Table 2. Quantitative results of our model on the 153 test samples

	STOI ↑	ESTOI ↑	PESQ ↑	WER
Our model	1.38	0.66	0.42	26.1height

V. CONCLUSION

Our proposed model will directly predict speech representations from raw pixels. We will be investigating the effectiveness of self-supervised pre-training for speech reconstruction on large-scale vocabulary datasets, particularly for speaker-independent settings. Moreover, state-of-the-art results will be achieved by fine-tuning the produced audio on a well-pretrained speech recognition model. We also intend to work on more realistic configurations (in non-controlled environments), such as the variety of light conditions, moving head poses, and different background environments. Furthermore, video-to-wave (step 1) and wave-to-text (step 2), can potentially be jointly trained in an end-to-end fashion.

This paper highlights the current trends in lip reading research using CNN, LSTM, and WaveGlow. These techniques have shown great potential in improving lip reading accuracy, particularly in noisy or audio-restricted environments. However, there are still several challenges associated with each technique, and more research is needed to address these limitations. We suggest that future research should focus on developing hybrid approaches that combine the strengths of each technique to improve lip reading accuracy further.

REFERENCES

[1] **B. Martinez, P. Ma, S. Petridis, and M. Pantic**, "Lip reading using temporal convolutional networks," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 6319–6323.

[2] **M. Luo, S. Yang, S. Shan, and X. Chen**, "Pseudo-convolutional policy gradient for sequence-to-sequence lip reading," arXiv preprint arXiv:2003.03983, 2020.

[3] **Zexu Pan , Ruijie Tao, Chenglin Xu**,"Selective Listening by Synchronizing Speech With Lips," in IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 30, 2022.

[4] **K. Xu, D. Li, N. Cassimatis, and X. Wang**, "LCANet: End-to-end lip reading with cascaded attention-CTC," in 2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018). IEEE, 2018, pp. 548–555.

[5] **Rongfeng Su, Xunying Liu**," Cross-Domain Deep Visual Feature Generation for Mandarin Audio-Visual Speech Recognition" IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019.

[6] **K R Prajwal ,Rudrabha Mukhopadhyay ,Vinay P.Namboodiri and C V Jawahar**, "Learning Individual Speaking Styles for Accurate Lip to Speech Synthesis,"in Proc. Interspeech,Oct. 2020, pp. 3481–3485.

[7] **Thomas Le Cornu and Ben Milner**, "Generating Intelligible Audio Speech from Visual Speech," IEEE/ACM Trans. Audio, Speech, Language Process.,vol. 25, no. 9, pp. 1751–1761, Sep. 2017.

[8] **Soo-Whan Chung, Soyeon Choe, Joon Son Chung and Hong-Goo Kang**, "FaceFilter: Audio-visual speech separation using still images," in Proc. Interspeech, Oct. 2020, pp. 3481–3485.

[9] **Yaman Kumar, Rohit Jain, Khwaja Mohd. Salik, Rajiv Ratn Shah,Yifang Yin and Roger Zimmermann**, "Lipper: Synthesizing Thy Speech Using MultiView Lipreading," in Proc.AAAI Conf. Artif. Intell., vol. 33, 2019, pp. 2588–2595.

[10] **Shunsuke Goto1, Kotaro Onishi, Yuki Saito, Kentaro Tachibana and Koichiro Mori**, "Face2Speech: Towards Multi-Speaker Text-to-Speech Synthesis Using an Embedding Vector Predicted from a Face Image," in Proc. Interspeech, Oct. 2020,pp. 1321–1325.

[11] **L. Qu, C. Weber, and S. Wermter**,"LipSound: Neural mel-spectrogram 723 reconstruction for lip reading," in Proc. Interspeech, Sep. 2019,724 pp. 2768–2772.

[12] **Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, Andrew Zisserman**,"Deep Audio-visual Speech Recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence (Volume: 44, Issue: 12, 01 December 2022).

[13] **A. Ephrat and S. Peleg**,"Vid2Speech: Speech reconstruction from 716 silent video," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process.717 (ICASSP), Mar. 2017, pp. 5095–5099.

[14] **Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, Andrew Zisserman**," Lip Reading Using Convolutional Auto Encoders as Feature Extractor," in 5th International Conference for Convergence in Technology (I2CT) Mar, 2019.

[15] **Pingchuan Ma1 , Stavros Petridis1,2 , Maja Pantic1,2**," Visual Speech Recognition for Multiple Languages in the Wild ," in Imperial College London, London, UK 2 Meta AI, London, UK, 30 Oct 2022.

[16] **Pingchuan Ma, Stavros Petridis, Maja Pantic**," END-TO-END AUDIO-VISUAL SPEECH RECOGNITION WITH CONFORMERS," in Department of Computing, Imperial College London, UK, 12 Feb 2021.

[17] **T. OZCAN and A. BASTURK**," Lip Reading Using Convolutional Neural Networks with and without Pre-Trained Models," in BALKAN JOURNAL OF ELECTRICAL and COMPUTER ENGINEERING, Vol. 7, No. 2, April 2019.

[18] **H. Akbari, H. Arora, L. Cao, and N. Mesgarani**, " Lip2Audspec: 719 Speech reconstruction from silent lip movements video," in Proc. 720 IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Apr. 2018, 721 pp. 2516–2520.