International Journal of Advanced Technology and Engineering Exploration, Vol 11(121)
ISSN (Print): 2394-5443 ISSN (Online): 2394-7454
http://dx.doi.org/10.19101/IJATEE.2023.10102613

Research Article

Lip2Voice: a sequence-to-sequence visual speech recognition system for predicting speech from silent video inputs

Aathira Pillai, Bhavana Mache and Supriya Kelkar*
Department of Computer Engineering, Maharshi Karve Stree Shikshan Samstha's Cummins College of Engineering for Women, Karvenagar, Pune, Maharashtra, India

Received: 28-November-2023; Revised: 19-December-2024; Accepted: 22-December-2024

Abstract
Lip reading refers to the understanding of speech without relying on auditory input and is beneficial for individuals with speech impairments, as it enables them to participate in social activities. In this work, a visual speech recognition (VSR) system was developed using a sequence-to-sequence (seq2seq) learning paradigm. The proposed model used a spatio-temporal encoder to capture the sequence of lip movements, which was then complemented by a decoder to generate speech of superior quality. The predicted mel spectrogram was reconstructed utilizing the Griffin-Lim algorithm. In addition, incorporating an inference module has enabled the creation of fixed-length speech from input videos of varying lengths. A different training method termed "alternative training" was adopted to instruct the model to prioritize the sentences themselves over speaker-specific qualities, hence leading to a quicker convergence. The model achieved a training loss of 36.6% on the dual speaker dataset and reduced the word error rate (WER) by 10% compared to the Vid2Speech model. A comprehensive human subjective evaluation was conducted on five audio sets, assessing two metrics—audibility and mispronunciation. The results showed that Lip2Voice had a lower overall percentage error than the Vid2Speech model. A comparative analysis between the proposed and existing models, focusing on audio spectrograms and frequency domain waveforms evaluated on power spectral density (PSD), demonstrated the similarity between the spectrograms generated by Lip2Voice model and the original audio. This research indicates that computer-based lip-reading systems for people with speech impairments are attainable.

Keywords
Lip reading, Visual speech recognition, Spatio-temporal encoder, Alternative training, Speech impairments, Word error rate.

1.Introduction
Humans have always paid high attention to lip movements for visually listening to speech in a noisy environment [1]. Furthermore, speech can be interpreted without auditory input. This may be required when audio recognition is complex, in cases of speech recovery during online video sessions or from corrupted videos, in military applications, and closed-circuit television (CCTV) surveillance. Advancements in computer vision have led to research efforts aimed at automating the process of lipreading, known as the visual speech recognition (VSR) system [2]. VSR primarily involves the analysis of visual aspects of speech. VSR systems aim to equip machines to comprehend speech even in noisy environments. Speech recognition systems can be modelled using audio-visual data, audio samples, and silent videos.

Audio-visual speech recognition (AVSR) systems include audio samples and video data to achieve better performance [3, 4]. Many proposed state-of-the-art models use attention mechanisms [1], teacher-student learning, and curriculum learning [3]. These improve the results while adding the overhead of increased complexity to the model architecture.

Lip reading is considered a video-to-text [5–7] and video-to-speech [8] task. However, developing a VSR model presents several intricate challenges. One significant hurdle lies in achieving accurate speech recognition where audio is absent, as prevalent audio-visual models face challenges to generalize effectively, resulting in speaker-dependent models [1, 9]. Additionally, this domain has the ongoing challenge of optimizing speech generation within VSR, which has historically received less attention than text generation. Training the models effectively with limited resources is another concern,

*Author for correspondence

emphasizing the need for streamlined approaches. Hence, further improvements are needed in generating speech.

The motivation for the research presented in this paper arises from the fundamental understanding of the critical role visual speech plays in communication, especially in challenging environments where audio recognition may be unreliable. There is a need for advancements in VSR due to human reliance on lip movements in noisy surroundings and when individuals face speech difficulties such as aphonia. The proposed research attempts to bridge the gaps associated with audio-visual model generalization, model complexity due to attention mechanisms [10], and the historical focus on text generation [9] over speech generation within VSR by proposing a novel sentence-level end-to-end VSR model.

The primary objective of this research is to propose and develop a sentence-level end-to-end VSR model capable of predicting speech from silent video inputs. The architecture of the proposed model leverages a convolutional neural network (CNN) combined with a long short-term memory (LSTM) encoder-decoder framework, providing an effective approach for spatial and temporal analysis. This methodological contribution ensures robust performance while emphasizing resource-optimized training, enhancing the model's practicality and accessibility. The addition of an inference layer further improves its real-time applicability, enabling swift speech prediction in diverse scenarios. Notably, the model addresses aphonia—a condition that impairs an individual's ability to speak. Potential extensions of this work include real-time speech generation, automated teller machine (ATM) personal identification number (PIN) recognition using silent lip movements, and speech generation from CCTV footage.

The subsequent sections of this paper are structured as follows: Section 2 presents the literature survey for connectionist temporal classification (CTC)-based models, seq2seq-based models, speech enhancement techniques, classification tasks in speech, and techniques addressing the homophones issue. Section 3 explores the methods for data pre-processing, feature extraction, model design, model selection, training the model, and inference layer. Section 4 provides a brief overview of the experimental setup, including dual-speaker categorization, results on WER, correlation analysis between mean squared error (MSE) loss and epochs, subjective human evaluation, and a comparative analysis of audio spectrograms. Section 5 presents the discussion and limitations. Finally, Section 6 concludes the proposed work along with future scope.

2.Literature review

With the rapid development of artificial neural networks, numerous methodologies have emerged and evolved to find a solution for the lip-reading problem. Recognizing the appropriate region surrounding the lip is vital in VSR to generate accurate speech. However, many obstacles prevent accurate lip region segmentation. Some obstacles include different lighting conditions, varying lip colours, and complex appearances related to an open mouth [11]. A fuzzy logic system organized as a deep neural network could extract the discriminative features at different scales. Fuzzy and convolutional units were integrated with this model. This network also provided comprehensive details for lip-segmentation at the pixel level [11].

Over the years, numerous datasets have been made available to facilitate the model training process for speech generation. These datasets can be broadly classified into two categories: constrained and unconstrained. The constrained datasets include lip reading in the wild (LRW) and Trinity College Dublin-Texas Instruments/Massachusetts Institute of Technology (TCD-TIMIT), Ouluvs2, and Grid audiovisual sentence corpus. In contrast, the unconstrained datasets consisted of lip reading sentences (LRS), Voxceleb, Lip2Wav, lip reading sentences 2 (LRS2), and lip reading sentences 3 (LRS3). Historically, the use of constrained datasets [5, 6] was prevalent; however, there was a surge in the popularity of unconstrained datasets [1, 7].

Various end-to-end CTC-based [7, 10] and seq2seq based [1, 3, 4, 8, 12] speech recognition systems were trained to predict text or speech sequences. A CTC-based model was generally used for tackling time-varying problems. It performed frame-wise predictions and tried to align the predicted results with the output sequences. The seq2seq model followed an encoder-decoder architecture. This model mapped the input sequence to an output sequence. It consisted of reading all the input sequences before predicting the output sequences.

Afouras et al. [7] showed no need for ground truth transcriptions for training a lip-reading system, as a pre-trained automated speech recognition (ASR)

model could generate text transcriptions from an unlabelled dataset. A teacher-student learning framework was adopted in this context, transferring knowledge extracted from a teacher ASR to a student VSR model. When combined with CTC, the cross-model distillation technique significantly improved training speed [7]. The performance of the ASR system was severely affected by surrounding acoustic noise. Tao and Busso [13] suggested a solution to address this problem by combining visual features of lip activity into audio-based ASR systems. This solution resulted in improvements concerning context word recognition compared to audio systems. An attention-based-CTC approach to generate text transcripts was proposed in [10, 14]. A hybrid lip-reading network (HLR-Net) proposed in [10] was an encoder-decoder architecture and performed CTC on a set of videos, producing corresponding subtitles. A CTC decoder based on cascaded attention generated an output text. This technique resolved the limitation associated with the independence assumption of CTC and effectively rectified the defect by introducing a CTC decoder.

A hybrid attention-based CTC model was based on a convolution-augmented transformer and residual network (ResNet)-18 [15]. An architecture without recurrent layers based on U-Net and extension of encoder-decoder methodology [9] was trained end-to-end to recognize phonemes. The work in [15, 9] indicated that encoder-decoder architecture embedded with existing networks improves performance and overcomes the issues related to phonemes. The research in [16] investigated the use of automatically generated transcriptions from unlabelled datasets to augment training sets. The study transcribed unlabelled data and combined it with established datasets using available pre-trained ASR models. The results demonstrated that augmenting the training set size reduces WER even while using noisy transcriptions.

The work in [17] combined convolutional three-dimensional (3D) neural networks with bidirectional LSTM networks and CTC. This model showed an accuracy of 15.8% WER and a 6.2%-character error rate, with less than 100 epochs. The proposed model may have performed differently than expected when the number of speakers was more and had varied accents. Kuriakose et al. [18] proposed a model using an attention architecture and an autoregressive encoder-decoder. This work included plots of the silent face expressions to mel-scale spectrograms. The intelligibility was measured with extended short-time objective intelligibility (ESTOI). The perceptual evaluation of speech quality (PESQ) was used to measure the quality of the generated audio signal. This system did not make use of human annotation. The earlier proposed fast conformer model was used to process audio and visual data with a hybrid CTC/ recurrent neural network (RNN)-T architecture [19]. The model used the LRS3 dataset and showed a WER of 0.8%. This work presented an audio knowledge-based visual speech recognition framework (AKVSR). This method used a pre-trained audio model with an LRS3 dataset and stored information in audio memory using quantization. There were two cross-attention layers with eight multi-head attention mechanisms. The AKVSR complemented the visual information of other VSR models. However, it required designing a time-efficient training method, as prior to the training of the VSR model, the construction of a compact audio memory was needed [20].

The advent of new text-to-speech (TTS) techniques centered around seq2seq learning acquired significant interest [1]. The state-of-art model Lip2Wav [1] was built on top of the TTS architecture. Modifications in the design of previous lip-to-speech tasks resulted in generating natural speech in unconstrained settings. Zhang et al. [21] proposed an acoustic voice conversion model that captured the connection between the audio and the articulation units in language. This work used seq2seq architecture to develop the model. This model performed two tasks for the voice conversion: one, auditory feature prediction, and two, generating waveform from the mel spectrograms of target speakers.

The spatio-temporal dynamics improved the lip-reading task combined with ResNet topology [12, 22]. Sequence-to-sequence-mapping by the alternating spatio-temporal and spatial convolutions (ALSOS) [12] module facilitated the conversion of sequences into clusters of feature maps. Zhang et al. [22] proposed a module that fused fine and coarse-resolution images to generate image sequences with adequate spatio-temporal resolution. The proposed module reduced feature dimensions and maintained the local spatial information. Here, the self-attention mechanisms reduced the training time. Lip reading was performed based on the experiments conducted in unconstrained natural settings [3, 23]. Realistic and unconstrained datasets were better than Grid audiovisual sentence corpus and TIMIT for learning accurate person-specific models. The work in [3] showed that generating sentences with or without

audio data in unconstrained natural settings was possible.

Using auditory and visual modalities is challenging for translating audio and video input streams into text. Afouras et al. [23] demonstrated that the seq2seq model performed better without audio, whereas the CTC model handled the noise in the background better. An attention mechanism was employed to dynamically align acoustic and visual modalities in the study by Sterpu et al. [24]. The audio-visual information fusion occurred within the encoder section by utilizing the seq2seq network. Notably, the visual modality proved to be more informative for longer sentences within the auditory modality.

The pseudo-convolutional policy gradient (PCPG) approach [25] overcame the inconsistency between the final evaluation metric and discriminative optimization target. Treating the decoder as an agent in a reinforcement learning environment and applying the PCPG approach to calculate the loss showed significant improvement in the model compared to the traditional approach. Vid2Speech [26] was one such model. Ephrat and Peleg [26] proposed a CNN-based model that takes an overlapping sequence of input frames and generates speech features. Waveform from digital speech's linear predictive coding (LPC) was synthesized. Although CNNs could generate speech, they needed to be more complex to learn the temporal features of the input sequence. This shortcoming of CNN was solved by incorporating LSTM architecture.

Arthur and Csapó [27] proposed a deep neural network-based system. Here, they performed the feature extraction with CNN and the final classification with LSTM. The work in [28] discussed two separate networks: one, an autoencoder network to extract feature vectors (FV) from a spectrogram, and the other, an LSTM lip-reading network to generate bottleneck features. Both networks were combined to reconstruct intelligible speech. The addition of LSTMs resulted in overall speech quality improvement. The difference between generated speech and target speech was calculated using MSE-based training of the speech recognition domain. A generative adversarial network (GAN) based training strategy provided more promising results than MSE-based training. Shandiz et al. [29] performed adversarial training with the help of a discriminator. This trained discriminator distinguished between mel-spectrograms generated by real input and those produced by generators. Application of the GANs framework in speech enhancement and voice conversion proved very successful.

Prajwal et al. [30] presented a sub-word level tokenization technique to address the ambiguities in speech generation tasks. An encoder-decoder transformer estimated sub-word probabilities based on frame-wise temporal features. This transformer incorporated an attention-based pooling mechanism. The neural network in [31] was a customized version of the three-dimensional residual network (3DResNet)-18 model enhanced by incorporating the squeeze-and-attention module. This modification enabled the extraction of highly informative visual features. The module used in this work exploited the multi-scale information for feature extraction.

The work in [32] recognized the significance of incorporating facial expressions as a salient feature in the recognition process. The study employed a CNN-LSTM deep learning video processing model. The results revealed that features extracted using the GoogleNet model exhibited superior classification accuracy compared to the AlexNet and ResNet models. A combined model with CNN, LSTM networks, and an adaptive interest mechanism was proposed [33]. CNNs were used to capture spatial capabilities, and LSTM networks were used to examine temporal dependencies. To determine the efficacy of these models, a comparative analysis of parameters such as WER, character error rate, and frame error rate was performed using a mixture of techniques.

Techniques for speech enhancement were also required to develop efficient ASR and AVSR models. Several studies [34–37] aimed to improve speech signal's perceptual fidelity. Visual information helped in the speech enhancement process. An audio visual (AV) framework based on speech enhancement is operated on multiple levels [34]. In this study, the first level employed a lip-reading model built on deep learning. The second level was about the audio power spectrum estimated by a visually derived Wiener filter (EVWF). Hou et al. [35] performed reconstruction of audio and visual streams based on a multi-task learning framework, audio-visual deep CNN (AVDCNN). AVDCNN followed an encoder-decoder architecture network.

Thimmaraja et al. [36] observed that encoding the noisy speech data with the LPC algorithm resulted in

lesser speech quality. The audibility enhancement of encoded speech data occurred due to combining LPC and spectral subtraction voice activity detection (SS-VAD). The visual information, in particular, enhanced speech even in noise-contaminated audio. Sadeghi et al. [37] proposed variational auto-encoders (VAEs) with audio-visual variants for speaker-independent and single-channel speech enhancement. Generating audio speech based on visual information of the lip region resulted in the development of conditional VAE.

Most of the VSR literature classified the output by label. The visual geometry group (VGG) network classified and recognized data [38]. This work used VGG16 to identify five words using the video input without audio. The proposed method surpassed the Hahn CNN architecture for performing the classification task. A Network for flow deformation deep functional network (DFN) [39] learned the deformation flow between adjacent frames. The prediction of probability for each of the word classes was done independently by each of the branches specified in the given work. Image feature extraction was an effective and fault-tolerant method to detect the desired region of interest for a VSR system [40]. The extraction of lip image features using the VGG network resulted in higher accuracy than the conventional method.

Homophenes are words that exhibit identical lip movements but distinct pronunciations. Distinguishing the homophones has been a prevalent issue in the speech recognition domain. Thangthai and Harvey [41] investigated the effect of feature transformation using phonetic class discriminant features. The work also identified whether the data were grouped according to speaker or linguistic similarity using Fisher's ratio. The work [42] combined a CNN with a model that handled multiple time-based data streams. It followed a re-synchronization procedure wherein the hand features aligned with the lip features. The model achieved a significant performance improvement in phoneme recognition. Techniques introduced in [2, 42] addressed the dominant issue of homophenes.

The work in [43] used CTC-based techniques for unsupervised phonemes and word segmentation. The performance of phoneme segmentation and classification together could be compromised. This effect was lessened by manually removing the offset from the representation or using an auxiliary contrastive loss between consecutive latent representations. The study [44] proposed a method for phoneme segmentation based on the regularization attention mechanism. The attention mechanism fused the speech and phoneme sequences. Speech feature representations were learned based on a pre-trained acoustic encoder. A phoneme encoder encoded the phoneme representations of the pronounced phoneme sequences. Phoneme conversion was suitable for continuous speech recognition. The work in [45] developed the k-nearest neighbour (KNN) automatic generation model to study the rules of morpheme-phoneme conversion. This work indicated that the independent conversion of each phoneme type could effectively avoid Gaussian mixture models (GMM) smoothing, thus making the converted speech closer to the target speaker.

In conclusion, most of the literature is categorized as audio [21, 42], visual [1, 12], and audio-visual [4, 23] models for generating speech. Various works [13, 15, 24] presented models for improving speech quality or speech recognition given audio and visual inputs. Some previous works performed speech generation by focusing on the visual inputs [8, 23, 26] without considering a continuous stream of video of varying lengths.

Future research needs to focus on explicitly solving the issue of speech impairment. The proposed work in this paper is a significant contribution to the field, closely aligning with solving the difficulty for speech-impaired individuals. The proposed work generates speech solely by visuals without considering audio signals as input. An inference layer generated audio for a continuous stream of variable length of input images. The proposed seq-2-seq model shows the ability to reduce the problem of speaker dependency at the sentence level. This research introduces a training technique to overcome overfitting in limited resources. This work also explores the potential of VSR models for real-time applications.

3.Methods

The methodology includes identifying the dataset for the experiments, implementing audio and video feature extractions and finalizing the architecture of the Lip2Voice model. The methodology is discussed in detail in the following subsections.

3.1Dataset

An English-based audio-visual corpus named Grid audiovisual sentence corpus [46], with a total of

33,600 sentences, is used for training the model. The dataset includes audio and video recordings of 34 speakers, including 18 males and 16 females, primarily with British English accents. However, regional variations result in phonological differences, vocabulary, and lexical variants. The dataset contains a set of predefined sentences that cover a range of phonemes and linguistic contexts. The recordings took place in constrained environment settings. The audio and video files are synchronized, and the speakers' movements are aligned with their corresponding audio recordings. The sentences, for instance, are “lay green by F 7 now” and “bin blue by M 1 please”. The speaker's age ranges from mid to late adulthood. This work conducted training and evaluation of the proposed model using videos from two male speakers (S2, S23). The data were pre-processed and then fed to the neural network.

3.2Feature extraction

The data samples consist of a set of silent video and audio files, each of 3 seconds long. The experiments used a total of 2000 videos as data samples. 80% of the video data samples serve for training purposes, and 20% for validation and testing. The data combines two speakers (S’, S”) to avoid speaker dependency. Training is conducted with two speakers simultaneously to avoid speaker dependency and to process the order of the same speaker. Let S’ = [S’$_1$, S’$_2$, S’$_3$ . . ., S’$_n$] and S” = [S”$_1$, S”$_2$, S”$_3$ . . ., S”$_n$] where n is the number of data samples training sets. The dataset can be represented as shown in Equation 1:

$$S_{(1-2n)} = \sum_{i=1}^{n} S' + \sum_{i=1}^{n} S'' \quad (1)$$

A shuffling operation is performed with the samples from both speakers to prevent any ordering bias during the training process. The obtained output is as shown in Equation 2.

$$S_{(1-2n)} = [S_1', S_1'', \ldots, S_{24}', S_{34}'', S_n', S_n''] \quad (2)$$

Here each sample is an audio and video file as shown in Equation 3.

$$S_{(1-2n)} \rightarrow V_{(1-2n)} + A_{(1-2n)} \quad (3)$$

Audio feature extraction is as shown in Equation 4.

$$A_{(1-2n)} = [a_1, a_2, \ldots, a_{2n}] \quad (4)$$

Audio feature extraction

This work uses the fast forward moving picture experts group (FFMPEG) tool to extract audio from sample video files. The extracted audio is resampled to fs = 16kHz. The number of audio samples recorded every second for fs/2 has resulted in fewer sample points. These sample points represent the audio that has resulted in an unclear speech. Concerning the previous speech recognition works [40, 41], the Lip2Voice model utilizes it to produce mel spectrograms. The short-time fourier transform (STFT) is used to convert the time domain to the frequency domain[1]. Mel-filter banks transform the STFT magnitudes to mel-frequency. The window size is 800 and the hop size 200, with 80 mel filters and a maximum of 900 mel frames. The mel step size is 240, and the mel overlap is 40. The expected mel-shape required is (239,80). Hence, the generated mel spectrograms are reshaped to a uniform size by padding with zero and flattening to form a one-dimensional (1-D) array. This output array represents the target audio values, and each row consists of the flattened mels for each audio file.

Video feature extraction

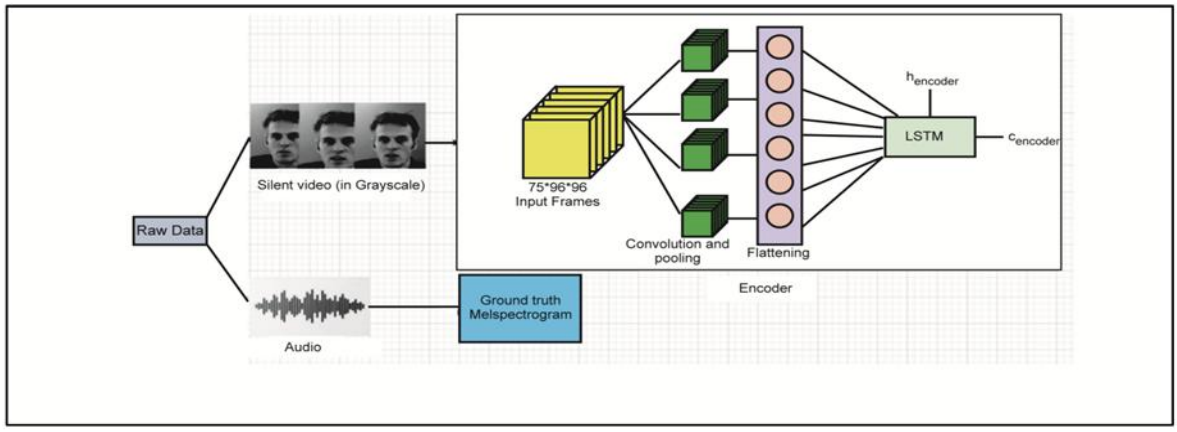
OpenCV is employed to extract individual frames from the provided input video files. The model generates frames using a fixed frame rate of 25fps. The video files have a duration of 3 seconds, resulting in a sequence length of 75 frames. In instances where the count of frames in the video is fewer than the sequence length, padding is applied using values from the second last frame. A Haar cascade classifier [21–25] performs face detection on each frame to isolate the desired region of interest around the captured frames. The captured frames are cropped and converted to grayscale to avoid wastage of processing power. Additionally, grayscale images provide adequate details for identifying the lip shapes associated with each word. The data is further normalized by resizing to (96, 96) and expanding the dimensions.

The audio and video features are processed separately and saved in NumPy files. At the start of the training process, these NumPy files are loaded and processed before feeding into the model. The array of mel spectrograms is normalized by subtracting the mean and then dividing by the standard deviation of the audio values in the sequence. The pixel values are normalized; this is done by subtracting the mean from each pixel value and then dividing by 255 to scale the values appropriately. The VSR model consists of a 3D convolutional encoder used to extract spatiotemporal features out of the frames of videos. The extracted spatiotemporal features undergo compression to form FVs. These FV are then given to the LSTM decoder. The LSTM decoder decodes the FV to learn the seq2seq linkage between the values. This LSTM output produced from the

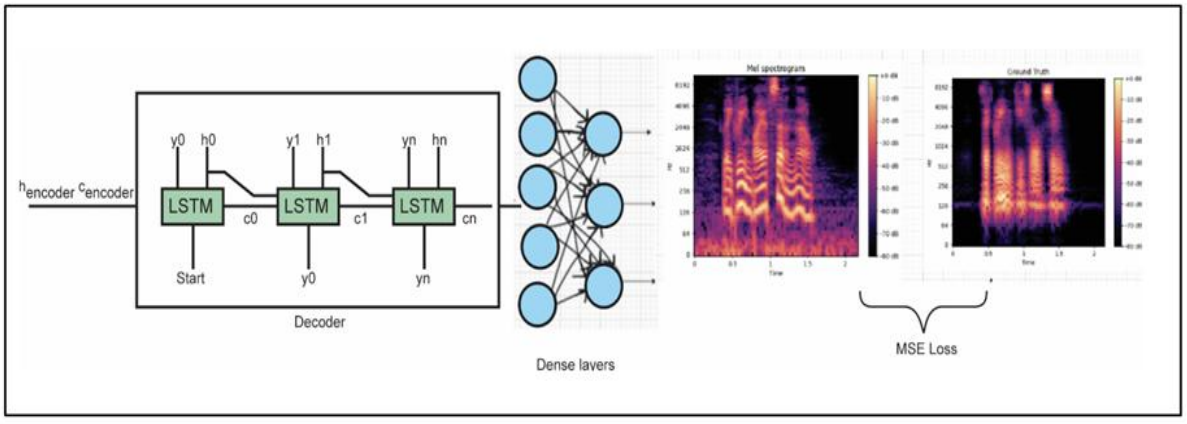
decoder generates the mel spectrogram. The model compares the ground-truth mel spectrograms obtained from the original audio with the generated mel spectrograms. The difference between the generated and original mel-spectrogram gives the loss function. The generated mel-spectrogram is used to reconstruct the speech.

3.3Model design

This section covers a detailed discussion of the model selection, training methodology, and the inference layer for continuous speech generation. *Figure 1* depicts the structure of the Lip2Voice model proposed in this study. Lip2Voice is an auto-encoder (AE). The model's architecture includes a 3D convolutional encoder for extracting spatiotemporal features, followed by an LSTM decoder and a set of dense layers for speech generation. The Griffin-Lim algorithm reconstructs the predicted mel spectrogram to generate speech. *Figure 2* represents the Workflow diagram of the Lip2Voice model.



A. Encoder



B. Decoder

Figure 1 High-Level architectural view of Lip2Voice model with (A) an encoder that combines CNN and LSTM to extract the spatial and temporal features from silent video and (B) a decoder that generates desired mel-spectrogram from the encoded FV

3.3.1Model selection

The proposed work conducted experiments with different models and compared their training performance to identify the best-fit model.

Table 1 shows experiments conducted based on various models. All models are layered above autoencoder architecture. Each consists of a set of 3D convolutions and LSTMs acting as encoders and a set of LSTMs for decoding the learned features. Lastly, two dense layers are included to generate predictions. Based on the performance of the different models studied, the Lip2Voice model selected the Bahdanau

attention mechanism, rectified linear unit (ReLu) activation function and MSE loss functions.

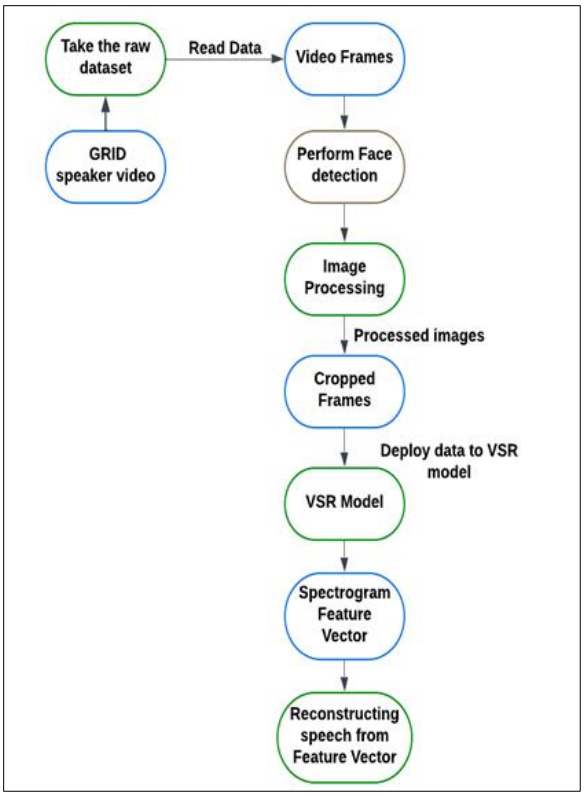


Figure 2 Lip2Voice workflow diagram

The first architecture follows the most widely used conventional techniques for building a seq2seq model. The experimental analysis of model-I leads to the development of a complex architecture with an extensive number of learning parameters, totalling around 385 million. Since there has been no learning, it is decided to decrease the number of parameters. The addition of three convolution layers in model-II resulted in a reduction in parameters. In the initial epochs, model-II shows some learning but later flattens out and stops learning altogether. Model-II performs better than model-I, as adding layers of convolutions reduces the FV dimensionality derived from the input image sequence. This results in a substantial decrease in the overall trainable parameters, reducing them to 91 million. This result indicates that the earlier problem faced during the experimentation still needs to be entirely resolved, and the model needs to be simplified for the data samples.

The subsequent experiments employ reduction in number of filters in each layer, in contrast to the earlier approach of incorporating additional convolution layers. From model-II to model-III, the number of parameters is reduced to 58 million. The parameter reduction resulted in a model showing a fair amount of learning for a few epochs, but the learning performance degraded in the later stage. The progress in initial learning in model-III compared to model-II follows the guiding principle. The principle is that the number of filters selected in each problem is directly correlated to the intricacy of the features the model learns. The above principle confirms that model complexity is not the only problem. The volume of data per sample passed to the model is insufficient for the model to learn the features. More data samples are needed to sustain the model's learning process. Hence, it is crucial to prioritize the quantity of data fed to the model.

The observations have indicated two significant challenges: model complexity and data volume per sample need resolution. The first three architectures experiment with the model complexity attribute. Model-IV shows some significant changes made to overcome the two challenges discussed. Hence, the number of 3D convolutions increases, and the number of LSTM layers of the decoder decreases. At each timestep, the volume of data per sample increases by changing the number of input image sequences. As a result, model-IV is less complex than model-III. The increase in the volume of data per sample reduces the model's dependency, as the entire context of the sample is now covered. As LSTM can handle these short-term dependencies, the reduced volume of data per sample led to the decision to remove the attention mechanism from the architecture altogether. This architectural change allows model-IV to continue learning for several epochs. However, it was later observed that the model is overfitting the dataset used for processing, and hence, a need to increase the dataset is felt.

The observations made in the previous architecture help us understand that more data is needed to train the model. Hence, model-V considers the dual speaker dataset over the single speaker dataset. Also, a new approach called “Alternative Training” trains the model. In this approach, the model trains for the first 200 epochs on the primary dataset, then for the following 100 epochs on a similar dataset, and alternates between the primary and similar datasets until reaching 507 epochs. The model converges at this point. This technique allows the model to focus more on the sentences being used than the speaker-dependent features, resulting in early convergence.

The model reduces the number of 3D convolutions from ten to six and adds dropouts and max pooling layers to reduce overfitting. All these changes eventually reduced the trainable parameters to 18 million, speeding up the training processes. The generated speech is audible, interpretable, and intelligent, and the model can make predictions on an unseen test set of two mixed speakers simultaneously. As a result, model-V, "Lip2Voice" is the best-fit architecture and has shown promising results compared to model-IV. The generated speech is audible, interpretable, and intelligent, and the model can make predictions on an unseen test set of two mixed speakers simultaneously.

Table 1 Experimental model configuration

Models	Trainable parameters	Conv layers	LSTM layers	Dense layers	Attention mechanism
Model-I	385,860,833	6	3	2	Bahdanau
Model-II	91,621,297	8	3	2	Bahdanau
Model-III	58,318,401	8	3	2	Bahdanau
Model-IV	39,466,334	10	2	2	-
Model-V	18,697,918	6	2	2	-

3.3.2Training the model

The input data loaded are pre-processed video and audio samples stored as NumPy files. The data is loaded, and normalization is performed as the first step to scale the data equally. The training is performed in batches using the Adam optimizer to adapt the learning rate dynamically based on the gradient of the loss function. During the experimentation, decreasing the learning rate progressively balanced fast convergence and stability in the training process. *Table 2* provides a detailed model configuration. The CNN, max pooling, LSTM, and fully connected layers are represented as Conv3di, MaxP3di, Lstmi, and FCi, respectively. In the first step, the batch of input data for each shape (75, 96, 96, 1) passes through the 3DCNN layers. Equation 5 indicates the FV.

$$FV = Conv3d6(Conv3d5(MaxP3d2(Conv3d4(Conv3d3(MaxP3d1(Conv3d2(Conv3d1(X_i)))))))) \quad (5)$$

The learned FV is reshaped and passed through the last layer of the encoder, which is an LSTM, as shown in Equation 6:

$$E = Lstm1(FV) \quad (6)$$

Table 2 Configuration details of Lip2Voice

Layer	Kernel Size/Pool size	No. of filters. neurons	Activation function
Conv 1	3×3×3	32	Relu
Conv 2	3×3×3	32	Relu
MaxP1	2×2×2	-	-
Conv 3	3×3×3	32	Relu
Conv 4	3×3×3	32	Relu
MaxP2	2×2×2	-	-
Conv 5	3×3×3	64	Relu
Conv 6	3×3×3	64	Relu
LSTM 1	-	1024	Tanh
LSTM 2	-	1024	Tanh
FC1	-	1024	-
FC2	-	478	-

The encoder's output states are preserved and serve as the decoder's starting state. The decoder architecture combines an LSTM layer followed by two dense layers. Equation 7 represents the decoder architecture.

$$D = FC2(FC1(Lstm2(E))) \quad (7)$$

The decoder predicts the mel spectrogram values by returning the output of the hidden layers at each time. The model generates the output iteratively to capture the temporal dependencies between the video frames. Equation 8 shows the model appending the predicted values at each time step to form the output vector.

$$Y_i = \sum_{t=1}^{T} D_i^t \qquad t = 1 \ldots . T \quad (8)$$

Here, T is the number of time steps. The last dense layer consists of 478 units, which make the output vector of shape (478,40) each. The output vector is reshaped into a one dimensional (1D) array to calculate the loss with the target mel spectrogram values. MSE function is applied to reduce the Lip2Voice loss, represented as L. Equation 9 shows the calculated loss used for backpropagation and training the neural network.

$$L = min\,\theta\left(\frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \widehat{Y_i}\right)^2\right) \quad (9)$$

The model applies dropouts at intermediate stages to reduce overfitting. The model is trained for 507 epochs to optimize the performance. The model adopts a different training approach to avoid overfitting. During training, the model trains the first 200 epochs on the input samples of 2 speakers. The following 100 epochs use a secondary dataset of 2 different speakers with the same set of words. This approach allows the model to learn different pixel values for the same word. This training approach prevents the model from entering an overfitting stage. Dropouts are also added between the networks to overcome overfitting. During each training iteration, a certain proportion of the neurons are removed randomly from the network. Algorithm 1 depicts the complete Lip2Voice model architecture.

Algorithm 1: Lip2Voice Model architecture

1. Preprocessing
   1.1 Normalize data
   $$X_t' = \frac{X_t}{255} - \mu_{x_t}, \qquad Y_t' = \frac{Y_t - \mu_{Y_t}}{\sigma_{Y_t}}$$
2. Model Architecture:
   2.1Input layer: $X_{input} \in R^{B_{batch}*75*96*96*1}$
   2.2 Convolution layers: $X_{conv} = Conv3D(X_{input})$
   2.3 Reshape for LSTM input:
   $$X_{reshaped} = reshape(X_{conv}, (-1, time\ steps, d))$$
   where d=(h*w*d)
   2.4 (h_{enc}, C_{enc}) =LSTM($X_{reshaped}$
   2.5 For each timesteps:
   2.5.1 $\left(h_{dec}^{(t)}, C_{dec}^{(t)}\right) = LSTM(X_{dec}^{(t)}, \left(h_{dec}^{(t-1)}, C_{dec}^{(t-1)}\right))$
   2.5.2 $Y_{out}^{(t)} = dense(h_{dec}^{(t)})$
   2.5.3 $Y_i.append\left(Y_{out}^{(t)}\right)$
   2.6 Loss calculation:
   $$L = \frac{1}{N}\sum_{i=1}^{N}(Y_i - Y_i^{\wedge})^2$$
   2.7 Gradient Update:
   $$\theta = \theta - \eta\nabla_\theta L$$
3. Predict
   $$Y = f(X_{input})$$

3.3.3 Hyperparameter optimization

Training is performed in batches of eight silent videos using the Adam optimizer. The initial learning rate is 0.001, and it is decreased from 0.001 to 0.0001 after 200 epochs to achieve smooth convergence of the loss function. The decoder's output is generated iteratively for 40 timesteps. Fifty percent dropouts are also added between the networks to overcome overfitting. Algorithm 2 outlines the step-by-step procedure of the complete inference layer.

Algorithm 2: Inference Layer

Input :
1. Data stream X = $\{x_n\}_{n=0}^{N}$
2. Saved weights of the model (W)

Output :
1. Generated speech: Y

Begin :
1. Load model.
2. Read data stream X
   2.1. Detect face : f = $\{f_n\}_{n=0}^{N}$
   2.2. Grayscale conversion: $f_n \leftarrow G(f_n)$
   2.3. Resize $f_n^{96*96} \leftarrow f_n^{p*q}$
   2.4. After every 3sec
      2.4.1. Y = Predict($\{f_n\}$n=0,n=75)
      2.4.2. Generate Speech : Y = GRIFFIN_LIM(Y)
3. End of data stream.

End

A user interface is available, offering the option to choose a preferred video and a pre-trained weight file. The model predicts using the resulting speech. The user can upload videos of any length, making predictions every 3 seconds. The inference layer reads 75 frames at a given time. It initiates the process by conducting face detection and converting the image to grayscale. The preprocessing module resizes the images to match the input shape of the Lip2Voice model, enabling predictions. Griffin-Lim's reconstruction technique transforms the generated mel spectrograms into wav files. This speech generation process continues until the end of the data stream.

4.Results

4.1Experimental setup

The experiments used cloud resources and the P5000 graphics processing unit (GPU) machine for processing, storage, and training the model. An i7

processor-based system with 16 gigabytes (GB) of random-access memory (RAM) was used for inference to generate speech on variable-length video.

4.2Experimentation with dual speaker categorization

To thoroughly evaluate the dual-speaker model's performance, three distinct datasets are employed to conduct the experiments. The datasets are prepared based on varying pitch levels consisting of two male speakers, two female speakers, and one male and one female speaker. Throughout the training process, the model achieved better speech generation outcomes when dealing with speakers of similar pitch levels, such as datasets of speakers of the same gender. In contrast, datasets containing a combination of male and female speakers demonstrated lower speech quality. An attempt was made to train the model with four to five speakers, but the data available per speaker in the Grid audiovisual sentence corpus was insufficient.

4.3Analysing architectures

This work compares various architectural improvements with those of Vid2Speech [26], CATNET [47], and AKVSR [48]. Vid2Speech and CATNET utilized CNNs for processing input data. In contrast, the research work of this paper emphasizes the importance of capturing both spatial and temporal dependencies within images, leading us to promote the use of three-dimensional convolutional neural networks (3DCNN). Although CNNs are proficient at extracting local features, they struggle to capture global dependencies throughout the image. This limitation affects their performance in tasks requiring a comprehensive understanding of the visual context. In contrast, the Lip2Voice model leveraged transformers that are well-suited for this purpose due to their self-attention mechanism. This mechanism allows them to model long-range dependencies more effectively. This capability enables the Lip2Voice model to understand and analyze the intricate relationships between various elements within the image over time. It leads to better performance in tasks requiring temporal coherence and spatial awareness.

Moreover, the novel training methodology used in this research enhances the Lip2Voice model's performance and ensures that it operates independently of the speaker. This dual speaker independence is a significant advantage over AKVSR proposed in [48], which inherently depends on a single speaker. By making the proposed approach dual-speaker-independent, the model's versatility and applicability across diverse speaker scenarios are enhanced. It is also a more robust solution for real-world applications in speech recognition and related fields.

The proposed Lip2Voice model is highly resilient and can operate effectively in different environmental conditions. A comparison of Lip2Voice's work with [48] highlights that [48] disregarded crucial non-linguistic elements such as noise. Analysis of these factors provides a context that helps interpret speech in ambiguous situations.

The hybrid CTC/attention loss approach [48] was employed to train the speech recognition model. Also, in [47], a CTC loss approach was utilized to analyze temporal information. However, the proposed Lip2Voice model employs a different methodology for utilizing MSE loss. Using MSE loss allows for the minimization of the average error, thereby improving the performance. Furthermore, the model is not constrained by the intricacies of complex sequential dependencies present in the hybrid CTC/attention loss approach [48]. The comparison of these approaches reveals that the use of MSE loss in the Lip2Voice model provides a more computationally efficient approach without compromising on performance. This efficient approach especially advantageous in scenarios where complex sequential dependencies are not required and in real-world applications with limited resources.

4.4Results on WER

This work is compared with Vid2Speech [26] for WER, which is employed to assess the accuracy of the words generated by the model. WER is shown in Equation 10.

$$WER = \frac{(S+D+I)}{N} \quad (10)$$

where, S = count of substituted words, D = count of deleted words, I = count of inserted words, N = total number of words.

The current work has chosen Vid2Speech for the comparative analysis due to the identical datasets used for training. The input conditions are similar due to the utilization of only silent videos without accompanying audio. For Lip2Voice, WER is calculated using the test split, which consists of 64 samples and 32 videos from each speaker. WER calculation includes dividing the number of substitutions (S), insertions (I), and deletions (D) by

the total number of words, which is consistently six across all videos in the Grid audiovisual sentence corpus. Observations show substitutions are more prevalent than insertions or deletions in the tested videos. These observations occur when the model incorrectly associates a lip pattern with a different word rather than the intended word, which often arises with homophones. On a positive note, the model does not produce extra or unnecessary words, indicating that it does not exhibit additional hallucinations. It accurately predicts words, showing reliability even in scenarios without lip movements.

Calculation of the Vid2Speech model's WER refers to [4]. The results in *Table 3* indicate a significant 11.59% reduction in WER compared to Vid2Speech. Training Lip2Voice on dual speakers allowed the model to learn more word combinations from the Grid audiovisual sentence corpus and to identify generalized lip patterns from the sentences instead of getting overfitted to just one speaker.

Table 3 WER score on the unseen test split of the Grid audiovisual sentence corpus

Method	WER
Vid2Speech	44.92%
Lip2Voice	33.33%

4.5Results on loss vs epoch

Figure 3 illustrates the correlation between MSE loss and the number of epochs throughout the training phase. It provides a comparative analysis between the Lip2Voice model and the pre-existing Vid2speech model. In the current research, the Vid2Speech model is simulated, as given in [26]. The model attained a 50.75% loss after training for 73 epochs on a single-speaker dataset. In contrast, experiments show that the Lip2Voice model attained a loss of 36.6% after training for 507 epochs on the dual speaker dataset. Continued training shows that both models overfit after a while. Thus, the training has been stopped at the respective minima. Vid2Speech, a simple CNN model, requires less training than Lip2Voice, which, as an autoencoder, requires more training to understand the dual speaker dataset.

4.6Results on subjective human evaluation

In this Experiment, fifty-three human participants listened to five predicted audio files generated from both Lip2Voice and Vid2speech models. They heard the five audio files in one sentence. The participants counted the number of words mispronounced out of the six words in each sentence for all five audios. Equation 11 represents the percentage error of mispronounced words.

% error of mispronounced words = count of mispronounced words/total participants (11)

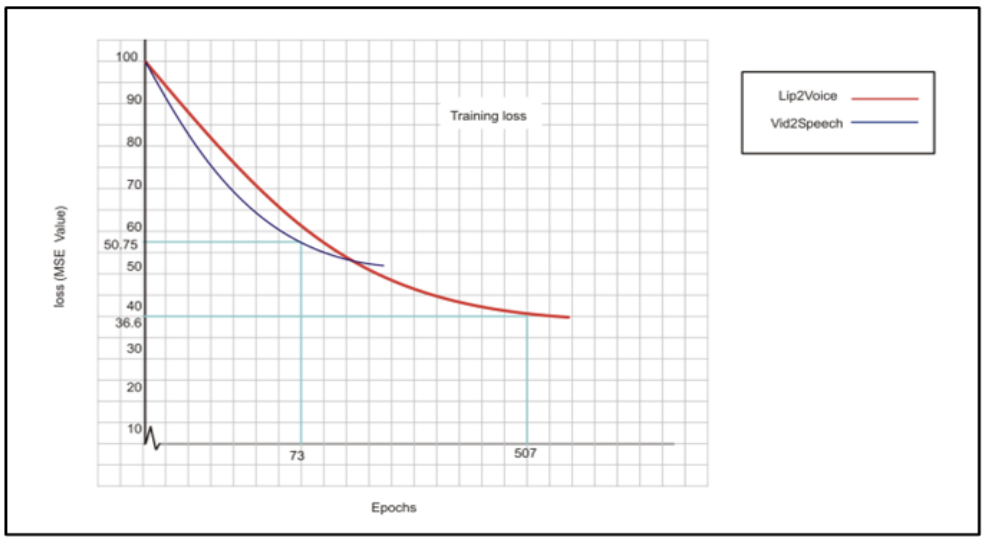


Figure 3 Training loss comparison of Lip2Voice and Vid2Speech models

The number of mispronounced words reported by each participant is summed across all fifty-three participants to calculate the percentage error for mispronounced words. This process is conducted separately for each of the five audio clips. *Figure 4* depicts the results of the survey conducted with fifty-three participants as mentioned above. As per the survey, the error percentage for all six mispronounced words is comparatively lower for the Lip2Voice model than for all five audios. In *Figure*

4(a), for audio 1, the average percentage error for the Lip2Voice model is found to be 13.84% as compared to the Vid2Speech model, where the average percentage error is 15.09%. In the case of audio 2 (*Figure 4(b)*), the average percentage error for both models is the same, which is 13.67%. Further, in *Figure 4(c)*, the average percentage error for Lip2Voice and Vid2Speech is found to be 8.80% and 12.26%, respectively, in the case of Audio 3. In the case of audio 4, the Lip2Voice model showed an average percentage error slightly higher than the Vid2Speech model, i.e., 14.46% and 13.83%, respectively, as depicted in *Figure 4(d)*. In *Figure 4(e)* for audio 5, the average percentage error for Lip2Voice and Vid2Speech is 11.32% and 12.58% respectively. The Lip2Voice model showcased a lower percentage of error overall than the Vid2Speech model.

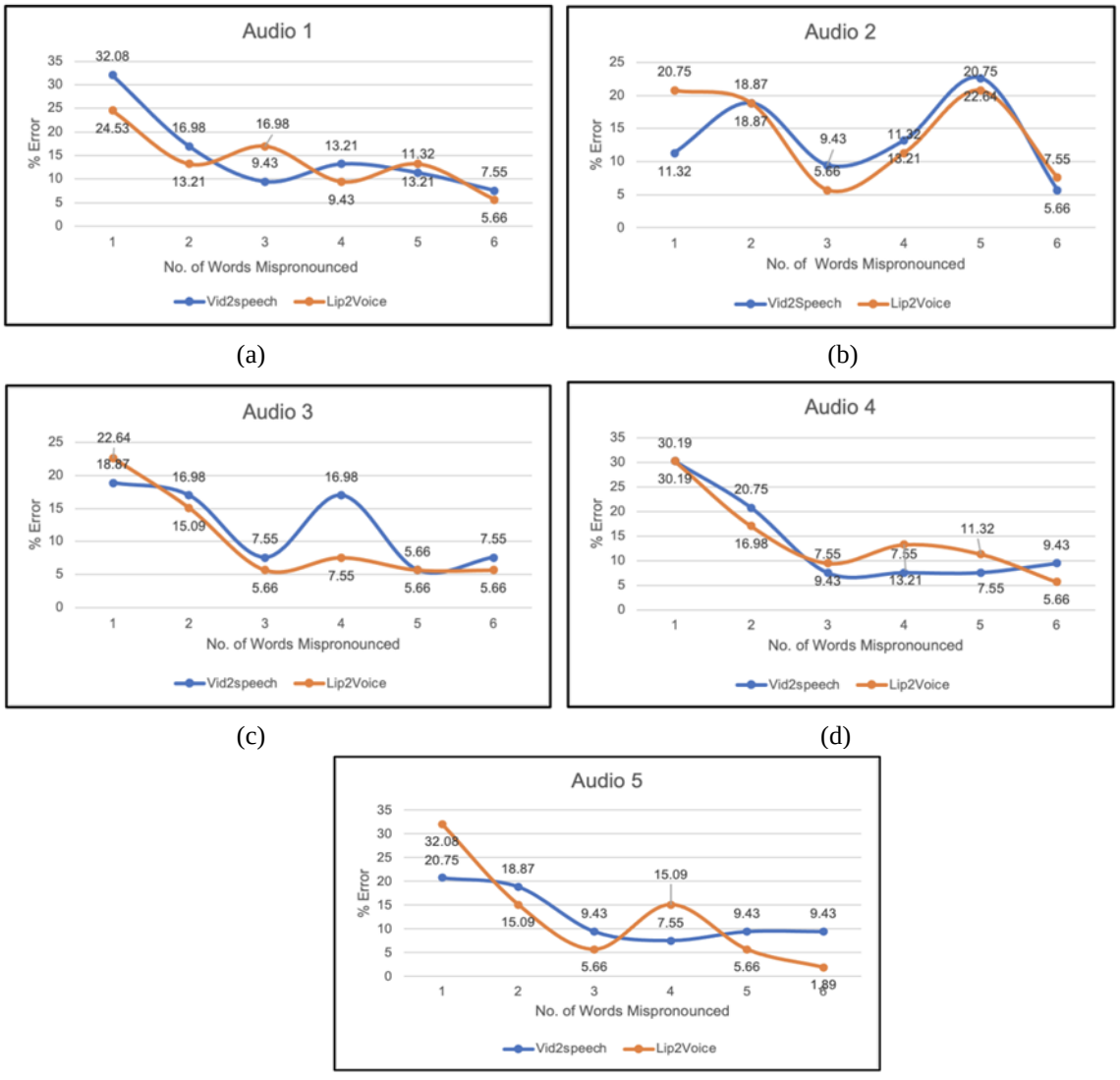


Figure 4 Percentage of errors in the number of mispronounced words for (a) Audio 1, (b) Audio 2, (c) Audio 3, (d) Audio 4, and (e) Audio 5

In *Figure 5*, the graphs show the audibility ratings for all five audio files as evaluated by the fifty-three participants in the survey. Audibility is defined based on the understandability of the generated sound for the human ear (clear/unclear) and background noise in the generated sound (noisy/without noise). The

participants provided quality ratings for the audio concerning four audibility ratings, namely,1) Unclear and/or Noisy, 2) Unclear without noise, 3) Clear but noisy, and 4) Clear without noise. *Figure 5* shows that the Lip2Voice model has lower percentage ratings than Vid2Speech for all five unclear and noisy or unclear without noise audios. However, for audios that are clear but noisy or clear without noise, the percentage ratings for Lip2Voice are higher than Vid2Speech, which signifies that Lip2Voice has resulted in clearer audibility with/ without noise for all five audios.

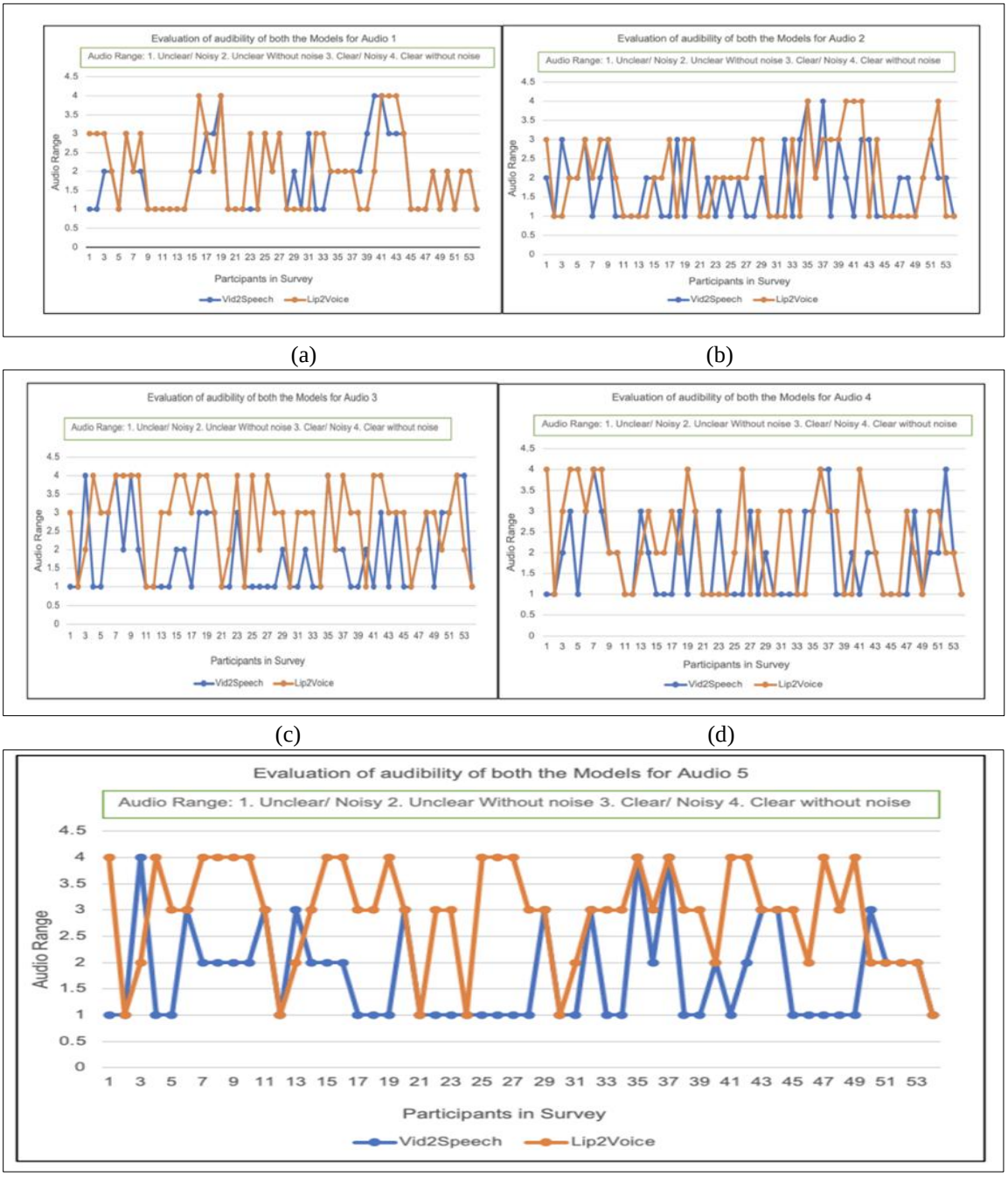


Figure 5 Audibility ratings for (a) Audio 1, (b) Audio 2, (c) Audio 3, (d) Audio 4, and (e) Audio 5

Figure 6 compares audibility ratings for Lip2Voice and Vid2Speech by considering all five audio files. It indicates the participants' perceptions of both models while listening to the audio files. For Lip2Voice, there is a substantial decrease in the number of participants voting for the ratings, namely unclear and noisy and unclear without noise. Also, for Lip2Voice, there is a considerable increase in the number of participants voting for the ratings, namely clear but noisy and clear without noise.

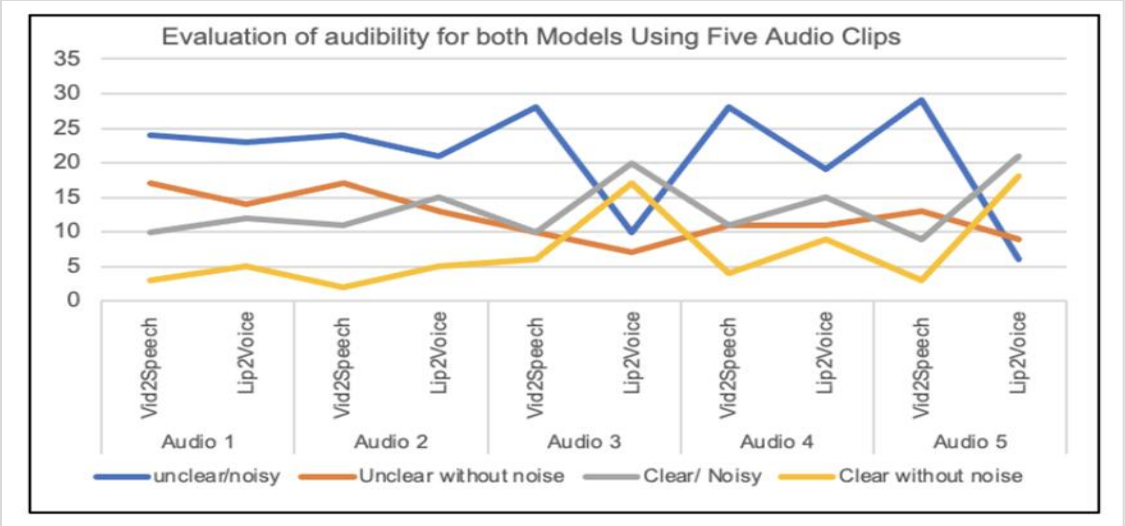


Figure 6 Evaluation of audibility of both models using five audio clips

4.7Comparison of audio spectrograms

Table 4 compares noise regions and noise bands of original and generated audios of Lip2Voice and Vid2Speech for an audio sample 'lay green by F7 now' present in the Grid audiovisual sentence corpus. Experiments included different frequency bands for different noise types, where the white band = (0,1000), the pink band = (1000,4000), the Brownian band = (4000, 8000), and the impulse band = (8000, 10000). The average power spectral density (PSD) is calculated within each frequency band, thus determining the dominant noise type based on the PSD in each frequency band [49]. The PSD for Lip2Voice audio resembles the original audio more closely than Vid2Speech.

The work analyzes the first three significant resonance peaks in the frequency spectrum of the speech signal, namely, the first formant (F1), the second formant (F2), and the third formant (F3). Each formant contains information about how the speaker's vocal tract shapes the sound. This work calculates formant features like power, frequency, width, and dissonance. Here, frequency is the central frequency of the formant; power is the intensity at the formant frequency; width is the bandwidth around the formant frequency, which indicates how spread the formant is; and dissonance is the measure depicting how much the formant deviates from a harmonic of the fundamental frequency. For each frame, the calculated average values of these features for the first three formants provide a detailed acoustic profile of the speech in the audio file. *Table 5* details the four formant features for original, Vid2Speech, and Lip2Voice audio. The focus on three formants is important in understanding speech intelligibility.

Noticeable differences arise when comparing the Lip2Voice and Vid2Speech models to the original model. Lip2Voice exhibits higher mean frequencies (20.60% increase), indicating potential shifts towards higher-pitched sounds, along with narrower mean width (4.34% decrease), suggesting sharper resonances and a significant increase in mean dissonance (104.68%), indicating higher roughness in generated speech. In contrast, Vid2Speech demonstrates lower mean frequencies (20.98% decrease) and mean power (15.63% decrease), implying shifts towards lower-pitched sounds with weaker resonances. While Vid2Speech also displays a slight increase in mean width (1.14%), possibly indicating broader spectral distributions, it significantly reduces mean dissonance (54.17%), suggesting improvements in speech quality. These differences underscore distinct spectral characteristics and potential perceptual implications, which are crucial for evaluating and refining speech synthesis systems.

Table 4 Comparison of noise regions and bands

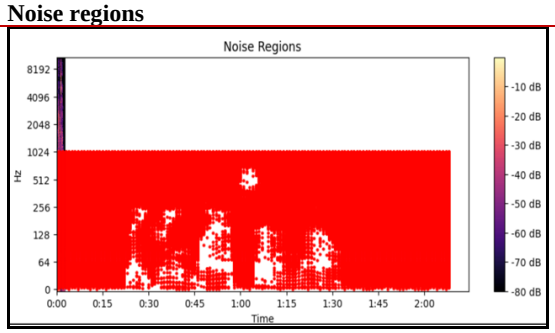
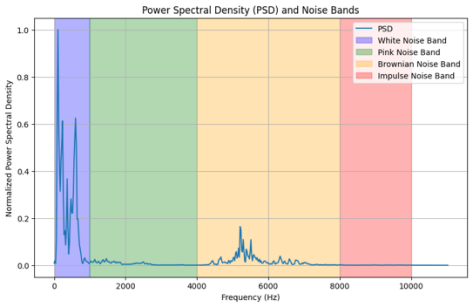
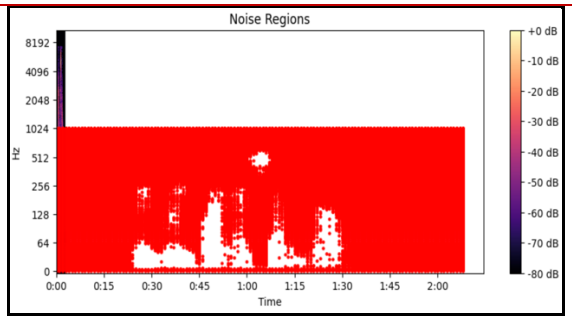
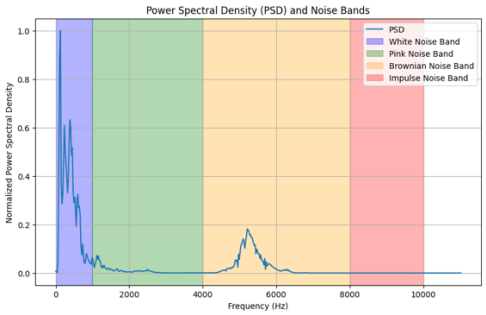
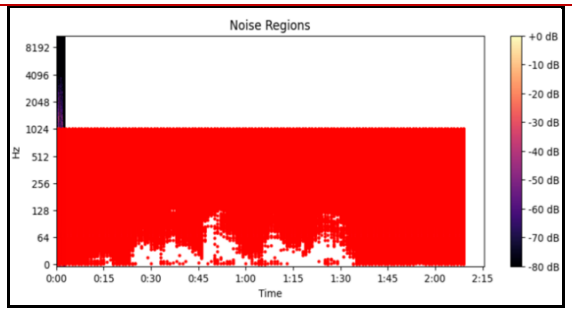
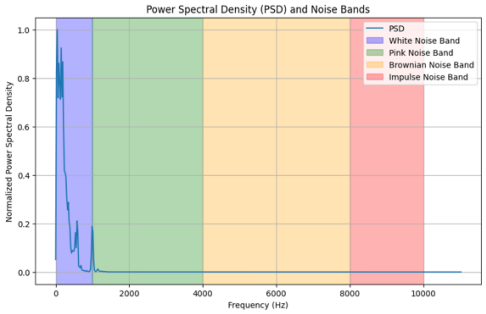
Type	Noise regions	Noise Bands
Original		
Lip2Voice		
Vid2Speech		

Table 5 Formant frequency, power, width and dissonance for original, Lip2Voice, Vid2Speech audio

Formant	Type	Mean formant frequency	Mean formant power	Mean formant width	Mean formant dissonance
0	Original	941.3	606.8	106.31	10.62
	Lip2Voice	1135.2	607.9	101.7	21.8
	Vid2Speech	744.0	512.0	107.5	4.9
1	Original	367.62	191.7	17.1	3.2
	Lip2Voice	556.3	250.7	22.4	7.3
	Vid2Speech	187.0	89.6	9.1	0.5
2	Original	139.7	64.7	2.3	0.7
	Lip2Voice	216.4	90.0	5.9	1.4
	Vid2Speech	32.0	12.0	0.9	0.04

Figure 7 compares mel-spectrograms and frequency domain waveforms of original and generated audios of Lip2Voice and Vid2Speech for lgbf7n silent video. The sentence spoken is “lay green by F 7 now”. For Lip2Voice, the inference layer generates audio on the corresponding silent video. For Vid2Speech, the pre-trained weights, as given in [26], were loaded to generate the spectrogram. The shape of the frequency domain waveform by Lip2Voice is very much like the original waveform compared to the one that Vid2Speech generates. This similarity confirms that Lip2Voice generates the correct voice type at the right time. Further, the generated mel spectrogram of Lip2Voice exhibits greater clarity compared to Vid2Speech, showcasing a reduction in audio noise.

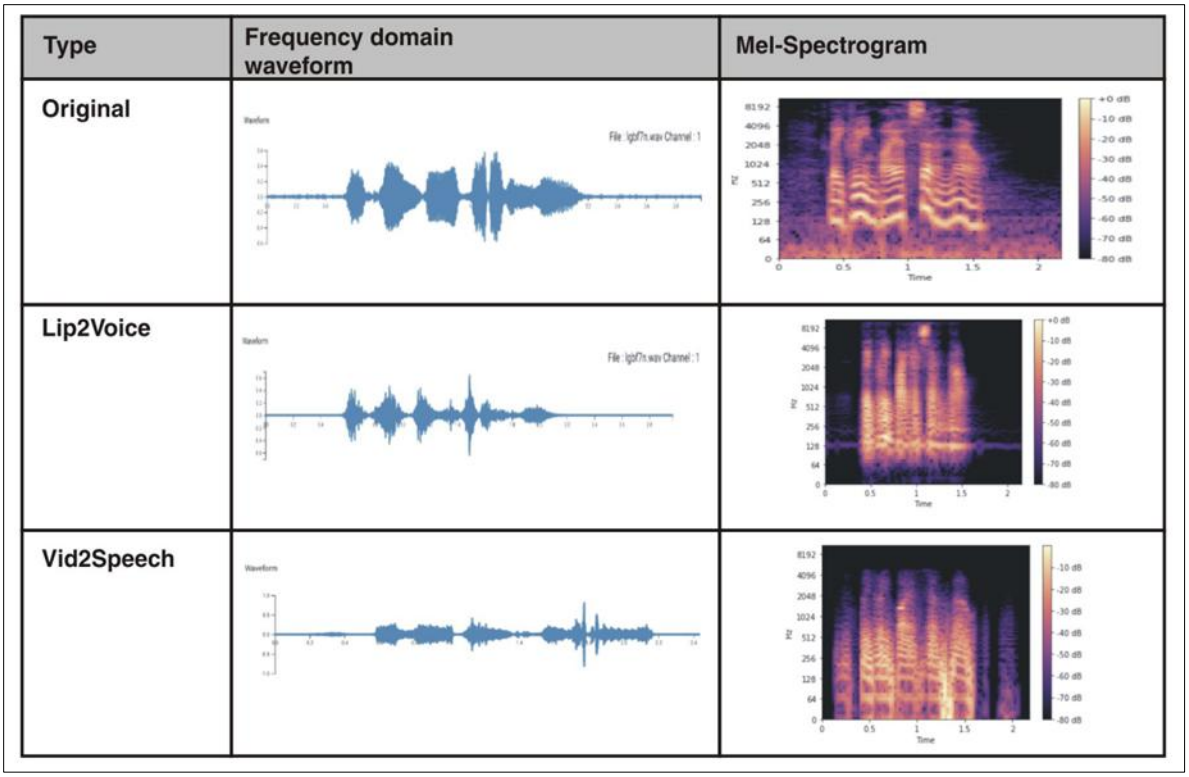


Figure 7 Comparison of audio spectrograms

5.Discussion

This research introduces "Lip2Voice", a sentence-level end-to-end VSR model designed to predict speech from silent video inputs. The key findings indicate that Lip2Voice significantly outperforms the existing Vid2Speech model, particularly in reducing the WER and enhancing the clarity of generated speech. The training loss for the Lip2Voice model is 36.6%, demonstrating a 10% decrease in WER compared to Vid2Speech. The dual speaker dataset primarily improves training loss, allowing the model to generalize better across different speakers.

Training VSR models using speakers with similar pitch levels further enhances the generalization and independence of the speaker. Results have shown that Lip2Voice achieved superior speech generation outcomes when handling speakers with similar pitch levels, such as those of the same gender. In contrast, mixed-gender datasets demonstrated comparatively lower speech quality. Additionally, alternately training on different speaker sets increases the model's robustness to varied input, promoting speaker independence.

The comparative study between Lip2Voice and Vid2Speech highlights critical differences, especially in WER and subjective human evaluations of audibility and pronunciation. The subjective evaluation with fifty-three participants provides deep insights into the model's real-world applicability. Participants have consistently rated Lip2Voice higher in clarity, both in noisy and noise-free conditions, which underscores its potential to integrate more effectively into daily life for individuals with speech impairments. Furthermore, Lip2Voice has exhibited greater clarity and reduced noise in the generated speech, as confirmed by detailed analysis of audio spectrograms and frequency domain waveforms. This is particularly evident when comparing the mel-spectrograms and resonance peaks of Lip2Voice with the original audio. Lip2Voice closely approximates actual spoken words, even detecting subtle nuances such as a minor spike near the 5kHz frequency in the PSD, a detail missed by Vid2Speech. The power distribution among the different frequency components of the original audio signal and Lip2Voice audio signals is nearly identical.

These findings have significant implications for the development of VSR systems, particularly for individuals with speech impairments. Lip2Voice's ability to generate accurate and intelligible speech from silent videos has the potential to improve

communication for this population. Lip2Voice represents a significant step forward in the field of VSR, offering improved accuracy and clarity over existing models. However, further research is needed to enhance its robustness and naturalness in speech synthesis to realize its full potential. Research is also needed to broaden its vocabulary and applicability in more varied and challenging environments. Future studies may also explore integrating advanced audio reconstruction techniques and expanding the dataset to improve the model's performance across diverse real-world scenarios.

5.1Limitations

Although Lip2Voice can produce more explicit speech, significant improvements are still required to achieve a human-like speech quality. Currently, the model uses the Griffin-Lim algorithm to reconstruct audio from mel spectrograms, resulting in a robotic-sounding voice. Further development in speech reconstruction is needed to attain a performance level that resembles human speech. Achieving such quality is essential for the practical application of the VSR system in the everyday lives of individuals with speech impairments. Enhancements include reducing discrepancies in mel-spectrograms and resolving issues with homophones. Recognizing that real-world conditions vary significantly from the controlled environments typically used in datasets is crucial.

The model is currently limited to recognizing only a few words, highlighting the need for a more diverse dataset. These datasets may include challenging scenarios such as corrupted or low-quality videos where facial features may not be noticeable. Moreover, the complexity of the model, with over 18 million trainable parameters, presents additional challenges for real-time processing. Optimizations are necessary to reduce the time it takes for the model to produce audio, which is currently every 3 seconds, as this delay could disrupt the flow of conversation and cause users to lose context in longer sentences.

A complete list of abbreviations is listed in *Appendix I.*

6.Conclusion and future work

This research proposed a new Lip2Voice model to generate speech on silent input videos. It describes the model selection approach in detail and explains the process of finalizing a model based on parameter optimization. This work provides a detailed chart of the number of filters, activation functions, and various neural network layers used in Lip2Voice. An inference algorithm is derived, which generates consecutive fixed-length speech for variable-length input videos. Comparisons performed with the existing Vid2Speech model based on metrics, including WER, Loss vs. Epoch, and subjective human evaluation, can validate the Lip2Voice model's effectiveness. The Lip2Voice model shows good performance with a training loss of 36.6%, marking a significant 10% improvement in word error rate (WER) over the Vid2Speech model. While Vid2Speech achieved a loss of 50.75% after training for 73 epochs on a single-speaker dataset, Lip2Voice outperformed Vid2Speech by reaching a loss of 36.6% after an extensive 507 epochs on a dual-speaker dataset. Additionally, in subjective human evaluations, Lip2Voice has an average percentage error of 13.84%, compared to 15.09% for Vid2Speech. This improvement in both loss and error rates highlights the effectiveness of Lip2Voice and its potential for applications in speech technology. The experimental results also indicate that generating speech for a continuous input stream is possible in real-time.

This work suggests future experiments to generalize Lip2Voice on a multi-speaker dataset. These investigations also show that researchers can use a generalized model for VSR in further studies. The dataset can be categorized based on different pitches to generate more accurate speech for a given set of people whose pitch level lies in a particular range. Moreover, different background noise and challenging scenarios can illustrate the robustness and versatility of the model. There can be more comparisons with other state-of-the-art models by locating the precise speaker and obtaining the appropriate set of trained weights, which are critical for reliable results. Additionally, investigations can be performed on the inference layer to provide usage of the VSR system as a real-time application.

Acknowledgment

None.

Conflicts of interest

The authors have no conflicts of interest to declare.

Data availability

The Grid audiovisual sentence corpus utilized in this study is publicly accessible and can be found at https://spandh.dcs.shef.ac.uk/gridcorpus/

Author's contribution statement

Aathira Pillai: Study conception, conceptualization, operationalize constructs, research design, and

investigation of challenges. writing – original draft, review, and editing. **Bhavana Mache:** Study conception, conceptualization, model data collection, model data preprocessing, and investigation of challenges. writing – original draft, review and editing, analysis and interpretation of results. **Supriya Kelkar:** Study conception, supervision, model data collection, result data processing. writing, reviewing, and editing, analysis and interpretation of results.

References

[1] Prajwal KR, Mukhopadhyay R, Namboodiri VP, Jawahar CV. Learning individual speaking styles for accurate lip to speech synthesis. In proceedings of the conference on computer vision and pattern recognition 2020 (pp. 13796-805). IEEE.

[2] Hassanat AB. Visual speech recognition. Speech and Language Technologies. 2011; 1:279-303.

[3] Son CJ, Senior A, Vinyals O, Zisserman A. Lip reading sentences in the wild. In proceedings of the conference on computer vision and pattern recognition 2017 (pp. 6447-56). IEEE.

[4] Fernandez-lopez A, Karaali A, Harte N, Sukno FM. Cogans for unsupervised visual speech adaptation to new speakers. In international conference on acoustics, speech and signal processing 2020 (pp. 6294-8). IEEE.

[5] Hao M, Mamut M, Yadikar N, Aysa A, Ubul K. A survey of research on lipreading technology. IEEE Access. 2020; 8:204518-44.

[6] Ma P, Wang Y, Shen J, Petridis S, Pantic M. Lip-reading with densely connected temporal convolutional networks. In proceedings of the winter conference on applications of computer vision 2021 (pp. 2857-66). IEEE.

[7] Afouras T, Chung JS, Zisserman A. Asr is all you need: cross-modal distillation for lip reading. In international conference on acoustics, speech and signal processing 2020 (pp. 2143-7). IEEE.

[8] Ephrat A, Halperin T, Peleg S. Improved speech reconstruction from silent video. In proceedings of the international conference on computer vision workshops 2017 (pp. 455-62). IEEE.

[9] Gao W, Hashemi-sakhtsari A, Mcdonnell MD. End-to-end phoneme recognition using models from semantic image segmentation. In international joint conference on neural networks 2020 (pp. 1-7). IEEE.

[10] Sarhan AM, Elshennawy NM, Ibrahim DM. HLR-net: a hybrid lip-reading model based on deep convolutional neural networks. Computers, Materials and Continua. 2021; 68(2):1531-49.

[11] Guan C, Wang S, Liew AW. Lip image segmentation based on a fuzzy convolutional neural network. IEEE Transactions on Fuzzy Systems. 2019; 28(7):1242-51.

[12] Tsourounis D, Kastaniotis D, Fotopoulos S. Lip reading by alternating between spatiotemporal and spatial convolutions. Journal of Imaging. 2021; 7(5):1-17.

[13] Tao F, Busso C. End-to-end audiovisual speech recognition system with multitask learning. IEEE Transactions on Multimedia. 2020; 23:1-11.

[14] Xu K, Li D, Cassimatis N, Wang X. LCANet: end-to-end lipreading with cascaded attention-CTC. In 13th international conference on automatic face & gesture recognition 2018 (pp. 548-55). IEEE.

[15] Burchi M, Timofte R. Audio-visual efficient conformer for robust speech recognition. In proceedings of the winter conference on applications of computer vision 2023 (pp. 2258-67). IEEE.

[16] Serdyuk D, Braga O, Siohan O. Audio-visual speech recognition is worth $32\times 32\times 8$ voxels. In automatic speech recognition and understanding workshop 2021 (pp. 796-802). IEEE.

[17] Shilaskar S, Iramani H. CTC-CNN-bidirectional LSTM based lip reading system. In international conference on emerging smart computing and informatics 2024 (pp. 1-6). IEEE.

[18] Kuriakose LK, Sinciya PO, Joseph MR, Namita R, Nabi S, Lone TA. Dip into: a novel method for visual speech recognition using deep learning. In annual international conference on emerging research areas: international conference on intelligent systems 2023 (pp. 1-6). IEEE.

[19] Burchi M, Puvvada KC, Balam J, Ginsburg B, Timofte R. Multilingual audio-visual speech recognition with hybrid CTC/RNN-T fast conformer. In international conference on acoustics, speech and signal processing 2024 (pp. 10211-5). IEEE.

[20] Liu X, Lakomkin E, Vougioukas K, Ma P, Chen H, Xie R, et al. SynthVSR: scaling up visual speech recognition with synthetic supervision. In proceedings of the conference on computer vision and pattern recognition 2023 (pp. 18806-15). IEEE.

[21] Zhang JX, Ling ZH, Liu LJ, Jiang Y, Dai LR. Sequence-to-sequence acoustic modeling for voice conversion. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2019; 27(3):631-44.

[22] Zhang X, Cheng F, Wang S. Spatio-temporal fusion based convolutional sequence learning for lip reading. In proceedings of the international conference on computer vision 2019 (pp. 713-22). IEEE.

[23] Afouras T, Chung JS, Senior A, Vinyals O, Zisserman A. Deep audio-visual speech recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2018; 44(12):8717-27.

[24] Sterpu G, Saam C, Harte N. Attention-based audio-visual fusion for robust automatic speech recognition. In proceedings of the 20th international conference on multimodal interaction 2018 (pp. 111-5). ACM.

[25] Luo M, Yang S, Shan S, Chen X. Pseudo-convolutional policy gradient for sequence-to-sequence lip-reading. In 15th international conference on automatic face and gesture recognition 2020 (pp. 273-80). IEEE.

[26] Ephrat A, Peleg S. Vid2speech: speech reconstruction from silent video. In international conference on

acoustics, speech and signal processing 2017 (pp. 5095-9). IEEE.
[27] Arthur FV, Csapó TG. Towards a practical lip-to-speech conversion system using deep neural networks and mobile application frontend. In the international conference on artificial intelligence and computer vision 2021 (pp. 441-50). Cham: Springer International Publishing.
[28] Akbari H, Arora H, Cao L, Mesgarani N. Lip2audspec: speech reconstruction from silent lip movements video. In international conference on acoustics, speech and signal processing 2018 (pp. 2516-20). IEEE.
[29] Shandiz AH, Tóth L, Gosztolya G, Markó A, Csapó TG. Improving neural silent speech interface models by adversarial training. In the international conference on artificial intelligence and computer vision 2021 (pp. 430-40). Cham: Springer International Publishing.
[30] Prajwal KR, Afouras T, Zisserman A. Sub-word level lip reading with visual attention. In proceedings of the conference on computer vision and pattern recognition 2022 (pp. 5162-72). IEEE.
[31] Ivanko D, Ryumin D, Markitantov M. End-to-end visual speech recognition for human-robot interaction. In AIP conference proceedings 2024, AIP Publishing.
[32] Bhaskar S, Thasleema TM. LSTM model for visual speech recognition through facial expressions. Multimedia Tools and Applications. 2023; 82(4):5455-72.
[33] Ajitha D, Dutta D, Saha F, Giri P, Kant R. AI LipReader-transcribing speech from lip movements. In international conference on emerging smart computing and informatics 2024 (pp. 1-6). IEEE.
[34] Adeel A, Gogate M, Hussain A, Whitmer WM. Lip-reading driven deep learning approach for speech enhancement. IEEE Transactions on Emerging Topics in Computational Intelligence. 2019; 5(3):481-90.
[35] Hou JC, Wang SS, Lai YH, Tsao Y, Chang HW, Wang HM. Audio-visual speech enhancement using multimodal deep convolutional neural networks. IEEE Transactions on Emerging Topics in Computational Intelligence. 2018; 2(2):117-28.
[36] Thimmaraja YG, Nagaraja BG, Jayanna HS. Speech enhancement and encoding by combining SS-VAD and LPC. International Journal of Speech Technology. 2021; 24(1):165-72.
[37] Sadeghi M, Leglaive S, Alameda-pineda X, Girin L, Horaud R. Audio-visual speech enhancement using conditional variational auto-encoders. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2020; 28:1788-800.
[38] Patilkulkarni S. Visual speech recognition for small scale dataset using VGG16 convolution neural network. Multimedia Tools and Applications. 2021; 80(19):28941-52.
[39] Xiao J, Yang S, Zhang Y, Shan S, Chen X. Deformation flow based two-stream network for lip reading. In 15th international conference on automatic face and gesture recognition 2020 (pp. 364-70). IEEE.
[40] Lu Y, Li H. Automatic lip-reading system based on deep convolutional neural network and attention-based long short-term memory. Applied Sciences. 2019; 9(8):1-12.
[41] Thangthai K, Harvey RW. Building large-vocabulary speaker-independent lipreading systems. In interspeech 2018 (pp. 2648-52).
[42] Liu L, Feng G, Beautemps D, Zhang XP. Re-synchronization using the hand preceding model for multi-modal fusion in automatic continuous cued speech recognition. IEEE Transactions on Multimedia. 2020; 23:292-305.
[43] Cuervo S, Grabias M, Chorowski J, Ciesielski G, Łańcucki A, Rychlikowski P, et al. Contrastive prediction strategies for unsupervised segmentation and categorization of phonemes and words. In international conference on acoustics, speech and signal processing 2022 (pp. 3189-93). IEEE.
[44] Lin B, Wang L. Learning acoustic frame labeling for phoneme segmentation with regularized attention mechanism. In international conference on acoustics, speech and signal processing 2022 (pp. 7882-6). IEEE.
[45] Wang Y. Research on automatic generation algorithm of phoneme conversion learning corpus based on KNN algorithm. In international conference on image processing and computer applications 2023 (pp. 1624-8). IEEE.
[46] Cooke M, Barker J, Cunningham S, Shao X. An audio-visual corpus for speech perception and automatic speech recognition. The Journal of the Acoustical Society of America. 2006; 120(5):2421-4.
[47] Wang X, Mi J, Li B, Zhao Y, Meng J. CATNet: cross-modal fusion for audio–visual speech recognition. Pattern Recognition Letters. 2024; 178:216-22.
[48] Yeo JH, Kim M, Choi J, Kim DH, Ro YM. AKVSR: audio knowledge empowered visual speech recognition by compressing audio knowledge of a pretrained model. IEEE Transactions on Multimedia. 2024; 26:6462-74.
[49] Liu ZT, Rehman A, Wu M, Cao WH, Hao M. Speech emotion recognition based on formant characteristics feature extraction and phoneme type convergence. Information Sciences. 2021; 563:309-25.

Aathira Pillai earned her BTech in Computer Science from MKSSS's Cummins College of Engineering for Women in Pune, Maharashtra, India, in 2021. Following this, she gained valuable experience as a Software Engineer at Societe Generale for two years. Presently, she is pursuing a Master of Science in Artificial Intelligence at Khoury College of Computer Sciences, Northeastern University, Boston, MA, USA. Her research interests include Computer Vision and Reinforcement Learning, including Robotics and Agent-Based Systems.
Email: aathira.pillai@cumminscollege.in

Bhavana Mache received her bachelor's degree from MKSSS's Cummins College of Engineering for Women, Pune, Maharashtra, India, in 2021. She is currently working as a Software Engineer at Intuit India. Her research interests include Computer Vision and Machine Learning.
Email: bhavana.mache@cumminscollege.in

Supriya Kelkar completed her bachelor's degree from Gogte Institute of Technology, Karnataka University, India. She worked as a Research and Development Engineer in the chemical industry for a few years. She obtained her Master's degree from the College of Engineering Pune, Pune University, India. She completed her Ph.D. at the Institute of Engineering and Technology, Devi Ahilya Vishwavidyalaya, Indore, India. Since 1994, she has been working at the Department of Computer Engineering, MKSSS's Cummins College of Engineering for Women, Pune, India. Her research interests include Machine Learning, IoT, Distributed and Automotive Networking. She is a Senior Member of IEEE.
Email: supriya.kelkar@cumminscollege.in

Appendix I

S. No.	Abbreviation	Description
1	AE	Auto-Encoder
2	ALSOS	Alternating Spatio-Temporal and Spatial Convolutions
3	AKVSR	Audio Knowledge-Based Visual Speech Recognition Framework
4	ASR	Automated Speech Recognition
5	ATM	Automated Teller Machine
6	AV	Audio Visual
7	AVDCNN	Audio-Visual Deep CNN
8	AVSR	Audio-Visual Speech Recognition
9	CCTV	Closed-Circuit Television
10	CNN	Convolutional Neural Network
11	CTC	Connectionist Temporal Classification
12	DFN	Deep Functional Network
13	ESTOI	Extended Short-Time Objective Intelligibility
14	EVWF	Wiener Filter
15	FFMPEG	Fast Forward Moving Picture Experts Group
16	FV	Feature Vector
17	GAN	Generative Adversarial Network
18	GB	Gigabyte
19	GMM	Gaussian Mixture Model
20	GPU	Graphics Processing Unit
21	HLR-Net	Hybrid Lip-reading Network
22	KNN	K-Nearest Neighbour
23	LPC	Linear Predictive Coding
24	LRS	Lip Reading Sentences
25	LRS2	Lip Reading Sentences 2
26	LRS3	Lip Reading Sentences 3
27	LRW	Lip Reading in the Wild
28	LSTM	Long Short-Term Memory
29	MSE	Mean Squared Error
30	PCPG	Pseudo-Convolutional Policy Gradient
31	PESQ	Perceptual Evaluation of Speech Quality
32	PIN	Personal Identification Number
33	PSD	Power Spectral Density
34	RAM	Radom-Access Memory
35	ReLu	Rectified Linear Unit
36	ResNet	Residual Network
37	RNN	Recurrent Neural Network
38	seq2seq	Sequence-to-Sequence
39	SS-VAD	Spectral Subtraction Voice Activity Detection
40	STFT	Short-time Fourier Transform
41	TCD-TIMIT	Trinity College Dublin - Texas Instruments/Massachusetts Institute of Technology
42	TTS	Text-to-Speech
43	VAEs	Variational Auto-Encoders
44	VGG	Visual Geometry Group
45	VSR	Visual Speech Recognition
46	WER	Word Error Rate
47	1D	One Dimensional
48	3D	Three Dimensional
49	3DCNN	Three-Dimensional Convolutional Neural Networks
50	3DResNet	Three-Dimensional Residual Network