A Robust and Integrated Speech Recognition Tool for Dysarthria Patients Using Lip Movement Recognition

May Altulyan
Department of Computer Engineering, College of Computer Science and Engineering, Prince Sattam Bin Abdulaziz University, Al Kharj, Saudi Arabia
m.altulayan@psau.edu.sa (corresponding author)

Received: 27 January 2026 | Revised: 7 March 2026, 4 April 2026, 11 April 2026, and 18 April 2026 | Accepted: 21 April 2026

 DOI: https://doi.org/10.48084/etasr.17799

ABSTRACT

Recent advances in human-computer interaction have led to marked innovations in the development of computer-aided tools for the disabled and survivors of neurological diseases. Dysarthria is a neurological disorder that affects muscles, impacting speech articulation and clarity. Brain tumors, Cerebral palsy, Parkinson's disease, and head injuries also affect the movement of the tongue, leading to unclear speech. Dysarthria speech is intelligible and poses an arduous challenge for voice recognition systems developed based on speech signal processing. This study presents a model based on lip movement recognition and transfer learning to develop a robust speech recognition tool for patients with dysarthria. Lip recognition is based on a 3D Convolutional Neural Network (CNN) and a Bidirectional Long-Short-Term Memory (BiLSTM) neural network for lip movement detection and speech recognition. The proposed speech recognition model was trained on the GRID sentence Corpus dataset. Dysarthria speech can be recognized using transfer learning. The speech data is converted to text by the lip recognition model, and the text data is analyzed by transformers for word prediction and grammar correction. The novelty of the proposed framework is that it not only recognizes speech data but also improves the text recognized with a sequence-to-sequence T5 transformer model to improve speech recognition. The lip movement recognition model had an accuracy of 98.29% and a precision of 99.58%. The accuracy of the transformer grammar correction model was 78% due to limited training. The proposed integrated model is a novel idea that uses lip movement recognition rather than speech data for speech recognition, demonstrating high performance.

Keywords-dysarthria; lip movement recognition; transformers; speech processing; human computer interaction; BiLSTM

I. INTRODUCTION

Dysarthria, a neurological speech disorder, affects the process of controlling and articulating speech sounds, leading to unclear speech and, in some cases, a substantial decrease in speech intelligibility [1]. Patients report stigmatization, changes in self-identity, and social and emotional disturbances due to post-stroke activities. Dysarthria patients show symptoms of speech with reduced intensity, uncontrolled pitch variations, and inconsistent speech rate [2]. The primary processes involved in speech production are respiration, phonation, resonance, articulation, and prosody [3]. Muscular dysfunction adversely affects speech, causing problems related to audibility and intelligibility, thereby affecting communication skills. Neoplastic diseases, such as central and peripheral nervous system tumors, can also lead to dysarthria [4]. Infectious diseases, such as Creutzfeldt-Jakob, can also lead to speech disorders. The presence of primary and metastatic brain and laryngeal tumors can lead to a lack of clear speech. Congenital birth conditions such as cleft lip or palate in babies are also causes for dysarthria [5]. In a recent study on 295 children suffering from pediatric neuromuscular diseases, the prevalence of dysarthria was found to be 31.5% [6]. Motor speech disorders are found in other neurological disease conditions such as Parkinson's Disease (PD), Progressive Supranuclear Palsy (PSP), Huntington's Disease (HD), and Amyotrophic Lateral Sclerosis (ALS) [7].

Speech intelligibility is a construct that depends on (a) a speaker that produces an acoustic signal within, e.g., conversational speech, and (b) a listener who receives the signal and interprets it; the success of the interpretation is a direct function of the intelligibility [8]. The objective of every speech therapist is to provide improved speech intelligibility for patients through therapy and treatment. The patient is clinically assessed by measuring speech intelligibility, which is the conventional method to monitor the status of a dysarthric patient and track treatment progress. Speech intelligibility assessment is performed with tests such as the Assessment of Intelligibility in Dysarthric Speakers (AIDS), the Sentence

Intelligibility Test (SIT), and the Word Intelligibility Test (WIT). The most common approaches for therapy include Lee Silverman Voice Treatment (LSVT) to improve loudness and intelligibility [9] and Pitch Limiting Voice Treatment (PLVT) [10]. When therapy and surgery fail to help patients, computer-aided tools are adopted to improve speech intelligibility. In this landscape, augmentative and alternative communication is a method of adopting high-tech devices such as voice synthesizers and speech-generating devices to help patients with dysarthria [11].

II. LITERATURE REVIEW

Although AI plays a crucial role in computer-aided detection and diagnosis, current research focuses more on therapy, treatment, and management of disease diagnosis. In [12], a speech-based analysis approach was based on acoustic features for early AD and PD diagnosis. Research in dysarthria is broadly categorized as dysarthria diagnosis using deep learning diagnostic models, speech enhancement with automated AI tools, AI-based tools for therapy, and voice recognition systems customized for dysarthria patients. Novel AI-based dysarthria classifiers integrate features from Mel-Frequency Cepstral Coefficients (MFCC) with Temporal Discriminative Bottleneck (TDBN) features. A Temporal Kolmogorov-Arnold Network (TKAN) layer in the dysarthria classifier model can extract temporal pattern differences in the speech of patients [13]. In [14], dysarthric speech was distinguished from normal speech with time-frequency image representations using a Deep CNN (DCNN). The classification accuracies for spectrogram, cepstrogram, mel-scalogram, and cochleagram were 98.39%, 99.24%, 90.79%, and 99.26%, respectively. In [15], an automated model for dysarthria detection and severity level assessment used the time-domain waveform as input, achieving accuracies of 88.5% and 91.80% for dysarthria detection and severity assessment on the TORGO dataset. In [16], a hybrid cross-attentive CNN-BiLSTM-Transformer network was adopted for dysarthria severity classification, achieving accuracies of 98.74% and 99.86% for binary classification on the TORGO and UA speech datasets [16]. Table I illustrates state-of-the-art methods for dysarthria classification.

TABLE I. . STATE-OF-THE-ART METHODS FOR DYSARTHRIA CLASSIFICATION

Study	Dataset	Input	Features	Method	Performance
[13]	TORGO	Audio input	MFCC and TDBN	PCA and DNN	Accuracy: 98%
[14]	TORGO	speech features	DCNN	Feature embedding and DNN	Cochleagram: 99.26
[15]	UA speech	Waveform input	Amor, Morse, and Bump wavelets	Wavelet transform layered CNN	Accuracy: 93.7%
[16]	TORGO	Audio input	Scalogram Images	CNN-BiLSTM-Transformer hybrid model	Accuracy: 98.74%
	UA speech				Accuracy: 99.86%

AI-supported Augmentative and Alternative Communication (AAC) technologies pave the way to assist adults and children with phonological or speech disorders. In [17], the generated synthetic speech effectively reproduced the disordered speech and resembled real speech. The model was tested by a licensed Speech-Language Pathologist (SLP) who misclassified around 30% of the synthetic speech as real speech. In [18], a customized AASC consisted of a collar microphone to record speech data, which was later processed by a transfer learning-based dysarthric speech recognition system. The recognized text was further synthesized using an HMM-based text-to-speech system and played using the speaker. In [19], a Kullback-Leibler divergence-based Hidden Markov Model (KL-HMM) was adopted to build a deep neural network-based acoustic model, achieving better performance compared to other methods using DNNs. Table II summarizes existing studies on voice recognition for dysarthria patients.

TABLE II. COMPARISON OF STATE-OF-THE-ART METHODS ON VOICE RECOGNITION FOR DYSARTHRIA PATIENTS

Study	Features	Method	Advantage	Performance
[17]	Synthetic voice generation	Voice cloning	Mitigate data scarcity and improve therapy	Identified dysarthria in 95% of the samples.
[18]	Text-to-speech synthesis system	HMM-based text-to-speech synthesis	Portable, affordable, and personalized speech aid	Speech delivery rate of roughly 4.4 s.
[19]	Automatic speech recognition systems	KL-HMM	Preserves speaker-specific information	Outperformed a CNN-based speaker-adapted system

In [20], DNN-HMM models with dropout and sequence discrimination were developed for speaker-normalized cepstral features. In [21], Hurst-based mode selection (EMDH) was adopted with a CNN to improve speech recognition for dysarthric patients. In [22], a Cycle-consistent Generative Adversarial Network (Cycle GAN) was studied under a simulated environment for dysarthria speech, analyzing its effect on dysarthria. Automatic Speech Recognition (ASR) for processing dysarthric speech is difficult due to limited data available, mitigating the speech nuances and the distorted spectral features [23]. In [24], a two-stage visual automatic speech recognition model used a Vision Transformer (ViT) with a Connectionist Temporal Classification (CTC) head to predict the sequence of phonemes from visuals, serving as input to a Large Language Model (LLM). In [25], a lip recognition model combined a lightweight CNN, named ShuffleNet, pre-trained on ImageNet, and an attention-GRU. In [26], a spatiotemporal CNN was combined with a BiLSTM and a CTC loss function to recognize individual words, achieving an accuracy of 96.4%.

The studies discussed above were based on public speech datasets. Annotated public datasets for dysarthric speech are very few, which poses the main hurdle for audio-based recognition systems. Audio data is often highly variable with abnormal pitch, slow speech rate, and inconsistent articulation. This makes it difficult for the model to generalize and achieve high performance with real-time data. Audio-based dysarthria

voice recognition systems exhibit poor performance for speakers with severe dysarthria and poor speech intelligibility. Noisy data recorded in real environments is another reason for the poor performance of audio-based models. The collection of large volumes of data from dysarthric speakers is not only difficult but sometimes an unrealistic task. Many studies used simulated data for developing AI models. These issues motivated the development of a model based on videos of lip movement of dysarthria patients rather than speech data. This study adopts transfer learning, in which the model is trained on the GRID audiovisual sentence corpus dataset and tested on videos of specific dysarthria patients. The proposed model includes auto-completion and grammar correction for improving the text detected by the lip recognition model.

III. METHODOLOGY

This section discusses the methodology adopted for the proposed AI-based speech improvement tool for patients with dysarthria. Speech from a dysarthria patient is recognized using two sub-modules. The patient's lip movement is preprocessed with the video preprocessing module, followed by the lip recognition module. The second module is a transformer-based text prediction and grammar correction module that converts speech directly to text. Figure 1 shows the proposed model.

A. Lip Movement-Based Speech Recognition Model

The lip movement recognition model is built with a CNN and a BiLSTM, combining spatial features extraction and temporal modeling. This model includes a preprocessing module and a deep learning module for lip recognition. The model is robust to speech variations and learns the complex spatiotemporal patterns from the dataset effectively. Figure 2 exhibits a block diagram for lip movement recognition.

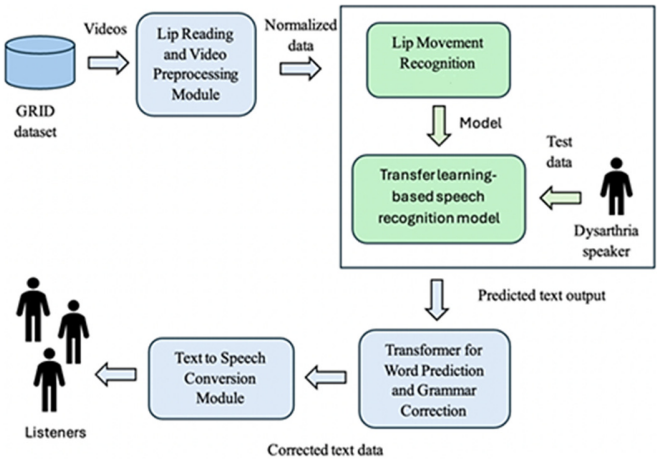

Fig. 1. Proposed model for an AI-based speech improvement tool for dysarthria patients.

1) Dataset

This study used the GRID audiovisual sentence corpus dataset, which is publicly available and commonly used, including a multitasker audio-visual sentence corpus consisting of high-quality audio and video (facial) recordings of 1000 sentences spoken by 34 speakers (male: 18, female: 16) [27]. The video files are categorized as normal quality videos with 360×288 pixel resolution and an average bit rate of ~1 kbit/s, and high quality ones with 720×576 pixel resolution and an average bit rate of ~6 kbit/s. The train and test data include 45 videos from the GRID audiovisual sentence corpus dataset. Audio, video, and related information, including word transcriptions, are available separately for each speaker. Audio files were scaled to have an absolute maximum amplitude value of 1 and downsampled to 25 kHz.

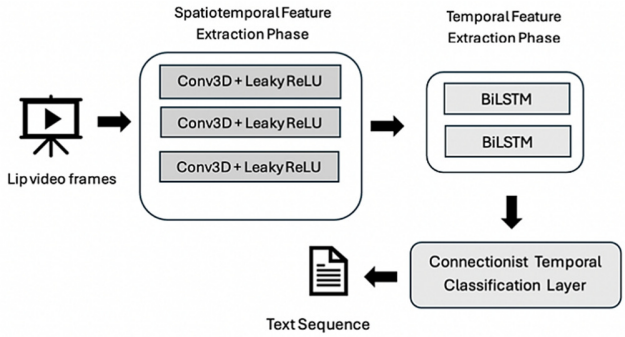

Fig. 2. Block diagram of lip movement recognition.

2) Algorithm for Lip Movement Recognition

The lip-reading module captures the lip movement of the patient as a video. The videos are converted into normalized data. Each video is segmented into a set of image frames. The frame count fr_{cnt} is computed. The frames are read from the video and converted to a gray-scale image fg_i. The mouth part of the images fg_i is cropped using a window size of [190:236, 80:220]. Each image is appended to a list. The mean and standard deviation are computed to normalize the pixel values. The data in the list is returned as normalized data. The alignment labels from the dataset are converted to a numeric sequence of tokens with the $find\_tokens\ (align\_path)$ method.

```
Algorithm 1: Lip_recognition()
1. The videos are converted into
   normalized data with the
   normalize_video (path) method
2. Convert alignment labels from the
   dataset to a numeric sequence of tokens
   with the find_tokens(align_path)
   method.
3. Prepare the data for the deep learning
   model.
4. Design the model
5. Train the model
6. Predict the data with the model.
7. Evaluate the model.
```

Each word in the image is formatted as the start time, end time, and the token. The tokens are determined and saved as a list with spaces in between. From the lookup table, the integer index of each token is found, and the numeric sequence of each token in the word is returned. The processed data is now fed into the proposed deep learning model.

Algorithm 2: normalize_video()
1. Load the video using the VideoCapture() method from open CV.
2. Compute frame count fr_{cnt}.
3. for (i=1; i <= fr_{cnt}; ++i)
3.1 Read the video frame by frame using the read method in f_i.
3.2 Convert f_i from RGB to Grayscale and save in fg_i
3.3 Crop the mouth part from each frame fg_i with a window sized $w_s = [190{:}236, 80{:}220]$, and save in f_{ci} of size $m \times n$.
3.4 Append f_{ci} in the list L_f.
end for
4. Create a tensor stack of all frames T_f from L_f with size T_{sz} $T_{sz} = fr_{cnt} \times m \times n$ where $m = 46$ and $n = 140$.
5. Compute the mean $\mu = \frac{\sum_{i=1}^{fr_{cnt}} \sum_{j=1}^{m} \sum_{k=1}^{n} X_{i,j,k}}{T_{sz}}$
6. Compute the standard deviation $\sigma = \sqrt{\frac{\sum_{i=1}^{fr_{cnt}} \sum_{j=1}^{m} \sum_{k=1}^{n} (X_{i,j,k} - \mu)^2}{T_{sz}}}$
7. Normalize the pixel values with $\hat{x} = \frac{x-\mu}{\sigma+\varepsilon}$, where $\varepsilon = 10^{-8}$
8. Return normalized data.

Algorithm 3: find_tokens (file_path)
1. Read the alignment files in the following format.
2. $F = \{l_1, l_2, l_3, \ldots, l_n\}$ where l_i is a line $l_i = (st_i, et_i, tk_i)$, where st_i is start time, et_i is the end time and tk_i is the token.
3. while (l_i in F)
$l_{si} = \text{split}(l_i)$
if$(l_{si}[2] != ' ')$
$tk_i = append(\{tk_i\}, l_{si})$
4. Insert a space between every token in tk_i and save as T_s, where $T_s = \{ l_{s1}, l_{s2}, \ldots, l_{si} \ldots., l_{sn} \}$, $N_s = \{f(l_{s1}), f(l_{s2}), \ldots. f(l_{si}), \ldots.. f(l_{sn})\}$ where N_s is the numeric sequence of T_s, f: integer index of tokens in T_s from the look up table, $f: C \rightarrow \mathbb{N}$; $N_s = f(l_{s1})$
5. Return the numeric sequence N_s.

These algorithms list the steps followed for lip movement recognition. The limitations of this model are the need for a large dataset and precise extraction of the mouth region. The model is computationally expensive as it includes two deep learning models. Figure 3 shows a cropped image for lip movement recognition.

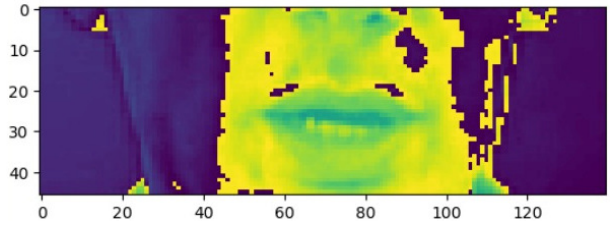

Fig. 3. Cropped lip image for lip movement recognition.

B. Transfer Learning-Based Speech Recognition Model

Noisy, inconsistent, or redundant data adversely affect the training and testing accuracy of the model [28]. Data scarcity is a critical issue faced by AI models in solving real-time medical diagnosis problems, as there is a lack of natural, unadulterated information based on direct human interactions [29]. Healthcare disciplines demand high precision and reliability, but ethical and unaddressed technical issues result in the unavailability of medical data. In this landscape, transfer learning is a potential solution to address the issue of data scarcity [30]. Transfer learning is deployed in scenarios where there is a lack of large datasets for effective training, leveraging models trained on other large and well-balanced datasets. The proposed lip movement recognition model is built using the GRID audiovisual sentence corpus dataset. Since there is no public dataset available for the lip movement of patients with dysarthria, this study adopted transfer learning to implement the proposed speech recognition model.

C. Word Prediction and Grammar Correction Module

The text output accuracy of the dysarthria speech recognition model can be improved with a transformer-based model concatenated to the lip movement recognition model. With the advent of LLMs [31], transformers are widely used to learn complex patterns from text data, paving the way to advanced tasks, such as sentence prediction and grammar corrections, related to Natural Language Processing (NLP). Latest LLMs also handle multi-modal inputs such as images and videos to wider applications, such as speech recognition, image classification, and code generation [32]. A transformer is a complex, high-dimensional language model that captures long-term structures in sequence data [33]. This study used the text-to-text T5 transformer model, which uses an encoder to interpret the input and a decoder to generate the output and, hence, can be referred to as a sequence-to-sequence model. The T5 model uses multi-head self-attention to prioritize the words from the input and previous output. The unique relative position embedding supports the model to deal easily with sequence positioning. The cross-attention layer replaces the self-attention mechanism, thereby reducing the number of parameters and memory usage [34].

The text data identified from the dysarthric speech is fed into the tokenizer. The tokenizer converts the data into vectors using an embedding technique, then fed as input to the topmost layer of the encoder stack. Since the position of the word in a sentence plays an important role in word prediction, relative positional encoding is the second layer in this model. The positional encoded vector is now given as input to the self-attention layer.

Self-attention works by relating and attending to every word in the sentence. It calculates the weighted representation of words in sentences considering their contextual priority. The focus is on tracking the dependencies and the relationship of the words in the sentence. The self-attention layer enables the decoder to convert the numeric vector of each word to a $< Q, K, V >$ vector, derived from:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_K}}\right)V \quad (1)$$

where Q is the Query vector, K is the Key vector, V is the Value vector, and d_K represents the dimensions of the Key vector. The attention output matrix is a context-aware input representation that captures the relationships among different parts of the input sentences [35]. The priority given to each word in other parts of the sentence is also calculated. The Softmax function is applied for normalization [36]. This procedure is repeated for all words in the sentence. A stack of multi-head self-attention cells with customized parameters captures complex relationships between the words in different locations in a sentence.

The next sublayer is a feed-forward neural network that processes the sentence token-wise. Normalization and dropout layers are critical for deep transformers. Word embedding, positional encoding, self-attention, and residual connections collectively train the model to understand the complex relationships between the words in a sentence. The decoder in the transformer is a stack of identical layers, each having self-attention, encoder-decoder, and feed-forward sublayers. Residual connections and layer normalizations follow each of these layers. Figure 4 shows the high-level architecture of the word prediction and grammar correction model.

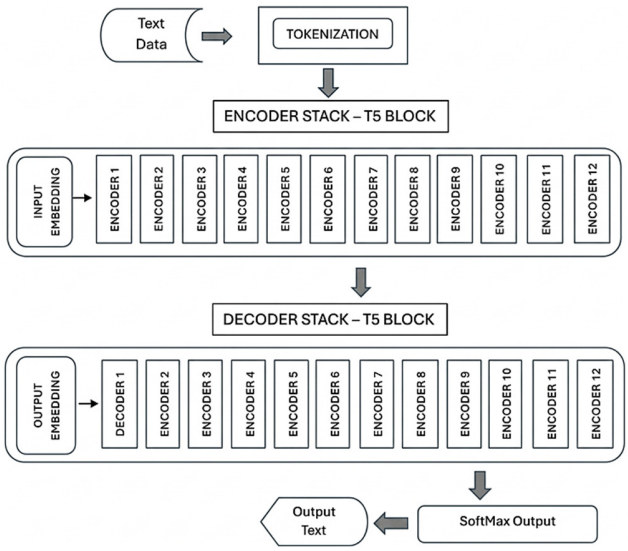

Fig. 4. High-level architecture of the transformer for word prediction and grammar correction.

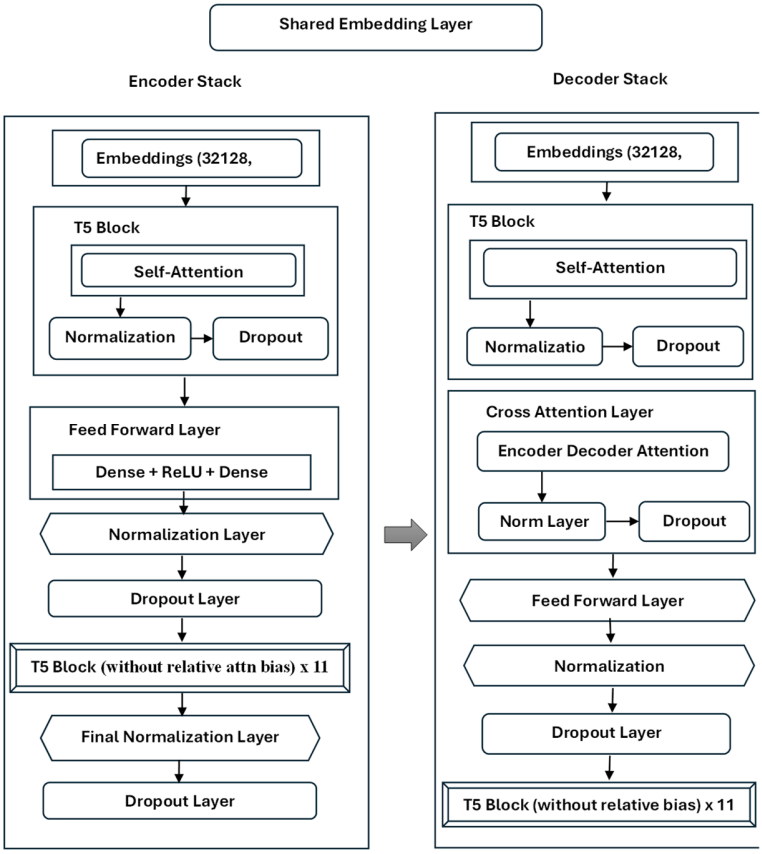

Fig. 5. Low-level architecture of the transformer for word prediction and grammar correction.

T5 leverages the transfer learning concept to improve model efficiency by training on large, versatile datasets. The model can generate the missing words and make grammatical corrections to the text [37]. The text data is initially tokenized, converted into smaller word units with a 32 k vocabulary. Each token is mapped to a dense vector. The shared encoder input embedding is of 32128 vocabulary size with a model dimension of 768. Each token is a 768-dimensional vector. Hence, the total parameters are 24.7 M. Unlike BERT, which uses an embedded matrix for input, T5 uses a shared embedding matrix that improves consistency. Each T5 block has a self-attention layer, a normalization layer, and a dropout layer. Figure 5 shows the low-level architecture of the transformer model for word prediction and grammar correction.

A feed-forward layer, with Dense and ReLU, follows. Residual connections prevent forgetting and stabilize the training process. The Normalization layer is key for deep transformers and stabilizes the model. This creates a contextual understanding using self-attention across the layers. The decoder stack processes the current information and predicts the best next token according to the grammar.

IV. EXPERIMENTAL SETUP

The lip detection model is built with a 3DCNN and BiLSTM. The input is videos that are processed as frames along a time sequence. The model includes four conv3D units and a BiLSTM layer. The model learns both the spatial details and the temporal motion of the lip movement. The 3DCNN kernels identify the phoneme articulation patterns of the lip movements, performing better compared to the 2DCNNs and RNNs. The model is built with three Conv3D units. The first Conv3D layer has 64 filters with an input shape of 75×46×140×1 with Leaky ReLU as the activation function. Leaky ReLU is preferable compared to normal ReLU, as it allows for having a negative slope value, allowing the gradient to flow, offering smoother learning and better convergence. A MaxPool3D layer is included in the unit, which selects the maximum value in time and spatial dimensions. Pooling ensures the selection of dominant spatio-temporal features. Max pooling ignores the movement by transitional invariance and maintains the temporal information constant, as the lip movement heavily relies on subtle temporal and spatial changes.

The second block has 128 filters with Leaky ReLU as the activation function and a MaxPool3D layer. Leaky ReLU preserves information, whereas ReLU loses it due to zero gradients. Real-time data contains subtle information that may be captured by Leaky ReLU. In a 3DCNN with multiple layers, ReLU allows neurons to die over a period, whereas leaky ReLU preserves them, stabilizing the feature learning process. The third block has 256 filters with ReLU as the activation function and a Maxpool3D layer. Leaky ReLU improves convergence speed and training stability and is highly recommended for real-time data with uneven illumination. The last layer is built with 75 filters with ReLU as activation function and a Maxpool3D layer.

The Conv3D block is followed by a flattened layer. Two BiLSTM models with 128 units follow the Conv3D block, sandwiched with dropout layers in between. The last layer is the fully connected Dense layer. 3DCNNs are more robust compared to other CNNs, as they are insensitive to slight camera movements, illumination changes, and minor positional changes. Capture of continuous motion is the key advantage of the 3DCNN model, ensuring accurate detection of the lip movement. The $TimeDistributed(Flatten())$ function applies the flatten operation to each frame in the video, converting the spatial feature map of the frame into a 1D vector while preserving the temporal sequence. This is essential when feeding frame-wise features into recurrent layers such as LSTMs for lip-movement detection. The CTC loss function is used, as there is ambiguity between the alignment of the input and output sequences. Its practical application is in scenarios where the data is continuous and sequential, and the output is short. Other loss functions demand a one-to-one alignment between the input and output. In a lip recognition model, the data are continuous video frames, and manually aligning them is an expensive task. The CTC loss function enables the model to learn alignments during the training phase. Multiple alignments between the input and the output are handled by blank symbols and the removal of unnecessary ones during the decoding phase.

The hyperparameters for the transformer were a learning rate of e^{-5}, a batch size of 8, two gradient accumulation steps, and 10 epochs for training.

V. RESULTS AND DISCUSSIONS

The transformed model was implemented on the Windows 11 operating system with 1 TB SSD, 64-bit 16 GB RAM, and Ryzen 3000 series. The lip recognition model was implemented using Google Colab, Python 3, Keras, and TensorFlow, and trained and evaluated on GRID Sentence Audiovisual Corpus dataset. Accuracy, precision, recall, and F1-score were used to evaluate the performance of the model.

$$Accuracy = \frac{TN+TP}{TN+TP+FN+FP} \quad (2)$$

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

$$F1-score = \frac{2 \times precision \times recall}{precision + recall} \quad (5)$$

Figure 6 shows the training and test accuracy graph for 100 epochs, indicating smooth learning, which elucidates important information regarding the model [38]. The lack of sharp zig-zag patterns indicates that the model learns the patterns in the data with more stability, consistence, and predictability. The stable learning curve indicates that the learning rate chosen is optimum for the model. The smooth, escalating curve indicates that the model is gradual in learning the complex relationships and patterns during the training phase. The curve is found to be stable after 60 epochs, indicating that the learning process is almost complete and the model is well-trained.

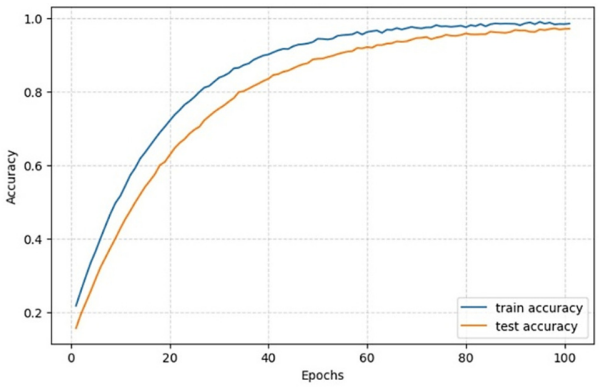

Fig. 6. Model training and test accuracy graph.

Figure 7 shows the confusion matrix of the model, demonstrating 6152 TP, 81 FN, 26 FP, and 64 TN. Accuracy is the number of accurate word predictions compared to the total number of samples. The model accurately identified 6152 samples in the test set, achieving an accuracy of 98.29%, a precision of 99.58, a recall of 98.70%, and an F1-score of 99.13%.

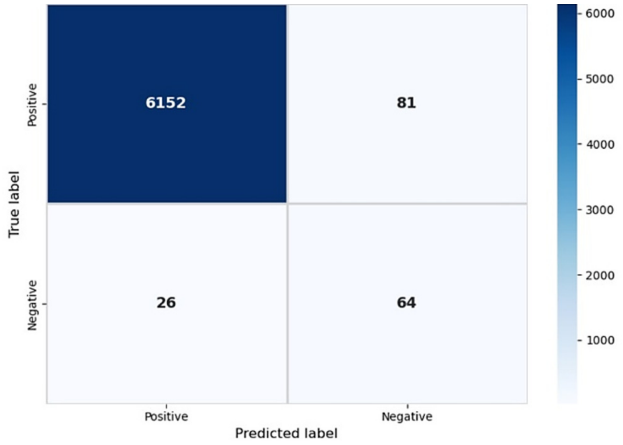

Fig. 7. Confusion matrix for lip recognition model.

A comparative analysis shows that the accuracy of models with speech data is not as high as that of the proposed one with lip movement recognition. Table III shows a comparison with state-of-the-art lip recognition models and their performance in terms of accuracy, demonstrating that the proposed model outperformed other models built on 3DCNN and Bi-GRU. The 3DCNN model captures the distinct low-level features of the lip movement along with the high-level features to recognize the lip movement, extracting spatial and temporal features from consecutive frames to accurately read the text from lip movement. The Bi-LSTM model exhibits superior performance to Bi-GRU models, as it excels in capturing and retaining the long-term temporal dependencies in the data and bidirectional contextual information in the visuals. Features from input sequences are processed simultaneously along the forward and backward directions to recognize speech from the lip movement. The Bi-LSTM network retains the long-term memory of the visual features extracted by the CNN and, hence, is an apt method for lip recognition. The intricate architecture of Bi-LSTM ensures stability of the gradients during training. The CTC loss function significantly improves the model's performance, as it enables the model to learn from previous and subsequent frames rather than a single one. All these modules play a key role in elevating the accuracy of the end-to-end training model that directly maps the video frames to text sequences.

TABLE III. STATE-OF-THE-ART METHODS FOR LIP RECOGNITION

Study	Dataset	Method	Accuracy
[39]	GRID Corpus	LipNet	95.2%
[40]	GRID Corpus	3 layers of STCNN, spatial max-pooling, Bi-LSTM	93.4%
[41]	GRID corpus	LSTM	96.9%
[42]	GRID Corpus	3D CNN, Bi-GRU	97.2%
[43]	GRID Corpus	Shuffle Net and Densely Connected Temporal CNN	98.8 %
This study	GRID Corpus	BiLSTM and 3D CNN	98.2%

The proposed T5-based model for grammar correction and sentence completion was developed with Python and evaluated using Bilingual Evaluation Understudy (BLEU). BLEU is an evaluation metric to assess the quality of the sentence generated, primarily calculating the similarity between the machine translation and the benchmark reference to assess the translation quality. BLEU ranges from 0 to 1. A higher value of BLEU indicates excellent translation or word prediction, while a lower value indicates a large gap between the predicted or translated word and the benchmark reference. The computations of the BLEU metric involve n-grams, which can take values of one, two, three, or four for unigram, bigram, trigram, and four-gram. The sequence is divided according to the number of grams adopted for the model, and model predictions are evaluated by comparing the similarity rate between the predicted results and the reference [44].

The T5-based model exhibited a moderate recall rate and modest precision for BLEU scores, indicating partial alignment with references. The score can be increased by scaling datasets. On a selected validation set, the model exhibited a BLEU score of 0.57. Table IV reports the results of the model with test data of a hundred samples of a custom-built dataset. The classification is for two classes: one is "no change", indicating no grammar correction needed, and the other is "needs correction", indicating sentences that need correction.

TABLE IV. EXPERIMENTAL RESULTS FOR A CUSTOM-BUILT DATASET OF SENTENCES

Class	Precision	Recall	F1-score
Class 1 - No change	78%	74%	76%
Class 2 - Needs correction	78%	81%	80%

Figures 8 and 9 show the confusion matrices for both classes of the grammar correction model. TP refers to cases where the model identified the error and corrected it. TN indicates the number of cases where the model identified no error and made no change. TP involves cases where a sentence was grammatically correct, but the model changed it. Finally, FN involves cases where the model characterized the sentence as correct, despite having grammatical errors.

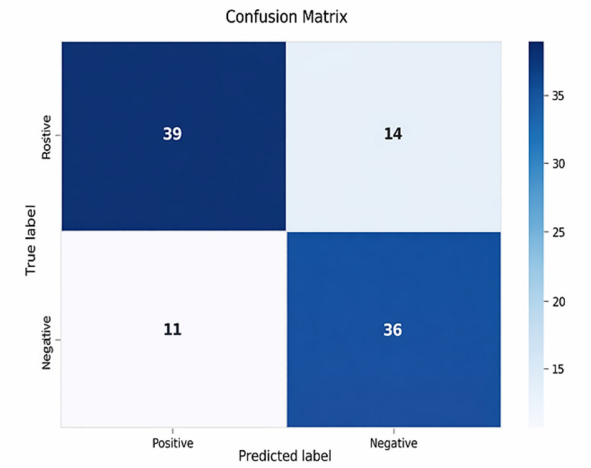

Fig. 8. Confusion matrix for class 1 for the grammar correction model.

In class 2, the number of samples identified as grammatically incorrect and corrected is high. The count of grammatical errors that the model could not correct was 11. The average accuracy was 78%, comparatively moderate for any classification model. The environment in which a model is trained has a great impact on its performance. Transformers demand high computational resources, which were unavailable in this case. Insufficient GPU memory was one of the main constraints that led to undertraining and low performance. Shallow-trained models are unable to predict words accurately. Training the model on a laptop restricted the number of epochs, resulting in unstable and insufficient training and poor performance. Hence, the proposed model could exhibit promising results if trained in a high-performance computing environment for more epochs.

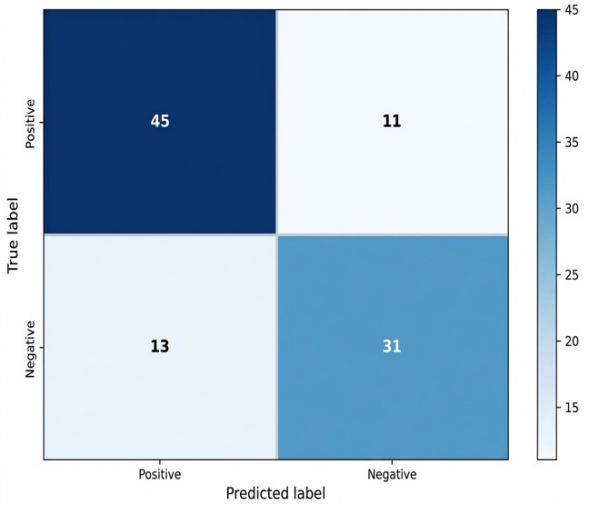

Fig. 9. Confusion matrix for class-2 for the grammar correction model.

VI. CONCLUSION

This study presented an innovative framework for a dysarthria speech recognition tool that achieved a high accuracy of 98% in speech recognition for the GRID corpus. The proposed model recognizes the speech and performs a grammar correction on the text data. The research is the first of its kind to integrate speech recognition with grammar correction to be effectively used by dysarthria patients in a professional environment. With transfer learning, the model can be pretrained and tested on lip movement videos of dysarthria patients. The proposed transformer model can predict the next word in the text generated by the lip recognition module and perform grammar correction. The text can be converted to speech by any existing deep learning-based model. The average accuracy of the proposed transformer model was 78%, which is satisfactory in the current environment with limited resources.

Since lip movements are affected by the disease condition only in a few patients, the proposed model can be an effective assistive tool. In addition, patient-specific customization would improve the performance of the model. The proposed speech recognition framework based on lip movement can be an aid for patients with motor disorders. Future work could make optimizations on the model to maintain an optimum rate of speech communication. The social life of dysarthric patients can be significantly improved with such speech recognition tools, which can also be used to help young patients who need speech assistance after accidents and surgeries.

DECLARATION OF COMPETING INTERESTS

The author declares no competing interests that could have influenced the results of this study.

ACKNOWLEDGMENT

This study was supported through funding from the Prince Sattam bin Abdulaziz University under project number PSAU/ 2026/R/1447.

DATA AVAILABILITY

The GRID audiovisual sentence corpus dataset used in this study is available at [27].

REFERENCES

[1] A. B. Kain, J. P. Hosom, X. Niu, J. P. H. van Santen, M. Fried-Oken, and J. Staehely, "Improving the intelligibility of dysarthric speech," *Speech Communication*, vol. 49, no. 9, pp. 743–759, Sept. 2007, https://doi.org/10.1016/j.specom.2007.05.001.

[2] R. Kumar, M. Tripathy, N. Kumar, and R. S. Anand, "Categorization of patients affected with neurogenerative dysarthria among Hindi-speaking population and analyzing factors causing reduced speech intelligibility at the human-machine interface," *Speech Communication*, vol. 175, Nov. 2025, Art. no. 103328, https://doi.org/10.1016/j.specom.2025.103328.

[3] A. Souky, "Making Speech Happen: The Five Processes Behind Every Word We Say," *Speech & Swallowing Solutions of the Capital Region LLC*, Feb. 10, 2025. https://speechswallowingsolutions.com/how-we-speak/.

[4] "Dysarthria in Adults," *American Speech-Language-Hearing Association*. https://www.asha.org/practice-portal/clinical-topics/dysarthria-in-adults/.

[5] D. K. Jayaraman and J. M. Das, *Dysarthria*. StatPearls Publishing, 2023.

[6] M. K. V. Es *et al.*, "Dysphagia and Dysarthria in Children with Neuromuscular Diseases, a Prevalence Study," *Journal of Neuromuscular Diseases*, vol. 7, no. 3, pp. 287–295, June 2020, https://doi.org/10.3233/JND-190436.

[7] K. Kang *et al.*, "Digital speech assessments and machine learning for differentiation of neurodegenerative diseases," *Clinical Parkinsonism &*

Related Disorders, vol. 13, 2025, Art. no. 100389, https://doi.org/10.1016/j.prdoa.2025.100389.

[8] J. Tröger *et al.*, "An automatic measure for speech intelligibility in dysarthrias—validation across multiple languages and neurological disorders," *Frontiers in Digital Health*, vol. 6, July 2024, Art. no. 1440986, https://doi.org/10.3389/fdgth.2024.1440986.

[9] T. Pu *et al.*, "Lee Silverman Voice Treatment to Improve Speech in Parkinson's Disease: A Systemic Review and Meta-Analysis," *Parkinson's Disease*, vol. 2021, pp. 1–10, Dec. 2021, https://doi.org/10.1155/2021/3366870.

[10] J. A. Russell, M. R. Ciucci, N. P. Connor, and T. Schallert, "Targeted exercise therapy for voice and swallow in persons with Parkinson's disease," *Brain Research*, vol. 1341, pp. 3–11, June 2010, https://doi.org/10.1016/j.brainres.2010.03.029.

[11] J. Mills, O. Duffy, K. Pedlow, and G. Kernohan, "Exploring Speech and Language Therapists' Perspectives of Voice-Assisted Technology as a Tool for Dysarthria: Qualitative Study," *JMIR Rehabilitation and Assistive Technologies*, vol. 12, Sept. 2025, Art. no. e75044, https://doi.org/10.2196/75044.

[12] A. Kehili, K. Dabbabi, and A. Cherif, "Early Detection of Parkinson's and Alzheimer's Diseases using the VOT_Mean Feature," *Engineering, Technology & Applied Science Research*, vol. 11, no. 2, pp. 6912–6918, Apr. 2021, https://doi.org/10.48084/etasr.4038.

[13] S. Salim, S. Shahnawazuddin, and W. Ahmad, "Enhancing voice biometrics for dysarthria patients using novel temporal discriminative feature embedding," *Digital Signal Processing*, vol. 168, Jan. 2026, Art. no. 105662, https://doi.org/10.1016/j.dsp.2025.105662.

[14] S. Aurobindo, R. Prakash , and M. Rajeshkumar, "Comparative analysis of different time-frequency image representations for the detection and severity classification of dysarthric speech using deep learning," *Results in Engineering*, vol. 25, Mar. 2025, Art. no. 104561, https://doi.org/10.1016/j.rineng.2025.104561.

[15] S. Sajiha, K. Radha, D. Venkata Rao, N. Sneha, S. Gunnam, and D. P. Bavirisetti, "Automatic dysarthria detection and severity level assessment using CWT-layered CNN model," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2024, no. 1, June 2024, Art. no. 33, https://doi.org/10.1186/s13636-024-00357-3.

[16] M. S. Remya, P. Ishwar, and P. Nedungadi, "A Hybrid Cross-Attentive CNN-BiLSTM-Transformer Network for Dysarthria Severity Classification," *Scientific Reports*, vol. 15, no. 1, Nov. 2025, Art. no. 42080, https://doi.org/10.1038/s41598-025-26049-2.

[17] B. Moell and F. S. Aronsson, "Voice Cloning for Dysarthric Speech Synthesis: Addressing Data Scarcity in Speech-Language Pathology." arXiv, 2025, https://doi.org/10.48550/ARXIV.2503.01266.

[18] M. T. A Celin., P. Vijayalakshmi, T. Nagarajan, and K. Mrinalini, "Augmentative and alternative speech communication (AASC) aid for people with dysarthria," *Computer Speech & Language*, vol. 92, June 2025, Art. no. 101777, https://doi.org/10.1016/j.csl.2025.101777.

[19] M. Kim, Y. Kim, J. Yoo, J. Wang, and H. Kim, "Regularized Speaker Adaptation of KL-HMM for Dysarthric Speech Recognition," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 9, pp. 1581–1591, Sept. 2017, https://doi.org/10.1109/TNSRE.2017.2681691.

[20] N. M. Joy and S. Umesh, "Improving Acoustic Models in TORGO Dysarthric Speech Database," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 3, pp. 637–645, Mar. 2018, https://doi.org/10.1109/TNSRE.2018.2802914.

[21] M. S. Yakoub, S. Selouani, B. F. Zaidi, and A. Bouchair, "Improving dysarthric speech recognition using empirical mode decomposition and convolutional neural network," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2020, no. 1, Dec. 2020, https://doi.org/10.1186/s13636-019-0169-5.

[22] S. Salim and W. Ahmad, "Advancing Voice Biometrics for Dysarthria Speakers Using Multitaper LFCC and Voice Conversion Data Augmentation," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 10114–10129, 2024, https://doi.org/10.1109/TIFS.2024.3484661.

[23] R. Vinotha, D. Hepsiba, L. D. Vijay Anand, J. Andrew, and R. Jennifer Eunice, "Enhancing dysarthric speech recognition through SepFormer and hierarchical attention network models with multistage transfer learning," *Scientific Reports*, vol. 14, no. 1, Nov. 2024, Art. no. 29455, https://doi.org/10.1038/s41598-024-80764-w.

[24] M. Thomas, E. Fish, and R. Bowden, "VALLR: Visual ASR Language Model for Lip Reading." arXiv, 2025, https://doi.org/10.48550/ARXIV.2503.21408.

[25] Y. Fu and Y. Lu, "Lip-Reading Research Based on ShuffleNet and Attention-GRU," presented at the 10th International Conference on Human Interaction and Emerging Technologies (IHIET 2023), 2023, https://doi.org/10.54941/ahfe1004024.

[26] C. Innocente *et al.*, "Deep Learning-Based Lip-Reading for Vocal Impaired Patient Rehabilitation," *Computer Modeling in Engineering & Sciences*, vol. 143, no. 2, pp. 1355–1379, 2025, https://doi.org/10.32604/cmes.2025.063186.

[27] "The GRID audiovisual sentence corpus." [Online]. Available: https://spandh.dcs.shef.ac.uk/gridcorpus/.

[28] S. Kumar, S. Datta, V. Singh, S. K. Singh, and R. Sharma, "Opportunities and Challenges in Data-Centric AI," *IEEE Access*, vol. 12, pp. 33173–33189, 2024, https://doi.org/10.1109/ACCESS.2024.3369417.

[29] H. B. Abdalla *et al.*, "The Future of Artificial Intelligence in the Face of Data Scarcity," *Computers, Materials & Continua*, vol. 84, no. 1, pp. 1073–1099, 2025, https://doi.org/10.32604/cmc.2025.063551.

[30] M. Hähnel, "Ethical challenges and solutions in AI-driven medical data management: a focus on distributed machine learning," *Discover Artificial Intelligence*, vol. 5, no. 1, May 2025, Art. no. 53, https://doi.org/10.1007/s44163-025-00266-0.

[31] H. Kheddar, "Transformers and large language models for efficient intrusion detection systems: A comprehensive survey," *Information Fusion*, vol. 124, Dec. 2025, Art. no. 103347, https://doi.org/10.1016/j.inffus.2025.103347.

[32] G. Antonesi, T. Cioara, I. Anghel, V. Michalakopoulos, E. Sarmas, and L. Toderean, "A systematic review of transformers and large language models in the energy sector: towards agentic digital twins," *Applied Energy*, vol. 401, Dec. 2025, Art. no. 126670, https://doi.org/10.1016/j.apenergy.2025.126670.

[33] S. Li and Y. Sung, "Transformer-Based Seq2Seq Model for Chord Progression Generation," *Mathematics*, vol. 11, no. 5, Feb. 2023, Art. no. 1111, https://doi.org/10.3390/math11051111.

[34] S. Grassi, "Examining the limitations and challenges of using Transformers for time series forecasting." ResearchGate, 2024, https://doi.org/10.13140/RG.2.2.32456.53765.

[35] L. Yang and S. Qiu, "BLEU Function Analysis of Machine Translation Based on Transformer Model," in *Proceedings of the 2024 International Conference on Artificial Intelligence, Digital Media Technology and Interaction Design*, Nov. 2024, pp. 230–236, https://doi.org/10.1145/3726010.3726046.

[36] A. Vaswani *et al.*, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, 2017, vol. 30.

[37] H. A. Z. S. Shahgir and K. S. Sayeed, "Bangla Grammatical Error Detection Using T5 Transformer Model." arXiv, 2023, https://doi.org/10.48550/ARXIV.2303.10612.

[38] Y. Jiang and R. Dale, "Mapping the learning curves of deep learning networks," *PLOS Computational Biology*, vol. 21, no. 2, Feb. 2025, Art. no. e1012286, https://doi.org/10.1371/journal.pcbi.1012286.

[39] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas, "LipNet: End-to-End Sentence-level Lipreading." arXiv, Dec. 16, 2016, https://doi.org/10.48550/arXiv.1611.01599.

[40] Y. Li, Y. Takashima, T. Takiguchi, and Y. Ariki, "Lip reading using a dynamic feature of lip images and convolutional neural networks," in *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*, June 2016, pp. 1–6, https://doi.org/10.1109/ICIS.2016.7550888.

[41] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip Reading Sentences in the Wild," in *2017 IEEE Conference on Computer Vision*

and Pattern Recognition (CVPR), July 2017, pp. 3444–3453, https://doi.org/10.1109/CVPR.2017.367.

[42] K. Xu, D. Li, N. Cassimatis, and X. Wang, "LCANet: End-to-End Lipreading with Cascaded Attention-CTC," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, May 2018, pp. 548–555, https://doi.org/10.1109/FG.2018.00088.

[43] Y. Li, A. S. Hashim, Y. Lin, P. N. E. Nohuddin, K. Venkatachalam, and A. Ahmadian, "AI-based visual speech recognition towards realistic avatars and lip-reading applications in the metaverse," *Applied Soft Computing*, vol. 164, Oct. 2024, Art. no. 111906, https://doi.org/10.1016/j.asoc.2024.111906.

[44] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "BLEU: A method for automatic evaluation of machine translation," presented at the Annual Meeting of the Association for Computational Linguistics, July 2002.