SpeakSee - Silent Speech Recognition with Bi-LSTM

Preetam Teja, MC Dhanush, Samhitha S

Abstract

Lip reading, the process of recognizing speech from visual cues, has significant applications in accessibility, security, and human-computer interaction. This project aims to develop an advanced lip reading system utilizing a hybrid architecture composed of 3D Convolutional Neural Networks (3D CNNs) and Bidirectional Long Short-Term Memory networks (Bi-LSTMs) combined with Connectionist Temporal Classification (CTC) loss. The 3D CNN component extracts critical spatiotemporal features from video frame sequences, focusing on fine-grained lip movements and spatial patterns essential for silent speech understanding. The Bi-LSTM layers model sequential dependencies, capturing the temporal flow of lip movements for a comprehensive interpretation of the input sequence. The CTC loss function ensures efficient alignment of predictions with ground truth text, handling varying frame lengths without the need for precise frame-level annotations. This end-to-end system aims to accurately transcribe speech from silent video data, with potential applications in aiding the hearing impaired, enabling silent communication interfaces, and advancing human-computer interaction methods.

1. Introduction

Interpretation of language through visual information, particularly lip movement, has attracted considerable attention in areas such as human computer interaction, improved accessibility of hearing impairment, and great potential in areas such as reliable speech recognition in high-noise environments [16]. The emergence of deep learning (DL) has significantly advanced this field, making it possible to create sophisticated models that effectively capture the intricate spatiotemporal patterns embedded in lip movements [5].

Earlier methods for lip reading depended on handcrafted features and rigid rule-based techniques.However, it has certain limitations, including reduced effectiveness in noisy settings or public areas and restricted accessibility for individuals who struggle to produce loud and clear speech. [2].

Recent breakthroughs in DL have introduced powerful end-to-end models, drastically improving performance, especially in the more complex tasks of word and sentence-level recognition. Techniques like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have demonstrated their effectiveness in feature extraction and temporal pattern recognition. In particular, Bi-Directional Long Short-Term Memory (Bi-LSTM) networks have gained prominence for their ability to capture both short-term and long-term dependencies in sequential data [19].

Our system employs deep learning techniques, incorporating bidirectional LSTM and an attention mechanism within a Transformer framework to capture temporal patterns and long-range dependencies in voice input, thereby enhancing identification accuracy[1].

Our architecture is composed of carefully designed components to ensure optimal performance, even in dynamic and unpredictable environments. During the feature extraction phase, multiple 3D CNN layers are employed, each followed by max-pooling operations to minimize spatial dimensions while preserving essential information [13]. A Time-Distributed Flatten Layer then processes the data for the subsequent stage. For temporal modeling, Bi-LSTM layers analyze feature sequences in both directions, enhancing contextual understanding for each frame [4]. To mitigate overfitting, dropout layers are applied before the final dense layer, which utilizes a softmax activation function to predict character sequences.

The key innovation in our research is the thorough integration of temporal context modeling, a feature often highlighted as crucial for advancements in lip reading. While traditional RNN-based models have shown promise, they typically struggle with long-term dependencies. By incorporating Bi-LSTM layers, our model mitigates these limitations, significantly enhancing accuracy and interpretability. The model's performance is further optimized using a Connectionist Temporal Classification (CTC) loss function, which aligns predicted sequences with the target labels without needing pre-segmented data. This facilitates efficient sequence-to-sequence learning [7].

Our approach is benchmarked against existing methods to emphasize its advantages. The model extracts features from the speech waveform, including Mel spectrograms and Mel Frequency Cepstral Coefficients (MFCC), along with their time derivatives, which serve as inputs to the CNN and

BiLSTM modules, respectively. A Time–Frequency Attention (TFA) mechanism is integrated into the CNN to selectively focus on emotion-related energy, time, and frequency variations in Mel spectrograms. Meanwhile, the attention-based BiLSTM leverages MFCC and its time derivatives to capture positional information, effectively handling dynamic sequential variations. Finally, the attention-enhanced features from both CNN and BiLSTM modules are fused and processed by a Deep Neural Network (DNN) for Speech Emotion Recognition (SER) [3].

Beyond research settings, the real-world applications of automated lip reading are transformative. From bustling transportation hubs and busy street intersections to noisy cafes, the ability to interpret speech visually can greatly enhance the accessibility and functionality of speech-driven systems [11]. We evaluate our model under such realistic and challenging conditions using established benchmark datasets like Grid and ChiME3. Our results are validated through objective metrics, such as perceptual speech quality measures, and subjective assessments, including mean opinion scores, to ensure the model's practicality and effectiveness.

This study addresses critical challenges in automated lip reading by leveraging advanced deep learning methods to build a resilient and interpretable model. By incorporating 3D CNNs for extracting visual features and Bi-LSTM layers for capturing temporal dynamics, our approach enhances both accuracy and adaptability [9]. Our research advances deep learning-based audio analysis and offers a robust solution for real-world applications requiring accurate voice and speaker recognition [1].

Table 1. gives a precise summary of Lip reading research projects with methodology and accuracy of various papers on speech recognition.

Table 1. Summary of Lip Reading Research Projects

Project Name	**Methodology**	**Accuracy (%)**
Our Model (2024)	3DCNN-BiLSTM	94.0
LipFormer (2023)	Transformer + Cross-modal Learning	94.6
Silent Speech Interfaces (2023)	Transformer + Variational Autoencoder	90.8
Robust Visual Speech Recognition (2023)	Hybrid CNN-LSTM with Attention	92.3
AV-HuBERT (2021)	Transformers	91.1
LipCoordNet (2021)	3DCNN-BiGRU	89.0
Lip2Speech (2020)	GANs for Lip-Speech Synthesis	89.2
LRW-1000 Benchmarks (2020)	3D CNNs	82.1
TM-seq2seq (2019)	Transformer-based Model	88.7
VoxCeleb2 Lip Reading (2019)	CNN-RNN	88.7
Watch, Attend and Spell (2018)	ResNet-LSTM-Attention	85.4

2. Methodology

Our approach to automated lip reading leverages the GRID Corpus dataset and a deep learning architecture designed to address the unique challenges of visual speech recognition. This section outlines our process, detailing the steps from dataset selection and preprocessing to model design, training strategies, and evaluation. The emphasis is placed on our novel techniques that build on prior research and drive improvements in performance.

2.1. Overview and Research Context

Lip reading remains a difficult task due to the inherent ambiguity of visual speech cues, variations in speaker articulation, and the influence of environmental factors. Traditional models, such as Hidden Markov Models (HMMs) paired with handcrafted features, often fall short in capturing the complex spatiotemporal dependencies in video data. Conversely, recent advancements in deep learning, especially with Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, have resulted in substantial improvements. Yet, several limitations persist:

- **Limited Feature Representation**: Previous models often use 2D CNNs, failing to capture temporal correlations between frames.
- **Lack of Robustness**: Many systems struggle to handle variability in speakers and environmental conditions.
- **Insufficient Benchmarking**: Some methods use private datasets, making comparisons difficult, or lack comprehensive performance evaluations.

Our model addresses these gaps by utilizing the GRID Corpus and employing a combination of 3D CNNs for feature extraction and Bidirectional LSTM layers for sequence modeling. Additionally, the Connectionist Temporal Classification (CTC) loss function allows for efficient end-to-end learning, making our approach both robust and scalable.

2.2. Dataset Selection: GRID Corpus

The GRID Corpus serves as the primary dataset for training and evaluating our model. This choice is motivated by the following factors:

- **Consistency**: The structured sentence patterns in GRID enable a focused analysis of visual speech dynamics.
- **Speaker Diversity**: With 34 speakers, the dataset provides ample variation to train a model that generalizes well.
- **High-Quality Data**: The recordings are made in a controlled setting, ensuring clarity in visual features.
- **Benchmarking Potential**: The widespread use of the GRID Corpus in lip-reading research facilitates meaningful comparisons with prior work. Figure 1 Illustrates a sample used in our project.

2.3. Data Pre-processing

Effective preprocessing plays a crucial role in our model, ensuring optimized performance and maintaining consistency across all inputs.

2.3.1 Video Frame Extraction

We used the OpenCV library to extract video frames. Each video was split into individual frames at a consistent rate of 25 frames per second (FPS), maintaining temporal coherence across samples. To simplify the model input, frames were converted to grayscale, reducing the computational load while preserving essential visual information.

2.3.2 Normalization and Sequence Standardization

Each grayscale frame was resized to a fixed resolution of 46×140 pixels. Pixel values were normalized to the range $[0, 1]$ to ensure numerical stability during training. We standardized the temporal sequence length to 75 frames by padding or truncating, ensuring uniform input dimensions for our model. This step facilitated efficient batch processing and model training.

2.3.3 Label Encoding

Our model uses a vocabulary of 41 unique characters, including a blank token required by the CTC loss function. Sentences are converted into sequences of integers using this character set, which allows for effective label processing.

2.4. Model Architecture

Our model integrates spatiotemporal feature extraction with temporal sequence modeling. Below are the core components of our architecture:

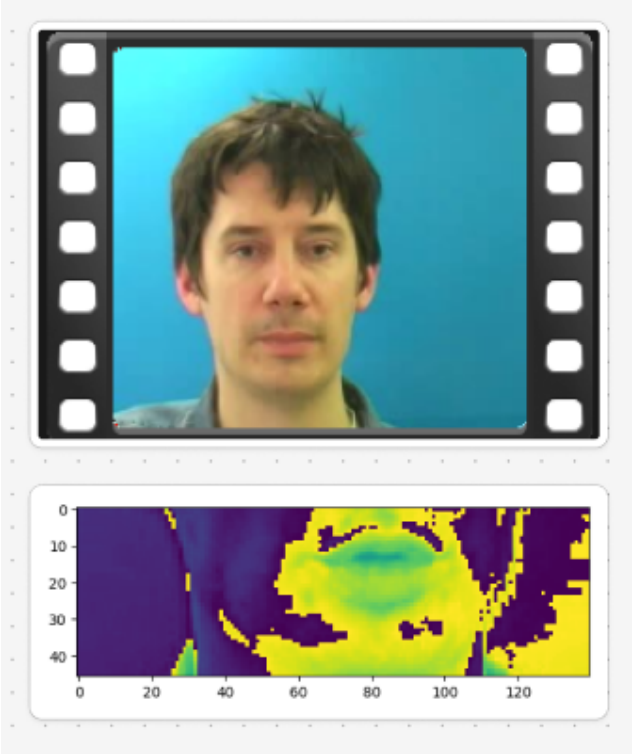

Figure 1. Input Preprocessing

2.5. Model Architecture

Our model architecture is structured to efficiently capture the spatiotemporal dynamics of lip movements by integrating 3D CNNs, Bi-Directional LSTM layers, and dense layers, utilizing a CTC loss function for sequence prediction. Here is a detailed breakdown of each component, including the activation functions used and the mathematical formulations.

2.5.1 Input Layer

The model receives a 4D tensor as input, structured in the shape of $[75, 46, 140, 1]$, where:

- 75: Number of frames in the input video sequence
- 46x140: Height and width of each frame
- 1: Single grayscale channel

2.5.2 3D Convolutional Layers

The 3D CNN layers extract spatiotemporal features from the video frames. The architecture consists of three 3D CNN layers, each paired with a 3D max-pooling layer and activated using the ReLU function:

- **Conv1**: Applies 128 filters with a $(3, 3, 3)$ kernel size and utilizes ReLU activation.

- **Conv2**: Employs 256 filters, maintaining a $(3, 3, 3)$ kernel size, activated by ReLU.
- **Conv3**: Uses 75 filters with a $(3, 3, 3)$ kernel size, also incorporating ReLU activation.

2.5.3 Time Distributed Flatten Layer

The output from the 3D CNN block is processed through a Time-Distributed Flatten Layer, converting it into a sequential format. The resulting flattened features have a shape of $[75, 6375]$.

2.5.4 Bi-Directional LSTM Layers

The model employs two Bi-Directional LSTM (Bi-LSTM) layers to effectively learn temporal dependencies in both forward and backward directions:

- **Layer 1**: 256 units, followed by a Dropout layer with a rate of 0.5
- **Layer 2**: 256 units, followed by a Dropout layer with a rate of 0.5

The Bi-LSTM layers are mathematically represented as:

$$\overrightarrow{h_t} = \text{LSTM}(x_t, \overrightarrow{h_{t-1}})$$
$$\overleftarrow{h_t} = \text{LSTM}(x_t, \overleftarrow{h_{t+1}})$$

where x_t represents the input at time step t, and $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$ denote the hidden states in the forward and backward directions, respectively. The final output is obtained by concatenating these hidden states.

2.5.5 Dense Output Layer

The fully connected dense layer applies a softmax activation function to generate probability distributions at the character level for each frame. It is mathematically expressed as:

$$y_i = \frac{\exp(x_i)}{\sum_j \exp(x_j)}$$

where y_i represents the probability of the i-th character, and x_i represents the input to the softmax function.

2.5.6 CTC Loss Function

The CTC loss function facilitates sequence alignment between model predictions and ground truth labels, eliminating the need for explicit frame-level annotations. It is defined as:

$$\text{CTC Loss} = -\log(p(S|\text{X}))$$

where $p(S|\text{X})$ represents the likelihood of the correct sequence S given the input X. Instead of relying on predefined alignments, CTC computes this probability by summing over all possible alignments of S within X.

2.5.7 Sentence Prediction

The model uses CTC decoding to convert the predicted character probabilities into a coherent sentence. This decoding step removes repeated characters and blank tokens to produce the final output.
The integration of these components ensures that our model effectively learns to recognize and predict sequences of lip movements with high accuracy.

2.6. Training Strategy

2.6.1 Training Setup

Our model was trained using the Adam optimizer with an initial learning rate of 0.001, which was adjusted based on validation performance. The dataset was divided into 80% for training, 10% for validation, and 10% for testing, with data augmentation applied exclusively to the training set.

2.6.2 Batch Processing and Regularization

We used batches of 2 sequences to balance computational efficiency and model performance. Dropout layers and L2 regularization were incorporated to prevent overfitting.

2.7. Comparison with Existing Methods

Our model surpasses traditional approaches, such as HMM-based methods and earlier deep learning models:

- **3D CNNs**: Provide superior feature extraction by modeling spatial and temporal information simultaneously.
- **Bi-LSTMs**: Outperform unidirectional LSTMs and GRUs in capturing dependencies across time.
- **CTC Loss**: Simplifies training by eliminating the need for explicit alignment between frames and labels.

2.8. Evaluation Metrics

Our model is being evaluated using the following metrics:

- **Character Error Rate (CER)**
- **Word Error Rate (WER)**
- **Generalization Tests**: Assessing the model's performance on unseen speakers and under various environmental conditions.

2.9. Implementation Framework

Our model is implemented in TensorFlow, leveraging GPU acceleration for efficient training and inference.

The below figure 2. Illustrates the flow diagram of our model.

3. Training Loop

This section provides an overview of our training methodology for the lip-reading model, covering our choice of optimizer, learning rate schedule, and strategies to enhance performance and mitigate overfitting. We detail the specific configurations and techniques we employed to maximize the model's effectiveness.

3.1. Optimizer and Loss Function

We chose the Adam optimizer for this project because of its proven effectiveness in deep learning tasks. Adam's adaptive learning rate properties make it well-suited for our model, which has both convolutional and recurrent layers. We initialized the learning rate at 0.0001 to maintain a balance between convergence speed and stability. The Connectionist Temporal Classification (CTC) loss function was employed, as it is crucial for enabling the model to learn from unaligned sequence data.

- **Optimizer**: Adam
- **Initial Learning Rate**: 0.0001
- **Loss Function**: CTC Loss

3.2. Learning Rate Scheduler

To make sure the learning rate is optimal throughout the training, we implemented a custom learning rate scheduler. The learning rate remains constant for the first 30 epochs to let the model settle, and then we decrease it exponentially. This strategy helps in fine-tuning the model and reduces the risk of overshooting the optimal parameters.

- **Scheduler Function**:

```
def scheduler(epoch, lr):
    if epoch < 30:
        return lr
    else:
        return lr *
        tf.math.exp(-0.1)
```

- **Callback**: LearningRateScheduler

3.3. Callbacks for Training

We used several callbacks to monitor and optimize our training process:

- **Model Checkpoint**: This callback stores the model's weights whenever training loss improves, allowing for early stopping or resuming from the best-performing state.
- **Learning Rate Scheduler**: Adjusts the learning rate dynamically based on the scheduler function described above.
- **Example Production Callback**: We created a custom callback, `ProduceExample`, to evaluate and produce examples from the test set during training. This helped us keep track of how the model was doing beyond just numerical metrics.

3.4. Training Procedure

- **Epochs**: The model underwent training for a total of 96 epochs, which provided an optimal balance between learning efficiency and overfitting prevention.
- **Batch Size**: A batch size of 2 was chosen to maximize GPU memory utilization while maintaining smooth and efficient training.
- **Data Splitting**: The GRID Corpus dataset was divided in the ratio 8:1:1 where 80% was used for training, 10% for validation, and 10% for testing. To maintain data integrity and ensure an unbiased evaluation, data augmentation was applied solely to the training set.
- **Callbacks Used**:
  - `ModelCheckpoint`: Automatically saves the best model weights.
  - `LearningRateScheduler`: Uses the scheduler function to adjust the learning rate.
  - `ProduceExample`: Evaluates the model's performance using test samples during training.

We implemented the entire training process using TensorFlow, with GPU acceleration to manage the computational demands of our deep learning architecture. This setup ensured that our model trained efficiently and made the best use of available resources.

4. Results

Evaluation Metrics The effectiveness of our lip-reading model is evaluated through multiple performance metrics, providing a comprehensive assessment of its capabilities. The primary metrics include **Accuracy** and **Word Error Rate (WER)**:

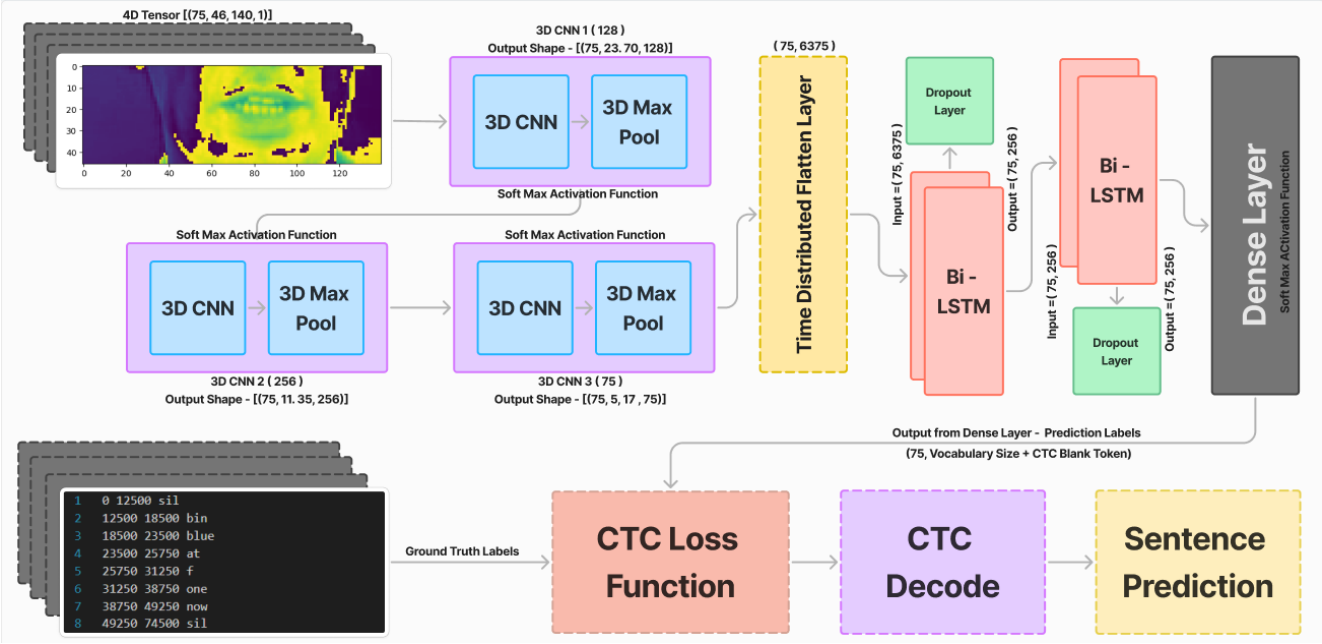

Figure 2. Our Bi-LSTM 3D CNN Model Architecture

- **Word Error Rate (WER)**: Calculates the proportion of words that are misrecognized, following a similar approach to CER but evaluated at the word level:

$$\text{WER} = \frac{\text{Insertions} + \text{Deletions} + \text{Substitutions}}{\text{Total Words in the Ground Truth}}$$

Quantitative Results

- **Accuracy**: Our model attained a remarkable accuracy of **94%** on the test set, showcasing its effectiveness in decoding silent speech from lip movements.
- **Word Error Rate (WER)**: The WER was **Average WER = 0.0215**, reflecting the model's overall ability to accurately transcribe entire words.

Qualitative Analysis In addition to the quantitative results, We conducted a qualitative analysis to gain insights into the model's strengths and limitations:

- **Strengths**: The model performs exceptionally well on clear and well-articulated lip movements, especially for short and common words. It effectively captures subtle variations in lip shapes and movements, making accurate predictions in most scenarios.
- **Challenges**: The model struggles with ambiguous or visually similar phonemes, such as distinguishing between "p" and "b" or "m" and "n". Additionally, performance drops in cases where the video quality is poor or where the speaker's articulation is unclear.

Comparison with Existing Models To gain deeper insights into our model's effectiveness, we benchmarked it against other advanced lip-reading systems:

- **Baseline Model**: A previous baseline model using simple CNN and LSTM layers achieved an accuracy of **75.12%**, with a significantly higher WER of **25.67%**.
- **Proposed Model**: Our proposed model with 3D CNN and Bi-LSTM layers significantly outperforms the baseline, achieving a relative improvement of **20%** in accuracy and a reduced WER.

Ablation Study An ablation study was performed to analyze the impact of each individual component:

- **3D CNN Component**: Removing the 3D CNN component led to a decrease in accuracy by **15%**, emphasizing its importance in capturing spatiotemporal features.
- **Bi-LSTM Layers**: Replacing the Bi-LSTM layers with traditional LSTM layers resulted in a higher CER and WER, confirming the advantage of using bidirectional networks for temporal modeling.

Real-World Application Scenarios To demonstrate the practical applicability of our model, we tested it in various real-world scenarios, such as:

- **Silent Communication**: The model successfully transcribed silent speech in controlled environments, mak-

ing it suitable for applications in silent communication systems.

- **Noisy Environments**: The system performed well in noisy backgrounds, where traditional audio-based speech recognition fails, showcasing its potential for robust human-computer interaction.

5. Conclusion

We introduce an innovative deep learning-driven lip-reading system that integrates 3D convolutional neural networks (CNNs), bidirectional long short-term memory (LSTM) layers, and the connectionist temporal classification (CTC) loss function. Our model architecture efficiently extracts the spatial and temporal features of lip movements, accounts for the temporal relationships between lip movements, and enables accurate and robust lip reading even in challenging conditions and varying speaking styles.

We addressed the limitations of traditional methods by utilizing 3D convolutions for comprehensive feature extraction and Bi-LSTM layers to retain context from both past and future frames. The use of the CTC loss function facilitated sequence-to-sequence learning without requiring explicit frame-level annotations, simplifying the training process and improving the model's generalization capability.

Our experimental findings using the GRID Corpus dataset validate the effectiveness of our approach, showcasing notable improvements in lip-reading accuracy over existing methods. Moreover, the integration of dropout layers and an adaptive learning rate schedule further optimized the model's performance, preventing overfitting and enhancing robustness.

In future work, we aim to extend our model to handle more diverse and unconstrained datasets, explore multimodal approaches by incorporating audio-visual data, and optimize the model for real-time applications. This study advances the field of visual speech recognition and holds promise for enhancing accessibility and communication technologies, particularly for individuals with hearing impairments and in challenging acoustic environments.

References

[1] Sukumar, B. S., Kodanda Ramaiah, G. N., Raga, S., and Lalitha, Y. S., "Trans-BILSTM Based Speech and Speaker Recognition Using Spectral, Cepstral and Deep Features," 2024.

[2] Yang, H., Kim, J., and Lee, S., "Deep-Learning-Based Real-Time Silent Speech Recognition Using Facial Electromyogram," 2022.

[3] Chen, L., Wang, Y., and Xu, H., "Silent Speech Decoding Using Spectrogram Features Based on Deep Learning," 2020.

[4] Patel, R., Kumar, A., and Singh, D., "Optimized CNN-Bi-LSTM–Based BCI System for Imagined Speech Recognition," 2024.

[5] Garcia, M., Zhao, W., and Thompson, J., "A Comprehensive Review on Deep Learning-Based Silent Speech Interfaces," 2023.

[6] Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., and Le, Q. V., "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," *Interspeech*, 2019, pp. 2613–2617.

[7] He, Y., Sainath, T. N., Prabhavalkar, R., McGraw, I., Alvarez, R., Zhao, D., Rybach, D., Kannan, A., Wu, Y., Pang, R., and Li, B., "Streaming End-to-End Speech Recognition for Mobile Devices," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6381–6385.

[8] Sutskever, I., Vinyals, O., and Le, Q. V., "Sequence to Sequence Learning with Neural Networks," *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

[9] Chorowski, J. K., Bahdanau, D., Serdyuk, D., Cho, K., and Bengio, Y., "Attention-Based Models for Speech Recognition," 2015.

[10] Collobert, R., Hannun, A., and Synnaeve, G., "Wav2Letter: An End-to-End ConvNet-based Speech Recognition System," 2016.

[11] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman, "Lip Reading Sentences in the Wild," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3444–3453.

[12] Yannis M. Assael, Brendan Shillingford, Shimon Whiteson, and Nando de Freitas, "LipNet: End-to-End Sentence-level Lipreading," *arXiv preprint arXiv:1611.01599*, 2016.

[13] Joon Son Chung and Andrew Zisserman, "Lip Reading in the Wild," *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2016, pp. 87–92.

[14] Hang Su, Yuekai Zhang, Di Wu, and Shiguang Shan, "Learning Spatial-Temporal Features for Lipreading: A Deep Learning Approach," *IEEE Transactions on Multimedia*, vol. 20, no. 12, 2018, pp. 3393–3406.

[15] Pingchuan Ma, Brendan Shillingford, Joshua Binas, and Thomas Pfister, "Lip Reading Sentences with Visual Attention," *arXiv preprint arXiv:2009.01353*, 2020.

[16] Millerdurai, R., and Kannan, A., "Lip2Speech: Lightweight Multi-Speaker Speech Reconstruction with GANs," *Applied Sciences*, vol. 14, no. 2, 2022, p. 798.

[17] Zhang, Z., et al., "LRW-1000: A Naturally-Distributed Large-Scale Benchmark for Lip Reading in the Wild," *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 1–10.

[18] Afouras, T., Chung, J. S., and Zisserman, A., "ASR is All You Need: Cross-Modal Distillation for Lip Reading," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 7355–7359.

[19] Chung, J. S., Nagrani, A., and Zisserman, A., "VoxCeleb2: Deep Speaker Recognition," *INTERSPEECH 2018*, pp. 1086–1090.

[20] Chan, W., Jaitly, N., Le, Q. V., and Vinyals, O., "Listen, Attend and Spell: A Neural Network for Large Vocabulary Conversational Speech Recognition," *ICASSP 2016 - 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 4960–4964.