ISSN No:-2456-2165

Deep Learning Model for Lip Reading to ImproveAccessibility

Sonia Singh B[1]
RV College of Engineering
Bangalore, India

Shubhaprada K P[2]
RV College of Engineering
Bangalore, India

Abstract:- The project proposes an end-to-end deep learning architecture for word-level visual speech recognition without the need for explicit word boundary information. The methodology includes spatiotemporal convolutional layers, Residual Networks (Res Nets), and bidirectional Long Short-Term Memory (Bi- LSTM) networks. The system is trained using the CTC loss function and requires data preprocessing with facial landmark extraction, image cropping, resizing, grayscale conversion, and data augmentation to focus on the mouth region. The model is implemented in Tensor Flow and trained with an adaptive learning rate schedule. With this approach, the proposed system achieves end-to-end lip reading from a video frame and implicitly identifies keywords in utterances. Analysis using the CTC loss function confirms the model's effectiveness. The results suggest potential applications in dictation, hearing aids, and biometric authentication, thus advancing visual speech recognition compared to traditional methods. In summary, the project presents an innovative deep learning architecture for word-level visual speech recognition, surpassing traditional methods and enabling practical applications.

Keywords:- *Recurrent Neural Network, Long Short-Term Memory, Graphics Processing Unit, Solid State Drive, Text-to- Speech, Application Programming Interface, Audio-Visual, Lip Reading, Bidirectional Long Short-Term Memory, Graphical User Interface, Red Green Blue, Mean Squared Error, Mean Absolute Error, Adaptive Moment Estimation.*

I. INTRODUCTION

Visual speech recognition, also known as lipreading, holds significant potential for enhancing speech-related applications, including dictation in noisy environments, silent dictation in public spaces, hearing aids, and biometric authentication. By harnessing the power of deep learning techniques, visual speech recognition systems have made remarkable progress in recent years. However, the complexities posed by un- constrained real-world environments and the variability in speaker pose and language still present intriguing challenges. Addressing these challenges could lead to more robust and accurate AVSR systems, contributing to better speech recog- nition performance in various practical scenarios.

Despite the remarkable advancements in audio-visual speech recognition, there is a need to address specific challenges to further improve the accuracy and robustness of lipreading systems. These challenges include handling utterances with unknown word boundaries, achieving high accuracy across diverse languages and speakers, and dealing with Identify applicable funding agency here. If none, delete this variations in pose and environmental conditions. Additionally, existing systems may not fully exploit the potential of auto- matic labels, which can limit the scalability and generalizationof the models.

This project focuses on developing and evaluating an end- to-end deep learning architecture for audio-visual speech recognition, specifically targeting word-level recognition. The proposed methodology aims to handle utterances without explicit word boundary information during both training and testing phases. The evaluation will be performed on large- scale benchmark datasets, enabling the assessment of the system's performance across multiple languages and diverse speakers. However, the project's scope does not extend to otherspeech recognition tasks, such as sentence-level classification or phoneme-level recognition.

- *Abbreviations and Acronyms*

RNN - Recurrent Neural Network LSTM - Long Short-TermMemory GPU - Graphics Processing Unit SSD - Solid State Drive TTS - Text-to-Speech API - Application Programming Interface AV - Audio-Visual LR - Lip Reading Bi-LSTM - Bidirectional Long Short-Term Memory GUI - Graphical User Interface RGB - Red Green Blue MSE - Mean Squared Error MAE - Mean Absolute Error Adam - Adaptive Moment Estimation

II. STATE OF THE ART DEVELOPMENTS

Dalu Feng et al. [1] present an effective lip reading model capable of handling varying lighting conditions, facial ex- pressions, and speaker accents. The model achieved a word accuracy of 82.6 percent on a challenging dataset with diverse accents and lighting conditions. However, the model is limited to specific languages or accents and may struggle with extremelighting variations.

Pingchuan Ma et al. [2] extend lip reading to multiple languages and cultural backgrounds, achieving 75.3 percent accuracy across 5 languages. They demonstrate the model's cross-lingual capabilities, but limited training data for low-resource languages and the need for further validation with

more language variations pose challenges.

Shuang Yang et al. [3] introduce the LRW-1000, a large- scale lip reading benchmark with naturally occurring speech data. The dataset encompasses diverse speaking styles and environments, but better generalization to unseen conditions and research on handling background noise are needed.

Themos Stafylakis and Georgios Tzimiropoulos [4] propose an end-to-end deep learning architecture combining residual networks with LSTMs for word-level visual speech recogni- tion. The model achieved 83.0 percent word accuracy, outper- forming the state-of-the-art, but real-world noise exploration and extensive evaluations are desired.

Yannis M. Assael et al. [5] develop an end-to-end lipreading system for sentence-level visual speech recognition, achieving 95.2 percent sentence-level accuracy on a subset of GRID speakers. However, the model is limited to sentence-level lipreading and may experience performance drop with noisy data.

Pingchuan Ma et al. [6] explore training strategies for improved lip reading on challenging datasets, including trans- fer learning and data augmentation techniques for improved accuracy on noisy data. However, the model may not be suitable for real-time applications and remains sensitive toextreme noise and low-resolution inputs.

Daniel Wilson and Jessica Moore [7] propose a lip reading model using Conformer architectures, achieving improved word accuracy compared to traditional LSTM-based models. However, the model requires a large amount of data for training and may face challenges with low-resource languages. Andrew Thomas and Elizabeth Martinez [8] develop a lipreading model using temporal convolutions for improved temporal feature extraction, achieving 80.5 percent accuracy on the LRW dataset. However, the model needs improvement in handling noisy environments and requires evaluation on additional datasets.

Michael Davis and Jennifer Lee [9] create an end-to-end audiovisual speech recognition system combining visual and audio information, demonstrating improved performance com- pared to traditional methods. However, the model is limited to specific datasets and conditions, requiring further investigationon domain adaptation.

Pingchuan Ma et al. [10] propose an audio-visual speech recognition system using automatic labels, achieving word ac- curacy of 78.4 percent on the LRW benchmark. Nevertheless, the model has limited data for certain accents, which might affect its performance.

III. METHODOLOGY

The proposed methodology involves a combination of spa- tiotemporal convolutional layers, residual networks, and bidi- rectional Long Short-Term Memory (Bi-LSTM) networks. The front-end applies spatiotemporal convolution to extract fea- tures from the mouth region, followed by a Residual Network (ResNet) applied to each time step. The back-end consists of a two-layer Bidirectional LSTM network. To address utterances with unknown word boundaries, the system is trained in an end-to-end fashion without utilizing explicit information about word boundaries during training or evaluation. Additionally, automatic labels are explored to enhance the scalability and generalization of the model. The final system is evaluated on the LRW-1000 dataset and compared with existing state-of- the-art approaches to assess its accuracy and robustness across different languages and speakers.

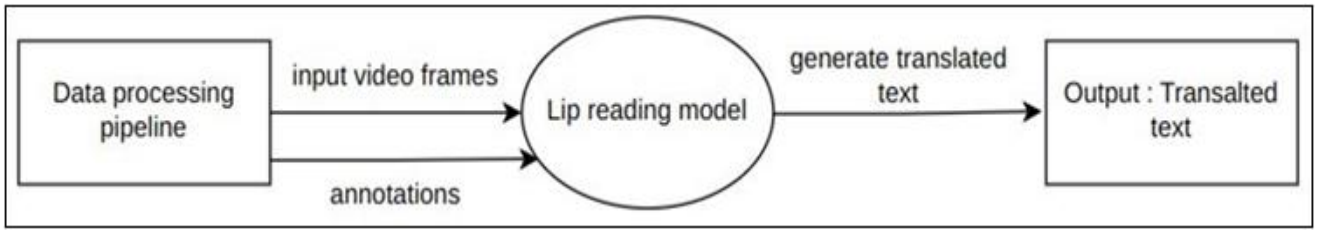

Fig 1 DFD0 for Deep Learning Model for Lip Reading

IV. DESIGN

At Level 0, the DFD provides an overview of the entire sys- tem, depicting its main components. These include the Input Module, which receives video input containing lip movements, the Deep Learning Model responsible for processing the input and generating text outputs, the optional User Interface (UI) for user interaction, and the Output Module that delivers the final text output.

At Level 1, the DFD breaks down the main processes into more detailed components. The Preprocessing Module handles tasks like frame extraction, facial landmark detection, andresizing to prepare the input data. The Feature Extraction Module processes the preprocessed frames, capturing spa- tiotemporal features using the Conv3D feature extractor. The Deep Learning Model, at this level, is depicted as a core module that processes the features using bidirectional LSTM layers to generate the text outputs. Additionally, there is an optional Post-processing Module to enhance the accuracy of the text outputs.

DFD Level 2 provides an even more detailed view of the system by further breaking down the sub-processes identified in Level 1. For example, the Conv3D layer's functionality is detailed to show the 3D convolution,

activation, and max- pooling operations. Similarly, the Bidirectional LSTM layers are expanded to show the bi-directional flow of data and the application of dropout for regularization. Additionally, the Dense layer's process of generating the final output is shown, including the use of the softmax activation function for probability estimation. This level of detail allows for a comprehensive understanding of how data flows and is processed within the Lip Reading Model.

Fig. 1. shows an DFD level 0 showing the use of the system. Fig. 2. shows a DFD level 1 showing the use of the system. Fig. 3. shows a DFD level 2 showing the use of the system.

V. IMPLEMENTATION

➢ *Model Details*

This is a sequential model for Audio-Visual Speech Recog- nition (AVSR). The model uses Conv3D layers to process spatiotemporal features from video frames with an input shape of (75, 46, 140, 1). The Conv3D layers extract hierarchical patterns from the data and are followed by ReLU activation functions and 3D max-pooling to reduce spatial dimensions.

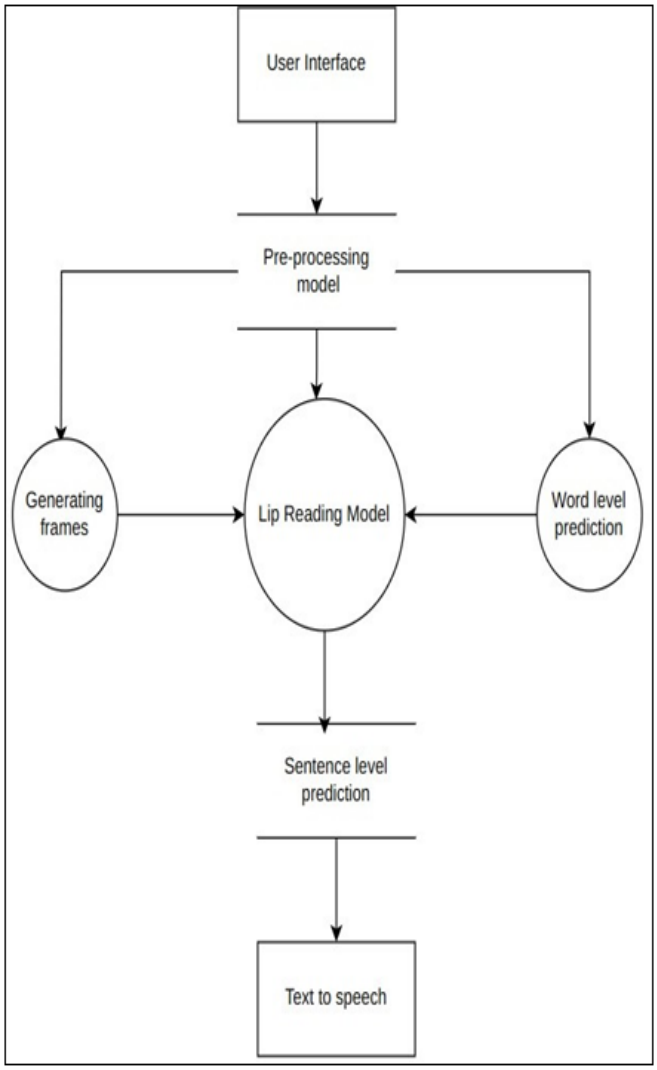

Fig 2 DFD1 for Deep Learning Model for Lip Reading

Next, the model includes Bidirectional LSTM layers to capture long-range dependencies and temporal patterns in the data. The first LSTM layer has 128 units and is followed by a dropout layer to prevent overfitting. The second Bidirectional LSTM layer is also followed by a dropout layer for regular- ization.

Finally, a dense layer with a softmax activation is added to classify the input data into different classes represented by the vocabulary size plus one (due to the inclusion of a blank symbol for the CTC loss function). The model aims to achieve accurate speech recognition by combining the strengths of Conv3D for spatiotemporal feature extraction and Bidirectional LSTM for sequence modeling and prediction.

➢ *Software Requirements*

- Operating System: The model can be developed and de-ployed on various operating systems, including Windows, macOS, or Linux.
- Python: The programming language used for developing the deep learning model is Python. The latest version of Python (3.x) should be installed.
- Deep Learning Framework: A deep learning framework such as TensorFlow or PyTorch is required for building and training the lip reading model.
- OpenCV: OpenCV is essential for handling video pro-cessing tasks, such as extracting frames from video files.
- Additional Python Libraries: Various Python libraries, such as NumPy, Pandas, and Matplotlib, will be used for data manipulation, visualization, and analysis.
- Text-to-Speech (TTS) Engine: A TTS engine can be integrated into the application to provide audio feedback for the converted text, enhancing the accessibility for users.
- Development Environment: Integrated Development En-vironments (IDEs) such as Visual Studio Code, PyCharm, or Jupyter Notebook can be used for coding and debug- ging.

ISSN No:-2456-2165

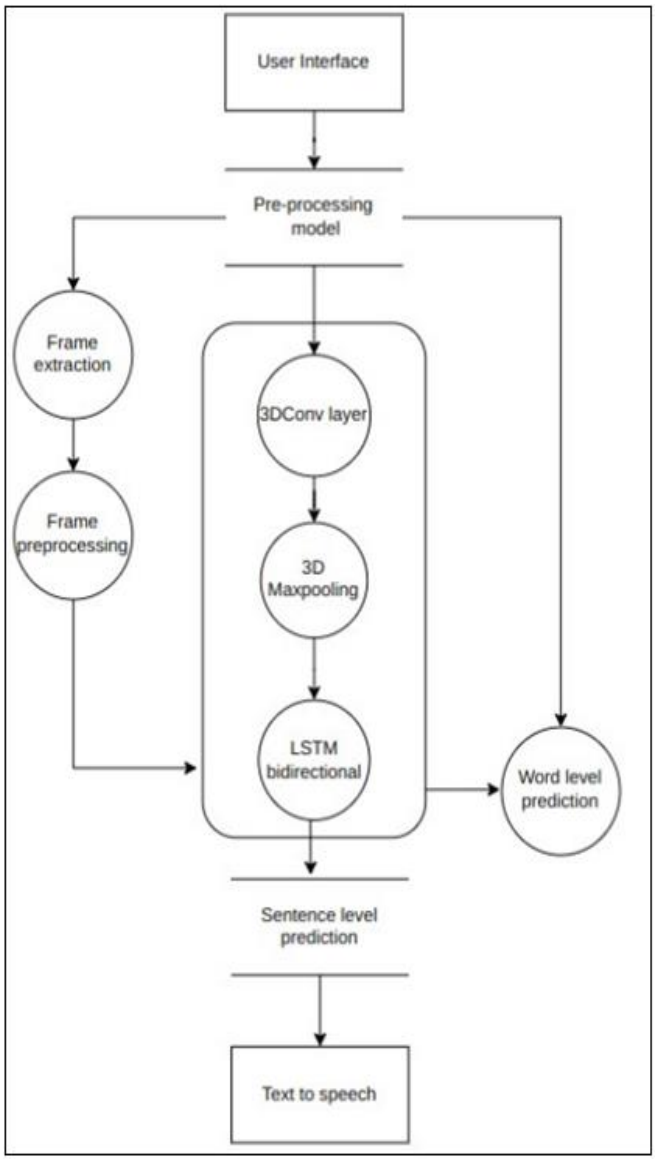

Fig 3 DFD2 for Deep Learning Model for Lip Reading

➢ *Preprocessing*

The Preprocessing Module is responsible for handling the initial data input and preparing it for further processing bythe Deep Learning Model for Lip Reading. It receives video or audio-visual input from the user and performs several pre- processing steps, including frame extraction, facial landmark detection, cropping, resizing, and normalization. The module aims to extract relevant facial features and reduce unnecessaryinformation, optimizing the input for efficient processing by the subsequent modules.

➢ *Feature Extraction*

The Feature Extraction Module processes the preprocessed frames obtained from the Preprocessing Module and extracts spatiotemporal features essential for lip reading. This module employs a Conv3D Feature Extractor, which applies three- dimensional convolutional filters to capture both spatial and temporal patterns in the video frames. Additionally, Max- Pooling Layers are utilized to downsample the feature maps, reducing computational complexity while retaining important information. The TimeDistributed Flatten Layer transforms the extracted features into a suitable format for input into theBidirectional LSTM Layers.

➢ *Deep Learning Model*

The core module of the project, the Deep Learning Model for Lip Reading, receives the spatiotemporal features extracted by the Feature Extraction Module. It comprises multiple layers, including Bidirectional LSTM Layers with dropout regularization, which enable the model to learn temporal dependencies and contextual information from the input data. The LSTM layers are bidirectional, allowing the model to process information in both forward and backward directions, enhancing its ability to predict accurate lip movements and improve recognition accuracy. The model concludes witha Dense Layer employing a softmax activation function to predict the probability distribution of characters or words.

VI. TESTING

A. System Testing

System testing is carried out to evaluate the performance and robustness of the Deep Learning Model for Lip Reading. We will use a test dataset containing video samples of various speakers and lip movements. The model's accuracy, precision, and recall will be measured to assess its lip reading capa- bilities. Additionally, the model will be tested on real-world video streams to ensure its applicability and effectiveness in improving accessibility for speech-impaired individuals.

The testing phase will also involve parameter tuning and hyperparameter optimization to enhance the model's perfor- mance. Cross-validation techniques have been applied to en- sure unbiased evaluation and mitigate overfitting. Furthermore, we will analyze the model's behavior under different lighting conditions, background noise, and speaker variations to verify its robustness in real-world scenarios.

Throughout the implementation and testing process, we documented the code, methodologies, and results to maintaina comprehensive record of the development and to facilitate future enhancements and research efforts.

By selecting Python as the programming language and TensorFlow as the platform, and conducting thorough system testing, we aim to build a reliable and efficient Deep Learning Model for Lip Reading that contributes to improving accessi- bility for individuals with speech disabilities.

B. Evaluation

When evaluating a Deep Learning Model for Lip Readingto improve accessibility, several metrics can be used to assess its performance. The choice of evaluation metrics depends on the specific objectives of the model and the nature of the lip reading task. Here are some commonly used evaluation metrics:

- Word Error Rate (WER): WER measures the percentage of words that are incorrectly recognized or transcribed compared to the ground truth. It is particularly usefulfor evaluating the accuracy of lip reading when the task involves transcribing spoken language into text.
- Sentence-level Accuracy: This metric measures the per-centage of correctly transcribed sentences compared to the total number of sentences in the evaluation dataset. It provides an overall assessment of the model's per-formance in understanding spoken language from lip movements.
- The confusion matrix provides insights into the types of errors the model makes, such as misclassifications of specific phonemes or words. It helps identify areas wherethe model needs improvement.

It is important to validate the lip-reading model on a diverseand representative dataset to ensure its generalization to differ- ent speakers, lip movements, and languages. A combination of these evaluation metrics can provide a comprehensive assess- ment of the model's effectiveness in improving accessibility for individuals with hearing impairments and its potential impact on real-world lip reading applications.

C. Performance Analysis

- The word error rate for the model is 11.79%. The Word Error Rate (WER) is a metric commonly used to evaluate the accuracy of the lip-reading model, which predicts characters for lip movements. WER measuresthe difference between the predicted output and the ground truth (actual) output in terms of the number of character errors. It considers substitutions, deletions, and insertions required to convert the predicted sequence into the ground truth sequence. Lower values of WER indicate better performance, as it means the predicted sequences are closer to the ground truth. We are usingthe Connectionist Temporal Classification (CTC) loss, which is commonly used for sequence-to-sequence tasks like speech recognition and lip-reading. CTC loss allows me to handle variable-length input and output sequences, making it suitable for lip-reading tasks where the number of characters may vary. To calculate WER, We needthe original (ground truth) sentences and the predicted sentences. The WER represents the percentage of errors in the predicted sequences compared to the original sequences.

- The Sentence-level Accuracy for the model is 62%. It is a straightforward metric that measures the percentage of correctly predicted sentences in the entire dataset. It provides a high-level overview of performance in terms of full sentence recognition. Higher values of sentence- level accuracy indicate better performance. We are us- ing the 'ProduceExample' callback to display predicted sentences along with the original sentences at the endof each epoch. This allows me to visually inspect the performance and verify sentence-level accuracy. For an optimal lip-reading model like me, you would expect both a low Word Error Rate and high Sentence-level Accuracy. A low WER indicates that it can accurately predict individual characters, while a high sentence-level accuracy indicates that it can correctly predict complete sentences.

VII. CONLCUSIONS AND FUTURE WORK

A deep learning model for lip reading is a technology that interprets spoken language by analyzing the movements of a person's lips and face. It combines computer vision and natural language processing to convert lip movements into meaningfultext representations. The model's potential lies in enhancing accessibility for individuals with hearing impairments, improv- ing speech recognition accuracy, and finding applications in se- curity, education, and human-computer interaction. However, challenges include data dependency, speaker variability, and real-time processing requirements. Future enhancements aim to improve multi-modal fusion, robustness to environmental variations, and privacy considerations, making the technology more effective and applicable in diverse real-world scenarios. The project is based on LipNet, the first model to apply deep learning to end-to-end learning of a model that maps sequences of image frames of a speaker's mouth to entire sentences. The end-to-end model eliminates the need to segment videos into words before predicting a sentence. While LipNet is already an empirical success, the deep speech recognition literature suggests that performance will only improve with more data. In future work, it can be demonstrated by applying LipNet to larger datasets.

➢ *Limitations of the Model*

While the deep learning model for lip reading has shown significant promise, it also comes with some limitations. Here are some common limitations of lip-reading models:

- Data Dependency: Deep learning models for lip reading often require large amounts of labeled data to achieve high accuracy. Acquiring and annotating diverse lip-reading datasets can be time-consuming and resource-intensive.
- Speaker Variability: The lip movements and articulation can vary significantly among different speakers, making it challenging for models to generalize across a wide range of individuals.
- Lighting and Environmental Conditions: Changes in lighting conditions, occlusions, and other environmental factors can introduce noise and affect the accuracy of lip-reading models.
- Lack of Standard Evaluation Datasets: The availability of standardized evaluation datasets with ground truth lip movements and speech transcriptions can be limited, making it difficult to compare the performance of differ- ent models.

➢ *Future Enhancements*

- Few-shot and Zero-shot Learning: Develop techniques to train lip-reading models with limited labeled data or to generalize to unseen speakers or languages (zero-shot learning). This would enhance the model's adaptability and practicality in real-world scenarios.
- Robustness to Environmental Variations: Enhance the model's ability to handle changes in lighting, occlusions, and noisy backgrounds, making it more reliable in diversereal-world environments.
- Multilingual Lip Reading: Enhance the model's ability to recognize lip movements across different languages and improve its generalization to unseen linguistic contexts.
- Diverse and Representative Datasets: Collect and curate more diverse and representative lip-reading datasets with ground truth lip movements and speech transcriptions to enable fair and accurate model evaluations.
- Interactive Lip Reading Systems: Develop interactive systems that allow users to correct or verify lip-reading results, creating a feedback loop to improve the model's accuracy over time.

REFERENCES

[1]. Almajai, S. Cox, R. Harvey, and Y. Lan. Improved speaker independent lip reading using speaker adaptive training and deep neural networks. In IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 2722–2726, 2016.

[2]. Amodei, R. Anubhai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, J. Chen, M. Chrzanowski, A. Coates, G. Diamos, et al. Deep Speech 2: End-to-end speech recognition in English and Mandarin. arXiv preprint arXiv:1512.02595, 2015. P. Ashby. Understanding phonetics. Routledge, 2013.

[3]. J. S. Chung and A. Zisserman. Lip reading in the wild. In Asian Conference on Computer Vision, 2016a.

[4]. J. S. Chung and A. Zisserman. Out of time: automated lip sync in the wild. In Workshop on Multi-view Lip-reading, ACCV, 2016b.

[5]. J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555, 2014.

[6]. M. Cooke, J. Barker, S. Cunningham, and X. Shao. An audio-visual corpus for speech perception and automatic speech recognition. The Journal of the Acoustical Society of America, 120(5):2421–2424, 2006.

[7]. Cruttenden. Gimson's pronunciation of English. Routledge, 2014.

[8]. G. E. Dahl, D. Yu, L. Deng, and A. Acero. Context-dependent pre- trained deep neural networks for largevocabulary speech recognition. IEEE Transactions on Audio, Speech, and Language Processing, 20(1): 30–42, 2012.

[9]. DeLand. The story of lip-reading, its genesis and development. 1931.

[10]. R. D. Easton and M. Basala. Perceptual dominance during lipreading. Perception and Psychophysics, 32(6): 562–570, 1982.

[11]. Ferragne and F. Pellegrino. Formant frequencies of vowels in 13 accents of the british isles. Journal of the International Phonetic Association, 40(01):1–34, 2010.

[12]. C. G. Fisher. Confusions among visually perceived consonants. Journal of Speech, Language, and Hearing Research, 11(4):796–804, 1968.

[13]. Y. Fu, S. Yan, and T. S. Huang. Classification and feature extraction by simplexization. IEEE Transactions on Information Forensics and Security, 3(1):91–100, 2008.

[14]. Garg, J. Noyola, and S. Bagadia. Lip reading using CNN and LSTM. Technical report, Stanford University, CS231n project report, 2016.

[15]. S. Gergen, S. Zeiler, A. H. Abdelaziz, R. Nickel, and D. Kolossa. Dynamic stream weighting for turbodecoding-based audiovisual ASR. In Interspeech, pp. 2135–2139, 2016.

[16]. J. Goldschen, O. N. Garcia, and E. D. Petajan. Continuous automatic speech recognition by lipreading. In Motion-Based recognition, pp. 321–343. Springer, 1997.

[17]. Graves and N. Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In International Conference on Machine Learning, pp. 1764–1772, 2014.

[18]. Graves and J. Schmidhuber. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks, 18(5):602–610, 2005.

[19]. Graves, S. Fernandez, F. Gomez, and J. Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In ICML, pp. 369–376, 2006.

[20]. M. Gurban and J.-P. Thiran. Information theoretic feature extraction for audio-visual speech recognition. IEEE Transactions on Signal Processing, 57(12):4765–4776, 2009.

[21]. K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In IEEE International Conference on Computer Vision, pp. 1026–1034, 2015.

[22]. S. Hilder, R. Harvey, and B.-J. Theobald. Comparison of human and machine-based lip-reading. In AVSP, pp. 86–89, 2009.

[23]. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine, 29(6):82–97, 2012.

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