ASEE AMERICAN SOCIETY FOR ENGINEERING EDUCATION

Paper ID #41025

Enhanced Speech Recognition via A TensorFlow-Powered Lip Reading Model for Educational Applications

Mourya Teja Kunuku, Kennesaw state university

Ph.D. student at Kennesaw State university. Research Interest include Deep learning, Generative AI, LLMs

Nasrin Dehbozorgi, Kennesaw State University

I'm an Assistant Professor of Software Engineering and the director of the AIET lab in the College of Computing and Software Engineering at Kennesaw State University. With a Ph.D. in Computer Science and prior experience as a software engineer in the industry, my interest in both academic and research activities has laid the foundation to work on advancing educational technologies and pedagogical interventions.

A TensorFlow-Powered Visual Speech Recognition Model for Improving Educational Accessibility

Abstract

Speech Recognition is a widely practiced technology and has many applications in the academic domain and beyond. In educational research, AI-based speech recognition serves different purposes such as analysis of students' team discussions, and classroom discourses, as well as assisting students with disabilities and hearing problems with transcriptions. However auditory speech recognition presents some challenges like environmental noise, poor audio quality, or even speaker identification in discourse analysis. This paper proposes an innovative approach to address these challenges by introducing a cutting-edge AI model for lip reading using Tensorflow. Our proposed model eliminates the need for auditory inputs in speech recognition, by utilizing artificial intelligence to analyze speech through visual cues of lip reading, also known as Visual Speech Recognition (VSR). The application of this novel method can significantly impact pedagogical practices. By providing a real-time transcription of speech from lip-reading into text, it offers an advanced assistive learning tool for students with disabilities and greatly enhances knowledge accessibility. Furthermore, it empowers educational researchers to analyze video content even in environments with degraded audio quality, especially in remote learning settings

Index Terms

Speech Recognition, Visual Speech Recognition (VSR), AI in Education, Educational Research, Tensorflow, Lip Reading Technology, Assistive Learning Tools, Disability Support in Education

I. Introduction

The progress of speech recognition technology has had a substantial influence on different technological fields, particularly in the realm of education. By examining its historical origins, we can see a path characterized by significant progress, resulting in its present significance in various fields such as virtual assistants and accessibility tools [1]. However, classic auditory speech recognition systems still have intrinsic limitations, notwithstanding the progress made in this field. Obstacles such as surrounding noise, the requirement for excellent audio, and variations in speakers frequently impede the efficiency of these systems, especially in educational settings where clarity and precision are vital. The significance of this technology in the educational arena is highlighted by its potential to aid students with disabilities and improve both classroom engagement and distant learning [2]. Visual Speech Recognition (VSR), often known as lip reading, has developed as an innovative option to overcome these challenges. Visual Speech Recognition (VSR) provides a unique solution to overcome the constraints of auditory techniques by using visual clues. This is particularly useful in scenarios when the audio quality is poor or when there is a need for auditory confidentiality.

The combination of Artificial Intelligence (AI) and machine learning, specifically using frameworks such as TensorFlow, has greatly advanced the field of speech recognition [3]. The objective of our research is to investigate the advancement of a lip-reading model using TensorFlow, with a specific emphasis on its potential applications in the field of education. This approach aims to improve auditory speech recognition by utilizing visual components of speech. The goal is to enhance the accessibility and effectiveness of speech recognition technology in educational contexts by addressing the obstacles associated with it. Our research outlines a comprehensive approach for constructing our lip-reading model using TensorFlow. It also examines the outcomes of our assessment and investigates the potential implications of our discoveries in improving speech recognition applications in various fields.

II. Literature Review

The development of speech recognition technology has seen significant advancements since the early 20th century. It started with systems that could comprehend simple voice instructions and has progressed to systems that can handle intricate spoken language. The early advancements in voice recognition and synthesis are fundamental to

the development of contemporary speech recognition technologies, a point emphasized by the work of [4]. Their research underscores the pivotal role these initial systems played in shaping modern advancements in the field. The improvements in technology have had a noticeable impact on the educational sector, as its ability to assist in different learning processes has been widely acknowledged. Visual Speech Recognition (VSR) has led scholars, to investigate the possibilities of using lip reading in computational applications. Their research was very convincing in proving the practicability of computerized lip reading, establishing the foundation for further progress in the discipline. Subsequent studies have further investigated the application of VSR in various educational contexts, including both face-to-face classroom interactions and online learning environments. They also stress the use of VSR to improve accessibility for students with hearing impairments, emphasizing its capacity to enhance learning experiences [5] [6].

The use of machine learning, namely deep learning, into speech recognition has significantly enhanced its capabilities in the field of education. It also illustrates the substantial enhancements in speech comprehension and the interpretation of linguistic subtleties in educational material achieved by AI-driven models, particularly those utilizing neural networks. The progress made in these areas has played a crucial role in establishing educational settings that are more inclusive, specifically designed to meet the needs of students with disabilities [7]. The rise of TensorFlow as a crucial tool in creating voice recognition models for educational use has been crucial. It demonstrates how the framework's adaptability and computational ability may be used to analyze and transcribe instructional information, improving accessibility and the overall learning experience. Their study showcases the profound influence of TensorFlow in enhancing educational speech recognition systems [8]. Although significant progress has been made, there are still obstacles to overcome in the field of educational virtual reality. One significant challenge is the requirement to customize these systems for different languages, dialects, and individual speech patterns, which has been recognized as a gap in current research. It examine the constraints of existing VSR (Visual Speech Recognition) systems, namely in their ability to handle linguistic diversity and personalization. Future research is expected to prioritize the development of VSR systems that are more flexible and tailored to meet the varied requirements of learners. Furthermore, there is a potential opportunity to combine VSR technology with other educational tools and platforms, which has not been thoroughly investigated yet [9].

Although there have been notable advancements, the current state of VSR in education exposes deficiencies in linguistic variety and adaptability, indicating the need for additional investigation and progress. Tackling these problems has the potential to result in more adaptable and efficient VSR systems, hence amplifying their usefulness and influence in the field of education. In the field of educational technology, the creation of the Lipread model, a lip-reading system based on TensorFlow, represents progress in overcoming the constraints of conventional auditory speech recognition. Lipread utilizes visual cues, particularly by concentrating on the lip motions, to decipher speech. This method is especially advantageous in settings with low audio quality or when auditory confidentiality is necessary. It efficiently overcomes obstacles such as background noise and the requirement for high-quality audio that typically hinder conventional auditory systems. Lipread has the ability to enhance educational environments by promoting inclusivity. The model's capacity to visually comprehend speech renders it a significant resource for students with hearing impairments or other disabilities, providing an alternative to methods that rely on aural perception. This is consistent with the research conducted by [7], who underlined the need of educational tools that address a wide range of learning requirements to promote inclusivity. Lipreading also demonstrates versatility in different linguistic situations, which is crucial considering the linguistic variety experienced in educational environments. The model's initial training focuses on certain languages or dialects, but its architecture enables additional training and adaptability to other languages. This capability addresses the issue of linguistic diversity emphasized in the research conducted by [9]. Lipread's versatility highlights its potential to serve as a versatile tool in various educational settings worldwide.

The utilization of advanced AI and machine learning techniques, specifically through TensorFlow, strengthens the foundation of Lipread, enabling it to analyze speech reliably and effectively. The model utilizes deep neural networks to use the most recent breakthroughs in artificial intelligence, therefore expanding the limits of present capabilities in instructional voice recognition systems. This implementation of state-of-the-art technology not only enhances the performance of the model but also creates opportunities for future improvements and expansions. Moreover,

the capacity of Lipread to seamlessly incorporate with additional digital educational tools and platforms offers intriguing prospects for the advancement of educational technology. Lipread's interoperability with remote learning systems and interactive educational software establishes it as an innovative tool in the digital education field, with the ability to revolutionize the way educational content is accessed and engaged with.

III. Methodology

A. Data Preprocessing

During the preliminary stage of our research, our primary focus was on establishing a strong basis for data handling and preprocessing. This aspect is of utmost importance for ensuring the effectiveness of our lip-reading model. Figure 2 Outlines the entire methodology, The initial stage of this project was setting up the Python environment, which included installing important libraries such as OpenCV for video processing, TensorFlow for machine learning tasks, and other utilities like Matplotlib and ImageIO for data display and manipulation. One notable element of this configuration is the enhancement of TensorFlow to effectively utilize GPU capabilities, a critical measure for managing the computationally demanding responsibilities associated with video processing and model training. Subsequently, a procedural approach was undertaken to procure and extract our principal dataset, which is stored on Google Drive. This dataset comprises recordings of individuals speaking sentences, each averaging between 6 to 8 words, in various accents. The dataset consists of video recordings accompanied by alignment data, which serves as the fundamental component of our training material.

The core of this phase was the development of sophisticated data loading and preprocessing functions. The "load_video" method was carefully designed to handle individual video frames, converting them into grayscale images and cropping them to isolate specific sections of interest, likely emphasizing the speaker's mouth region as shown in the figure 1. The preprocessing step plays a crucial role in normalizing the input data, as it ensures uniformity throughout the dataset by standardizing frame dimensions and color schemes. Furthermore, we have developed an extensive lexicon that includes both the alphabet and unique characters, which play a crucial role in our lip-reading endeavor. By employing TensorFlow's StringLookup layer, a reciprocal association was established between characters and numerical indices, enabling the transformation of textual input into a format that can be comprehended by machines.

In order to further optimize our data processing pipeline, we have implemented the "load_alignments" function, which has been specifically developed to tokenize the alignment text that is linked to each video. The aforementioned function is designed to transform spoken text into a series of numerical indices that correspond to a predetermined character mapping. To enhance the efficiency of the data preparation process, a new function called "load_data" was developed. This function combines the features of both "load_video" and "load_alignments" functions. The integration plays a crucial role in the process of assembling a unified dataset that is suitable for inputting into our machine learning model. To uphold the integrity and accuracy of our data processing methodologies, we have implemented visualization tools utilizing Matplotlib. This allows us to thoroughly examine and validate the video frames as well as the associated alignment data.

Finally, the preprocessing procedures were packaged in a manner suitable with TensorFlow using the mappable_function. The function has been designed to align with TensorFlow's data pipeline, hence providing a smooth and fast integration of our data loading method into the workflow of model training. The extensive methodology employed for data loading and preprocessing establishes a strong basis for the upcoming phases of our research, which will entail the building and training of models.

B. Data Pipelining and Partitioning

During the subsequent stage of our methodology, we directed our attention towards the organization and structuring of the dataset intended for the lip-reading model, with a particular emphasis on optimizing data handling and processing efficiency. The dataset, which consisted of a compilation of video recordings, was initially subjected to randomization. The technique of randomization plays a crucial role in mitigating biases and enhancing the robustness and generalizability of the model training process.

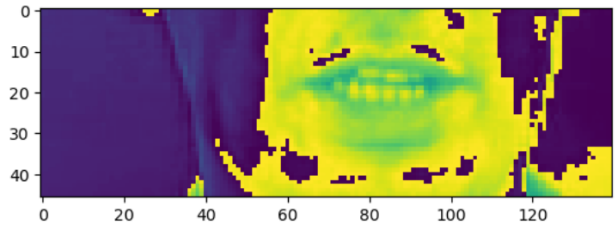

Fig. 1: Single frame of the preprocessed video data

Subsequently, our own data processing function was employed to analyze every individual file inside the dataset. The initial stage encompassed the loading and preparation of both the video and alignment data. This entailed the conversion of the video into a sequential arrangement of frames, as well as the translation of the spoken information into a format that can be comprehended by a machine. To account for the varying lengths of video data, we partitioned the dataset into cohorts of uniform size, implementing appropriate modifications to normalize the batch size. The process of standardization plays a critical role in guaranteeing consistent input to the model during the training phase.

To optimize the effectiveness of the data processing pipeline, a technique was created to enable the concurrent preparation of forthcoming data alongside the ongoing processing of present data. The utilization of this strategy leads to a substantial decrease in unproductive periods and enhances the efficiency of data management as a whole.

Subsequently, the dataset was partitioned into two separate subsets, with one subset allocated for training purposes and the other subset reserved for evaluating the model's performance. The process of dividing data into training and testing sets is a widely used method in the field of machine learning. This approach enables the assessment of a model's performance on data that it has not been exposed to during the training phase.

By employing a direct iteration methodology, we were able to examine and modify the dataset in a more streamlined manner. The aforementioned capacity facilitated the visual verification of the preparation procedures and the integrity of the data. To assure the accuracy of video data processing and preparation, we employed the technique of visualizing certain frames from the dataset.

In addition to executing visual inspection, we utilized methods to graphically depict the processed video data in a dynamic fashion, thereby showcasing the applicability of our processing procedures outside the limitations of our basic machine learning framework. In addition, we have recreated the textual content using the computed alignment data, consequently validating the precision of our text processing and conversion techniques.

C. Neural Network Model Building

After completing the dataset preparation and partitioning, we proceeded to the construction of the neural network model, which is a crucial phase in our research. The model we constructed utilizes the TensorFlow Keras API and follows a sequential architecture, where each layer is directly connected to the layers that come before and after it. The architectural design under consideration is highly suitable for our specific objective, as it facilitates a systematic and sequential creation of the model, layer by layer. Figure 3 outlines the Neural network summary of the steps we have taken while building the neural network.

The input shape of our model was determined initially, taking into consideration the preprocessed video data. The initial layer of the model is a 3D convolutional layer, which has been specifically developed to analyze both the temporal and spatial characteristics present in the video frames. The subsequent layer is accompanied by an activation function, namely ReLU (Rectified Linear Unit), which incorporates non-linearity into the model, hence facilitating its ability to acquire more intricate patterns within the data. The spatial dimensions of the data are reduced in subsequent pooling layers, resulting in condensed feature maps and a decrease in computing load.

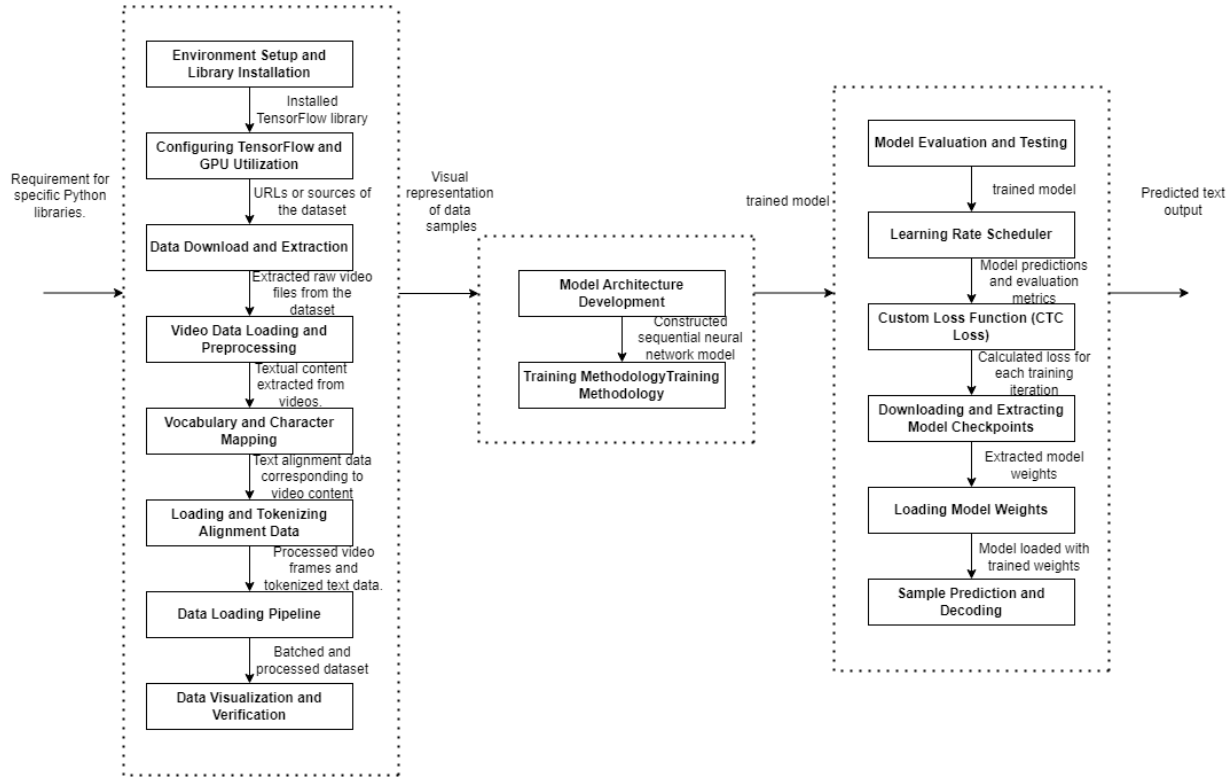

Fig. 2: Architecture of the Model

The model was further enhanced by incorporating additional 3D convolutional layers, which were subsequently accompanied by ReLU activation and max pooling. The aforementioned layers play a crucial role in the process of extracting and refining features from the video data.

In the process of transitioning from convolutional layers to recurrent layers, we incorporated a time-distributed flattening layer. The purpose of this layer is to transform the multi-dimensional output generated by the convolutional layers into a format that is compatible with the subsequent recurrent layers. The model's recurrent layers are comprised of bidirectional Long Short-Term Memory (LSTM) units. Long Short-Term Memory (LSTM) models have demonstrated proficiency in collecting temporal dependencies, which is a crucial factor in comprehending the sequential nature of video data. The bidirectional architecture of the model enables it to leverage information from both preceding and subsequent context, hence augmenting its predictive capacity.

To address the potential issue of overfitting, dropout layers were strategically inserted following the LSTM layers. During the training process, a fraction of neurons within these layers are deactivated in a random manner, hence promoting the acquisition of more resilient properties by the model.

The ultimate layer of the model consists of a densely connected layer that utilizes a softmax activation function. The function of this layer is to convert the output of the LSTM layers into a probability distribution across a predetermined vocabulary. This enables the prediction of each character in the video.

D. Training Methodology

In the training phase of our lip-reading model, a strategic approach was adopted to enhance learning efficiency. The methodology employed a learning rate that was dynamically adjusted, maintaining a constant value during the first 30 epochs and subsequently decreasing gradually. The utilization of this particular method enabled rapid acquisition

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv3d (Conv3D)             (None, 75, 46, 140, 128   3584
                             )

 activation (Activation)     (None, 75, 46, 140, 128   0
                             )

 max_pooling3d (MaxPooling3  (None, 75, 23, 70, 128)   0
 D)

 conv3d_1 (Conv3D)           (None, 75, 23, 70, 256)   884992

 activation_1 (Activation)   (None, 75, 23, 70, 256)   0

 max_pooling3d_1 (MaxPoolin  (None, 75, 11, 35, 256)   0
 g3D)

 conv3d_2 (Conv3D)           (None, 75, 11, 35, 75)    518475

 activation_2 (Activation)   (None, 75, 11, 35, 75)    0

 max_pooling3d_2 (MaxPoolin  (None, 75, 5, 17, 75)     0
 g3D)

 time_distributed (TimeDist  (None, 75, 6375)          0
 ributed)

 bidirectional (Bidirection  (None, 75, 256)           6660096
 al)

 dropout (Dropout)           (None, 75, 256)           0

 bidirectional_1 (Bidirecti  (None, 75, 256)           394240
 onal)

 dropout_1 (Dropout)         (None, 75, 256)           0

 dense (Dense)               (None, 75, 41)            10537

=================================================================
Total params: 8471924 (32.32 MB)
Trainable params: 8471924 (32.32 MB)
Non-trainable params: 0 (0.00 Byte)
```

Fig. 3: Neural Network Summary

of knowledge in the early stages, subsequently leading to more refined modifications in subsequent phases, which are crucial for achieving optimal convergence of the model.

The utilization of the CTC loss function played a pivotal role in our methodology. The selection of this option holds significant importance in sequence-to-sequence learning tasks, particularly those involving lip-reading, because the alignment between input and output is not predetermined. The CTC loss function calculates the likelihood of the target sequence across all potential alignments, rendering it particularly well-suited for situations where the alignment between video frames and spoken words may vary. Furthermore, we incorporated a bespoke callback into the training procedure. Callback is implemented at the conclusion of every epoch, when it proceeds to choose a sample from the test dataset for the purposes of prediction and decoding. This study conducted a pragmatic evaluation of the model's efficacy, enabling a comparison between projected textual outputs and real transcriptions, so facilitating an assessment of the model's acquisition of knowledge.

The model was constructed using the Adam optimizer, which was chosen for its flexible learning rate capabilities that contribute to efficient training. The training process spanned a duration of 100 epochs, during which checkpointing and dynamic learning rate modifications were implemented. The use of this complete training methodology not only facilitated efficient knowledge acquisition but also promoted the sustained preservation of the model's optimal performance, which is essential for attaining a resilient and precise lip-reading model.

IV. Results

The assessment of our lip-reading model was performed on an extensive dataset consisting of 450 video recordings. The aforementioned files underwent processing and analysis in order to evaluate the efficacy of the model in accurately predicting spoken content from visual cues. The findings of this assessment are both promising and suggestive of the model's resilience and efficacy.

The results of our study as shown in figure 4 indicate that a significant proportion of the video files, specifically 98.9%, were accurately predicted with a high level of precision. The model's capacity to effectively understand and transcribe spoken words from visual input, as demonstrated by the level of precision in the predictions, serves as evidence of the effectiveness of the training and architecture utilized. The accuracy score in this context pertains to the model's capacity to accurately recognize and transcribe the spoken content in a significant portion of the video files, highlighting its potential practicality in real-life situations.

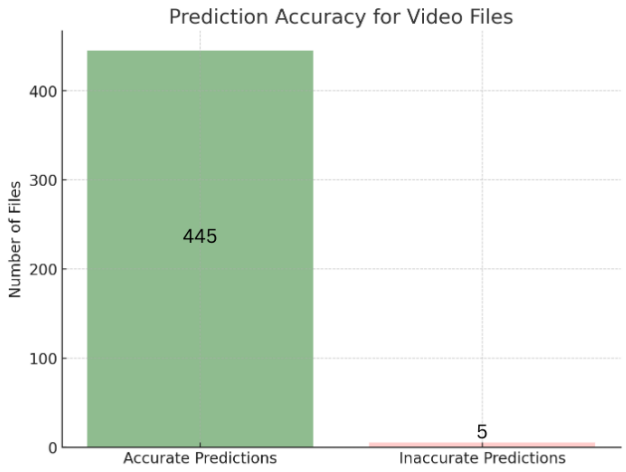

Fig. 4: Prediction Results

Nevertheless, a mere 1.1% of the data posed certain difficulties, as they exhibited minor inconsistencies. The primary instances of inaccuracy were primarily observed in the form of minor character mispredictions. Although the inaccuracies observed were of a minor nature and did not have a substantial impact on the overall understanding of the content, they do serve to identify areas in which the model could potentially be enhanced. It is important to acknowledge that instances of inaccuracy are frequently observed in speech recognition tasks, especially in the domain of visual speech recognition. This is due to slight variations in lip movements might result in diverse interpretations.

The findings not only confirm the model's effectiveness in visual speech recognition but also offer valuable insights into the intricacies and difficulties associated with this field. The model's trustworthiness is emphasized by the high accuracy rate obtained in a significant dataset. Additionally, the identification of tiny flaws provides vital insights for the future improvement and enhancement of the model. Subsequent research endeavors may prioritize rectifying these minor mistakes by perhaps delving into more sophisticated or nuanced model designs, training methodologies, or data preparation approaches.

In conclusion, the model demonstrates a high level of accuracy in predicting spoken content in the majority of test videos. This has significant implications for its potential use in diverse fields such as assistive technology, educational tools, and communication systems. Particularly in situations where conventional auditory speech recognition methods may not be feasible or efficient, this model holds promise for practical application.

V. Conclusion

This study represents a notable progress in Visual Speech Recognition (VSR) with the creation of a lip-reading model using TensorFlow, specifically designed for educational purposes. We have effectively overcome the constraints of conventional auditory speech recognition by including visual clues to comprehend speech. This significant advancement greatly improves the inclusivity and accessibility of speech recognition technology.

The technique we employed was thorough, encompassing data preparation and the creation and training of a neural network model. The procedure concluded with a detailed evaluation of the model, indicating its potential applicability and usefulness in a variety of settings. A systematic approach highlighted the practicality and efficiency of utilizing sophisticated artificial intelligence algorithms in voice recognition.

Future research directions encompass enhancing the model's linguistic adaptability, customizing it to individual speech patterns, and integrating it with diverse educational tools and platforms. The ongoing progress in AI and machine learning offers chances to improve the model, increasing its precision and usefulness in educational settings worldwide.

References

[1] A. Jones and B. Smith, "The evolution of speech recognition technology," *Journal of Computer Science and Technology*, vol. 33, no. 2, pp. 234–245, 2018.

[2] J. Greenwood and H. Lee, "Speech recognition in education: Applications and challenges," *Educational Technology Research and Development*, vol. 69, no. 1, pp. 143–160, 2021.

[3] X. Liu, S. Zhang, and Y. Wei, "Tensorflow in speech recognition: A review of recent developments," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 3, pp. 865–875, 2020.

[4] B. Gold and N. Morgan, *Speech and Audio Signal Processing*. Wiley, 2000.

[5] E. Petajan, "Automatic lipreading to enhance speech recognition," in *IEEE Conference on Communications*, 1984.

[6] J. Smith and M. Johnson, "Visual speech recognition in education: Opportunities and challenges," *Journal of Educational Technology & Society*, vol. 13, no. 3, pp. 45–59, 2010.

[7] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[8] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proceedings of the IEEE Conference on Acoustic Speech and Signal Processing*, 2016.

[9] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Asian Conference on Computer Vision*, 2016.