Accents & Agnostic Speech Transcription Model for Enhanced User Accessibility

***Abstract*—In the digital learning environment, accent diversity can hinder comprehension and accessibility, posing a significant barrier for non-native English speakers and individuals with hearing impairments. This project addresses the challenge of accent variability in educational content by developing an accent-agnostic speech transcription model. Utilizing the Mel-frequency cepstral coefficients (MFCC) for feature extraction and exploring Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU) models, we aimed to enhance speech recognition accuracy across diverse accents. The models were trained and tested using a dataset from Mozilla's Common Voice project, adjusted to represent various global English accents. Our results indicate that the GRU model outperformed others, achieving the highest accuracy and demonstrating substantial potential for reducing the educational accessibility gap caused by accent bias. This project not only advances the field of speech recognition but also contributes significantly to making e-learning platforms more inclusive and accessible.**

***Index Terms*— e-learning, speech transcription, accent recognition, Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTMs), Gated Recurrent Units (GRUs), Mel-frequency cepstral coefficients (MFCCs), speech recognition, neural networks, machine learning in education.**

I. Introduction

The rapid evolution of technology in recent years has significantly transformed the educational landscape, facilitating the emergence and growth of e-learning platforms. These platforms offer learners from diverse geographical, economic, and cultural backgrounds unprecedented access to educational resources. However, as digital education becomes more widespread, the need for inclusivity within these platforms becomes increasingly critical. One of the major hurdles to achieving this inclusivity is the challenge posed by the diversity of accents in speech recognition systems used within these platforms.

Speech recognition technology is pivotal in making learning content accessible, especially through features like automated subtitles and voice-driven commands. However, traditional speech recognition systems often perform optimally on a narrow range of accents—typically those that are well-represented in their training datasets. This creates a significant accessibility issue for users with accents that deviate from the dataset norm, including non-native English speakers and regional dialects, who may find that these systems misinterpret or fail to recognize their speech effectively.

The objective of our project is to develop an accent-agnostic speech transcription model that can recognize and accurately transcribe a wide variety of accents, thereby reducing the educational disparities caused by speech recognition inaccuracies. By integrating advanced neural network architectures such as Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM) units, and Gated Recurrent Units (GRU) with Mel-frequency cepstral coefficients (MFCCs) for audio processing, our model aims to enhance the inclusivity and accessibility of e-learning platforms, making educational content more universally comprehensible and engaging. This initiative not only supports equitable education but also promotes a more inclusive digital learning environment.

II. Problem Statement

The burgeoning growth of e-learning platforms has facilitated unparalleled access to educational resources globally. However, this digital revolution has also spotlighted significant inclusivity challenges, particularly in speech recognition technologies used within these platforms. Traditional speech recognition systems often struggle with accent diversity, primarily due to their training on datasets predominantly composed of standard accents. This bias leads to lower accuracy rates when transcribing speeches with non-standard or regional accents, thereby impeding comprehension for a substantial user base, including non-native speakers and individuals with hearing impairments.

The specific problem our project addresses is the accent bias in speech recognition within e-learning platforms. Current systems' limited ability to accurately transcribe diverse accents results in an inequitable distribution of learning opportunities and resources. This failure not only affects user engagement and learning outcomes but also exacerbates educational inequalities.

Our model seeks to rectify this by being accent-agnostic, thereby ensuring high transcription accuracy across a broad spectrum of accents. By leveraging advanced neural network architectures—specifically RNN, LSTM, and GRU—combined with feature extraction techniques such as MFCCs, our approach aims to create a more robust and inclusive speech recognition system. The model's objective is to minimize the educational barriers posed by accent diversity, thus enhancing accessibility and equity in digital education platforms.

II. Data Preparation

The dataset utilized in our project was sourced from Mozilla's Common Voice project, an open-source collection of voice data from speakers worldwide. This extensive dataset is characterized by its diversity in accents, making it an ideal resource for training an accent-agnostic speech transcription model. The Common Voice dataset includes thousands of hours of spoken audio, annotated with transcriptions, which provides a robust foundation for training and validating our speech recognition models.

Dataset Characteristics

- Source: Mozilla Common Voice
- Content: Audio recordings paired with text transcriptions
- Accents: Includes a wide range of global English accents
- Format: Audio files primarily in MP3 format

Preprocessing Steps

The pre-processing of the audio data involved several key steps to convert the raw audio files into a format suitable for feeding into our neural network models:

Conversion to WAV Format: Initially, the audio files, originally in MP3 format, were converted to WAV format using the FFmpeg software. This step was necessary because WAV files are uncompressed and provide consistent, high-quality audio data that is more suitable for processing and analysis.

Sampling Rate Adjustment: The audio files were resampled to a consistent sampling rate of 16,000 Hz, which is standard for speech recognition tasks. This ensures that the data across all files maintains uniformity in terms of temporal resolution.

Feature Extraction - MFCCs:

Why MFCCs?: Mel-frequency cepstral coefficients (MFCCs) are a feature widely used in speech and audio processing. The rationale behind using MFCCs lies in their ability to mimic the human auditory system's response more closely than other raw audio forms. They effectively represent the power spectrum of sound based on a logarithmic Mel scale of frequency, which closely approximates the human ear's resolution.

Process: To compute the MFCCs, each audio clip was first divided into short frames (typically 20-40 ms long), to assume stationarity within each frame. For each frame, the power spectrum was computed, followed by the application of the Mel filter bank to the power spectra, logarithm of the filter bank energies, and finally, the discrete cosine transforms (DCT) of the log filter bank energies. These steps convert the raw audio signals into a set of features that effectively capture the important characteristics of speech necessary for recognizing different phonemes across varied accents.

Data Normalization: Post-extraction, the MFCC features were normalized to have zero mean and unit variance. Normalization helps in speeding up the training process and improving the model's convergence behaviour.

III. Methodology

In this project, we employed three different recurrent neural network (RNN) architectures to tackle the challenge of accent diversity in speech recognition: the basic Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU). These architectures were selected due to their demonstrated capability to manage sequential data effectively, such as speech, which necessitates retaining information across time for successful processing.

Model Architectures

A. Recurrent Neural Network (RNN)

Architecture Details: This model configuration consists of two layers, each comprising 200 RNN cells. This architecture facilitates time-series data processing by updating its hidden state sequentially with each timestep.
Configuration: To mitigate the risk of overfitting, a dropout rate of 0.5 was implemented between the RNN layers. Following each RNN layer, a TimeDistributed Dense layer maps the recurrent layer outputs to the vocabulary size.
Justification: Basic RNNs are adept at sequence prediction tasks; however, they are susceptible to challenges like vanishing gradients, which impair their ability to handle long-term dependencies.

B. Long Short-Term Memory (LSTM)

Architecture Details: The LSTM model is structured with two layers, each containing 100 LSTM cells. LSTMs enhance the basic RNN model by incorporating mechanisms to avoid long-term dependency issues.
Configuration: Like the RNN model, this configuration includes a dropout rate of 0.5 between the layers to prevent overfitting and uses a TimeDistributed Dense layer for output generation.
Justification: LSTMs are particularly effective for tasks requiring a comprehensive understanding of the entire input sequence, as they are designed to maintain relevant information over extended periods.

C. Gated Recurrent Unit (GRU)

Architecture Details: Comprising two layers of GRU cells, each with 200 units, the GRU model simplifies the LSTM architecture by reducing the number of gating mechanisms used.
Configuration: The model employs a dropout rate of 0.5 and includes a TimeDistributed Dense layer for the final output, like the other models.
Justification: GRUs offer a balance between computational efficiency and model performance, making them suitable for large-scale applications such as speech recognition.

Model Training:

Optimization: All models were optimized using the Adam optimizer, noted for its efficiency in handling large datasets and model parameters due to its adaptive learning rate capability.

Loss Function: The categorical cross-entropy loss function was selected, appropriate for multi-class classification problems inherent in speech recognition tasks.

Performance Metrics: Model performance was primarily evaluated based on accuracy, alongside monitoring the convergence behavior and loss metrics during training/validation phases.

The diverse configurations and layers in these models were strategically chosen to optimize both model complexity and the capacity to generalize speech patterns effectively. This comprehensive approach not only illuminated the distinct behaviors of various RNN architectures but also facilitated the identification of the most suitable model in terms of accuracy and operational performance for this specific application.

III. Results

The evaluation of the three different neural network models—RNN, LSTM, and GRU—yielded insightful results regarding their performance in accent-agnostic speech recognition. The following sections detail the comparative results, highlighting key findings and any unexpected model behaviors.

Comparative Performance

The performance metrics for each model were primarily assessed through accuracy and loss during both training and validation phases. The table below summarizes these results:

TABLE I
Performance Comparison of rnn, lstm and GRU Model

Model	Training Accuracy	Validation Accuracy	Training Loss	Validation Loss
RNN	75.2%	72.8%	0.88	0.93
LSTM	81.4%	79.5%	0.65	0.70
GRU	83.9%	82.3%	0.59	0.62

Observations and Discussions

A. *General Performance*

The GRU model exhibited the highest accuracy and lowest loss among the three models, both during training and validation, suggesting better handling of dependencies and robustness against overfitting compared to RNN and LSTM.

The LSTM model also performed significantly better than the basic RNN, reinforcing the importance of advanced gating mechanisms in managing long-term dependencies in speech data.

B. *Unexpected Behavior*

The basic RNN model demonstrated a higher-than-expected validation loss compared to its training loss, indicative of overfitting despite the application of dropout. This may suggest that RNNs, without additional mechanisms to handle long sequences, are less suited for complex tasks like accent-agnostic speech recognition.

Both LSTM and GRU models showed a narrower gap between training and validation metrics, indicating better generalization to unseen data. However, the LSTM's performance, while good, was not as close to GRU's, which was an unexpected finding given their similar architectural complexity.

C. *Significant Findings*

The superior performance of the GRU model suggests that its simpler gating mechanisms are sufficiently powerful for the complexity of the speech recognition task, with the added benefit of being more computationally efficient than LSTM.

The dropout strategy was effective in managing overfitting in LSTM and GRU models but less so in the basic RNN, indicating the need for more sophisticated regularization techniques or model architectures for basic RNNs in similar tasks.

D. *Analysis*

The results clearly favor the GRU model in terms of both efficiency and effectiveness, making it the most suitable choice for the accent-agnostic speech transcription task in our project. The findings from this comparative analysis not only contribute to the theoretical understanding of RNN architectures but also have practical implications for the design of more inclusive and effective e-learning tools.

IV. Discussion

The analysis of the results from the RNN, LSTM, and GRU models provides a comprehensive understanding of their respective strengths and weaknesses, which is crucial for future improvements and applications. Here, we discuss these aspects in detail, including how the best-performing model aligns with the project's objectives.

Strengths and Weaknesses

A. *Recurrent Neural Network (RNN)*

Strengths: The basic RNN model is relatively simple and quick to implement. It is capable of processing sequences by updating its hidden state with incoming data, making it suitable for time-series prediction tasks.

Weaknesses: The primary limitation of the basic RNN is its difficulty in learning long-term dependencies due to the vanishing gradient problem. This was evident from the high validation loss indicating overfitting and poor generalization to unseen data.

B. Long Short-Term Memory (LSTM)

Strengths: LSTMs are designed to overcome the limitations of RNNs by incorporating gates that regulate information flow, thus enabling them to retain information over longer periods. This feature made the LSTM model perform significantly better than the basic RNN in terms of both accuracy and stability.

Weaknesses: Despite their advantages, LSTMs are computationally intensive due to their complex architecture, which can lead to longer training times. They may also require more data to train effectively due to the increased number of parameters.

C. Gated Recurrent Unit (GRU)

Strengths: GRUs simplify the LSTM architecture by using fewer gates, which helps in reducing the computational burden while retaining most of the advantages regarding long-term dependency management. This model showed the best performance in our tests, balancing computational efficiency with high accuracy and low loss.

Weaknesses: While GRUs are generally more efficient than LSTMs, they might still suffer from some of the issues typical to RNNs, such as sensitivity to hyperparameter settings and the need for ample training data to achieve optimal performance.

Alignment with Project Goals

The GRU model, with its high accuracy and robust performance across various accents, aligns well with our project's goal of developing an accent-agnostic speech transcription system. By effectively handling the variability in speech patterns and accents, the GRU model enhances the inclusivity and accessibility of e-learning platforms.

Limitations and Potential Improvements

Data Dependency: All models, particularly LSTM and GRU, require large amounts of data to train effectively. Expanding the dataset or employing data augmentation strategies could help improve model robustness.

Complexity and Resources: The computational resources required for training complex models like LSTM and GRU could be prohibitive. Optimizing the models to run efficiently on less powerful systems or employing model compression techniques could help mitigate this issue.

Hyperparameter Optimization: Further tuning of model parameters, such as learning rate, batch size, and number of layers, could potentially enhance model performance. Automated hyperparameter optimization techniques could be employed to find the optimal settings more efficiently.

Advanced Regularization Techniques: While dropout was used, other regularization methods like L1/L2 regularization, batch normalization, or advanced techniques such as dropout variants tailored for RNNs could be explored to improve generalization further.

IV. Conclusion

This project embarked on addressing the crucial challenge of accent diversity in e-learning platforms through the development of an advanced speech transcription model. By utilizing recurrent neural network architectures such as RNN, LSTM, and GRU, the project aimed to create an accent-agnostic system capable of accurately transcribing speech from diverse linguistic backgrounds. Here, we summarize the key points and the overarching contributions of this project.

Key Points

Data Preparation: We employed the Mozilla Common Voice dataset, known for its diversity in accents, and preprocessed this data into Mel-frequency cepstral coefficients (MFCCs), which effectively capture the essential characteristics of spoken language suitable for machine learning applications.

Model Implementation: Three neural network models were tested: basic RNN, LSTM, and GRU. Each model was designed to handle the sequential nature of speech data, with specific configurations aimed at overcoming the limitations of traditional speech recognition systems.

Performance Evaluation: The GRU model demonstrated the highest performance in terms of accuracy and loss, indicating its superior capability to handle long-term dependencies and generalize across different accents compared to RNN and LSTM.

Project Contributions: The GRU model's success underscores our project's contribution to enhancing accessibility in e-learning platforms by offering a more reliable and inclusive tool for speech recognition. This model ensures that users, regardless of their accent, can engage with digital educational content more effectively, thereby democratizing access to education.

Contribution to Accessibility in E-Learning

The accent-agnostic model developed through this project significantly contributes to reducing the barrier that accent diversity presents in automated speech recognition within e-learning platforms. By improving the accuracy of speech transcription across a multitude of accents, the project helps in making educational content more accessible to a broader audience. This inclusivity is vital for educational equity, ensuring that every learner can benefit from digital education regardless of their linguistic background.

References

[1] S. Young et al., "The HTK Book (for HTK Version 3.4)," Cambridge University Engineering Dept., 2009.

[2] D. Povey et al., "The Kaldi Speech Recognition Toolkit," in Proc. IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society, Dec. 2011, IEEE Catalog No.: CFP11SRW-USB.

[3] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," in Proceedings of the IEEE, vol. 77, no. 2, pp. 257-286, Feb. 1989.

[4] Mozilla Foundation, "Common Voice: Mozilla’s initiative for an open voice database," Mozilla Foundation, 2020. [Online]. Available: https://commonvoice.mozilla.org. [Accessed: Sept. 15, 2023].

[5] A. Graves et al., "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks," in Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, 2006, pp. 369-376.

[6] F. Chollet, "Deep Learning with Python," Manning Publications, 2017.

[7] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," IEEE Transactions on Neural Networks, vol. 5, no. 2, pp. 157-166, March 1994.

[8] I. Goodfellow, Y. Bengio, and A. Courville, "Deep Learning," MIT Press, 2016, pp. 321-359.

[9] "LibriSpeech: an ASR corpus based on public domain audio books," in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015.

[10] "FFmpeg," FFmpeg Development Team, 2020. [Online]. Available: https://ffmpeg.org. [Accessed: Oct. 5, 2023].