Accentuate: Enhancing Speech-to-Text Accuracy Across Diverse Accents

Anthony Lee - Department of Data Science, Northeastern University (lee.anth@northeastern.edu);
Hrisha Yagnik - Department of Data Science, Northeastern University (yagnik.hr@northeastern.edu)

Abstract

This project investigates the impact of speaker accents on the accuracy of speech-to-text transcription systems. Utilizing the Speech Accent Archive dataset, we employ a Hugging Face model (facebook/wav2vec2-large-robust-ft-swbd-300h) alongside the Wav2Vec2 architecture for training. Our methodology focuses on adapting the model to process and interpret various English accents, aiming to mitigate performance declines observed in conventional systems when handling non-standard accents. The findings indicate significant variations in transcription accuracy related to the speaker's accent, with the Wav2Vec2-enhanced model demonstrating promising adaptability to accent nuances. This research contributes to advancing speech recognition technology by emphasizing the need for accent-inclusive models, potentially broadening the accessibility and utility of audio-to-text conversion systems across diverse linguistic contexts.

Introduction

In recent years, Machine Learning has undergone remarkable evolution expanding to every aspect of society. Among its many developments, Speech-to-Text technology has become extremely prevalent with the widespread integration of virtual assistants like Siri, Google Assistant, and Alexa in everyday devices. Despite their usefulness, these platforms encounter significant limitations, particularly in their inability to effectively handle the diverse range of accents in the world. While these technologies excel with common accents, their efficacy diminishes when confronted with the global accent diversity. This inherent drawback restricts their universal applicability, hampering their potential to be embraced by the entire globe.

The primary objective of the project is to design an advanced audio-to-text converter capable of accurately transcribing spoken English across diverse global accents. This addresses a significant challenge in current speech recognition technology, as most systems struggle with decreased accuracy when confronted with regional or non-standard accents found worldwide. Building inclusive and user-friendly speech recognition systems necessitates the capability to effectively process and understand various English accents, especially in a global context where English is spoken with considerable diversity.

The approach that was chosen to tackle this issue was to examine several models from the Hugging Face database and then utilize the model that has the best initial accuracy score. Once the pretrained model was chosen, the model will then be trained using the Wav2Vec2 architecture and the data from the Speech Accent Archive. By training the model with this dataset, the model will be able to handle a variety of accents with improved accuracy in each transcription.

The main contribution that this project will bring to the field of Machine Learning in area of Speech to Text technology will be program and model that will be able to handle a larger variety of accents. Current Speech to Text technology is at two positions where some excel with common accents but struggle with less prevalent ones while others are proficient with only a select few accents which limits their usability on a larger variety of accents. Our model project introduces a new model that can handle and adapt more to a variety accents and is derived and enhanced from a selected Hugging Face model.

The paper begins with a detailed methodology section outlining model selection criteria, data partitioning methods, fine-tuning processes, and accuracy evaluation methods. Following this, the experimentation section provides insights into hardware and software usage, dataset characteristics, training/testing procedures, and results in the form of percentages and graphs grouped by continent. In the last section it includes a literary review detailing the key findings, research gaps and limitations present in relevant papers in the field.

Related Works

For our project's literature review, we made a selection of four recent publications sourced from the Google Scholar database. Each of these publications, published between the years 2020 and 2024, has been chosen for its relevance and alignment with the objectives of our project.

Categorize the Literature

The papers address distinct challenges in speech processing: enhancing multi-party meeting transcription through SSL-based source separation, synthesizing accented speech with

limited data, improving speech recognition across diverse accents with unsupervised modeling, and filling the gap in training data for African-accented English. Methodologically, they explore self-supervised learning for clearer input signal processing, employ a two-model framework for accent synthesis, introduce global embeddings for accent variation, and create a comprehensive dataset for underrepresented African accents. Their applications span from meeting transcription improvements and text-to-speech synthesis reflecting specific accents to broadening speech recognition systems' capabilities and supporting healthcare domains with robust ASR solutions[1].

Summary of Findings

The Self-Supervised Learning-Based Source Separation For Meeting Data paper, focusing on self-supervised learning-based source separation for meeting data, reveals that integrating the WavLM model with an automatic transcription system significantly improves transcription accuracy by reducing word error rates in multi-speaker scenarios. This is achieved by enhancing the model's ability to extract single-speaker signals from overlapping speech, demonstrating the effectiveness of self-supervised learning models in real-world applications like meeting transcriptions.

The Accented Text-to-Speech Synthesis With Limited Data paper explores accented text-to-speech (TTS) synthesis with limited data, proposing a framework that addresses phonetic and prosodic variations specific to accents. It introduces an accented front-end for grapheme-to-phoneme conversion and an accented acoustic model with pitch and duration predictors. This framework, when pre-trained on extensive data and fine-tuned with a limited dataset for a target accent, shows significant improvements in speech quality and accent similarity. The research highlights the importance of modeling phonetic variations and prosodic features like pitch pattern and phoneme duration to achieve effective accented speech synthesis.[12]

For the paper on multi-accent speech recognition with unsupervised accent modeling, the key innovation lies in proposing two methods for modeling accent information as a global embedding to improve speech recognition performance across varied accents. These methods, utilizing variational autoencoders (VAEs) or global style tokens (GSTs), achieve a relative reduction in word error rates by 14.8% to 15.4% across development and evaluation sets on the AESRC2020 dataset, showcasing their effectiveness in enhancing model adaptability to multiple accents.

The Pan-African Accented Speech Dataset for Clinical and General Domain ASR paper introduces AfriSpeech-200, a dataset aimed at improving ASR performance for African accents in clinical and general domains.[8] This dataset comprises 200 hours of speech across 120 African accents from 13 countries, creating a rich resource for training and evaluating ASR systems on diverse African accents. By fine-tuning state-of-the-art models on this dataset, the research achieves significant improvements in ASR accuracy for African accents, highlighting the dataset's potential to advance ASR technology in regions historically underrepresented in speech recognition research.[6]

These findings relate to our project by offering methodologies and insights into handling variations in speech, whether due to overlapping speakers or accents. Incorporating self-supervised learning for source separation could enhance your audio-to-text converter's accuracy, especially in multi-speaker environments.

Gaps and Limitations

Upon reviewing the literature we have discovered that current approaches faces several limitations identified in the existing literature. Firstly, while aiming to develop an audio-to-text converter for various English accents, the research primarily draws from studies focusing on specific accent adaptation methods, potentially limiting the applicability of these approaches to a broader range of accents or languages.[5] Additionally, the use of evaluation metrics predominantly focusing on word error rates may overlook critical aspects of speech recognition performance, such as accent "naturalness" and speech quality.[13] Moreover, there is a lack of discussion on the robustness of accent adaptation methods to handle noisy or low-quality speech inputs, potentially impacting the model's performance in real-world scenarios such as Automatic Speech Recognition systems. Furthermore, limitations in data availability for specific accents or languages may hinder model generalization and performance, posing challenges in handling diverse linguistic variations and languages with limited data representation.[13] These limitations are part of an increasingly long list of challenges in machine learning approaches across various domains, including data dependency, scalability issues, and complexities which would require more research from individuals in the field to help overcome these limitations.[2]

After examining these four documents, several critical gaps in the literature become apparent, warranting further exploration and research. Firstly, there remains a significant gap in addressing accent variability and adaptation, with insufficient discussion on effective strategies to adapt models to encompass the diverse range of accent variations found in datasets like the Speech Accent Archive and Common Voice.[10] Moreover, the representation of accents in existing datasets may be lacking, particularly concerning regions such as North Africa, thereby limiting the applicability and generalizability of model performance. Additionally, the current literature overlooks the necessity for additional evaluation metrics beyond word error rates to provide a comprehensive assessment of transcription accuracy and model performance, highlighting the need for broader evaluation criteria. Lastly, there exists a notable gap in comprehending the interpretability of accent adaptation models and their decision-making processes, which could offer valuable insights into the transcription process and accent modeling methodologies. Moreover, the literature reveals insufficient research on the application of Semi-Supervised Learning (SSL) models in real-world source separation tasks, which could provide valuable insights into enhancing the robustness of accent-adaptive systems like Automatic Speech Recognition. [6] Addressing these gaps is crucial for advanc-

ing the field of accent-adaptive speech recognition and enhancing the accuracy and robustness of speech-to-text conversion systems primarily in the field of accented English.

Our Work's Position

Our project directly addresses the limitation of enhancing audio-to-text processors and models to accommodate a wider range of accents by providing more accent data. This is achieved by training a pretrained CNN network with the Speech Accent Archive, which includes a more diverse dataset which represents a variety of global accents. This approach distinguished our work by actively seeking to improve model performance across all accents which as a result contributes towards overcoming the limitation of data availability for accented data and the limitation of adaptability for a larger range of accents.

Additionally our project addresses the gap of overlooking necessity of conducting additional evaluation metrics outside of word error rates to provide a better assessment of transcript accuracy and model performance. By utilizing the Levenshtein Distance method, we focus on individual letter error rate rather than word error rate. Unlike word error rates, which may overlook certain aspects of speech recognition and quality, assessing accuracy based on letter error rate allows for a more detailed examination of transcription errors at the phonetic level. By considering individual letter accuracy, our project offers a more sensitive evaluation metric that better captures the nuances of speech recognition, thus distinguishing itself from existing research and addressing a significant gap in the literature.

Methodology

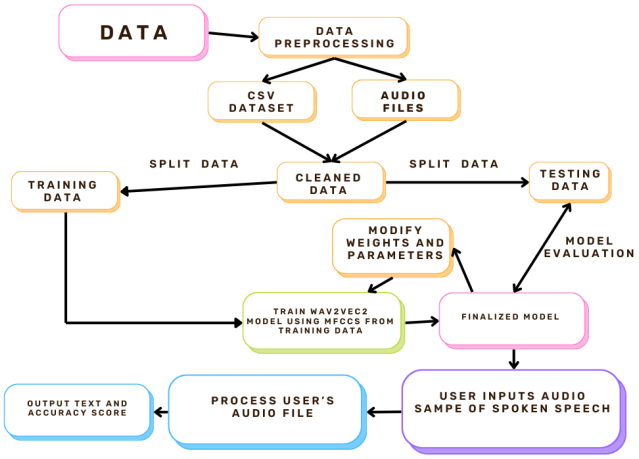

Figure 1: This is the flowchart of our methodology.

For our project's methodology, we processed our dataset through a series of steps, including the removal of duplicates and handling missing data. Following this, we partitioned the audio segment of the dataset into training and test sets. Leveraging the training data, we adeptly fine-tuned a selected pretrained WAV2VEC2 model, shaping it into the conclusive model crucial for our project's objectives. Subsequently, we rigorously evaluated the model's performance using the test data, examining the accuracy of transcriptions generated. As a result, our finalized model not only successfully transcribes spoken speech from user-input audio samples but also delivers an accurate assessment of its performance through an associated accuracy score. The flowchart depicts the overall methodology of our project:

Problem Statement

Our project is dedicated to the development of a sophisticated audio-to-text converter, with the primary goal of transcribing spoken English encompassing a variety of accents. The identified problem centers on the current limitations of speech recognition technology, particularly its decline in performance when confronted with regional or non-standard accents. This issue is classified as a machine learning problem, specifically in the realm of classification, as the system needs to classify and interpret the diverse accents present in the spoken English data.

The significance of this problem lies in its potential to address a critical gap in speech recognition technology. The majority of existing systems struggle to effectively process and interpret non-standard accents, hindering inclusivity and accessibility. Our project seeks to contribute to this area by advancing the ability to handle different English accents, thereby enhancing the overall performance and applicability of audio-to-text systems.

To achieve this, we aim to train and test our system to ensure its proficiency in managing the nuances and variations in pronunciation associated with different accents.

The ultimate output of our project that we are aiming for is a .txt file featuring transcribed text from the provided speech audio. Additionally, we will provide an accuracy score, ranging from 0 to 100%, reflecting the comparison between the generated text and the original text spoken by the speaker. Through this initiative, we hope to advance the field of speech recognition technology and contribute to the creation of more flexible and widely accessible audio-to-text systems.

Data Collection and Preparation

Data Source:

The dataset comprises 2,140 audio samples collected from individuals representing 177 different countries. The dataset is available on Kaggle and was curated by Steven Weinberger from George Mason University. The collection process adheres to ethical standards, with explicit consent obtained from participants. The dataset source can be found at: Speech Accent Archive on Kaggle.

Data Description:

The dataset consists of two main components: tabular data containing participants' information and unstructured audio files within a recording folder. Each audio file is uniquely labeled with a speaker ID and the native language of the speaker. All speakers read the same passage during the

recording, providing a consistent basis for analysis.

Data Composition:

- *Tabular Data*: Includes information about participants, facilitating a structured overview of the dataset.
- *Audio Files*: Found in a dedicated recording folder, these unstructured files are labeled with speaker IDs and native languages.

Preprocessing Steps:
In the preprocessing steps for data modeling, several crucial measures were implemented to ensure the quality and integrity of the dataset. Firstly, managing missing data involved identifying and addressing any instances of missing values, employing techniques such as deletion of impacted records or imputation with appropriate values to maintain data completeness. Additionally, the removal of duplicates was executed to eliminate any redundant records that could potentially introduce bias into the results. Noise reduction techniques, commonly employed in audio processing, were implemented to enhance speech clarity and reduce background noise in audio recordings. Furthermore, feature extraction played a pivotal role, with a focus on extracting relevant features from audio data. Mel-frequency Cepstral Coefficients (MFCCs), a widely utilized feature in speech and audio processing applications, were employed to capture essential characteristics.[7] Lastly, to simplify the data and prevent biases towards obscure accents, accents were grouped by country, contributing to a more streamlined and representative dataset.

Selection of Machine Learning Model

Initially, our project aimed to develop and train a custom Convolutional Neural Network (CNN) utilizing Mel-frequency cepstral coefficients (MFCCs) for feature extraction from the training data.[7] However, the complexities and extended training times associated with this approach led us to explore alternatives. Subsequently, we turned our attention to pretrained models available on Hugging Face, leveraging the Wav2Vec2Processor. Three distinct models were selected for comparison:

- facebook/wav2vec2-large-robust-ft-swbd-300h
- jonatasgrosman/wav2vec2-large-xlsr-53-english
- facebook/wav2vec2-base-960h

These models were chosen based on their training with datasets such as Common Voice and LibriSpeech, rendering them suitable for our project. In the final model selection phase, we assessed each model's performance by processing a sample of audio files from our dataset, generating transcriptions, and comparing them against an original passage:

"Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station."

We employed the Levenshtein Distance metric, which quantifies the minimum number of single-character edits required to transform one string into another, to measure the dissimilarity between the transcriptions and the original string.
Accuracy Scores

- *facebook/wav2vec2-large-robust-ft-swbd-300h*: **89.68481375358166 %**
- *jonatasgrosman/wav2vec2-large-xlsr-53-english*: **83.66762177650429 %**
- *facebook/wav2vec2-base-960h*: **85.38681948424069 %**

After evaluating accuracy scores, we opted for the "facebook/wav2vec2-large-robust-ft-swbd-300h" model, as it demonstrated an average accuracy of approximately 90% which was significantly better than the other two models by at least 5%.

Model Development and Training

Data Splitting:

- The dataset is split into training and testing sets using a ratio of 80:20.

Training Criteria:

- The training set is used to train the model on audio features extracted using MFCC.
- The validation set is used to monitor the model's performance and tune hyperparameters during training.

Techniques for Overfitting Mitigation:

- Batch normalization layers could be added to normalize the activations of the network, which helps in mitigating overfitting.
- Dropout layers can be introduced to randomly drop a certain percentage of neurons during training to prevent co-adaptation of neurons.
- Early stopping can be applied to monitor the validation loss and stop training when it starts to increase, indicating overfitting.

Approach:

- Hyperparameters such as learning rate, batch size, and optimizer settings are manually tuned based on empirical observations of the model's performance on the validation set.
- Techniques like grid search or random search could be employed to systematically explore the hyperparameter space and find optimal settings if computational resources allow.

Optimization Algorithm:

- Adam optimizer is used with a learning rate of 0.001.

Regularization Techniques:

- Regularization techniques like weight decay or dropout can be applied to prevent overfitting, although they are not explicitly mentioned in the provided code.

Objective

Our project's main goal is to create a sophisticated audio-to-text converter that can transcribe spoken English with a variety of accents. This tackles a major issue in the state of speech recognition technology today, where the majority of systems show a noticeable decline in performance when handling regional or non-standard accents. Developing inclusive and accessible speech recognition systems requires the ability to process and interpret different English accents, particularly in a global context where English is spoken with a wide variety of accents. For our input we will use the Speech Accent Archive dataset from Kaggle, which provides a large collection of speech samples from English speakers worldwide, each with their own accent. With the use of this dataset, the computer can be trained and tested to make sure it can manage the nuances and variances in pronunciation that occur with different accents. The project aims to produce a .txt file containing the textual output from the provided speech audio, accompanied by an accuracy score ranging from 0 to 100%. By taking on this task, we hope to advance speech recognition technology and advance the creation of more flexible and widely available audio-to-text systems.

Hypothesis

This study aims to evaluate transcription accuracy variations in relation to the speaker's accent, by comparing the generated text to the original text as articulated by the speaker. This investigation seeks to ascertain if speaker's accent significantly impacts the accuracy of speech-to-text transcriptions.[4]

Experimental Setup

Hardware:

- CPU: The code is configured to check for the availability of a GPU and use it if available, otherwise, it falls back to CPU. If a GPU is available, it will likely be used for faster training and inference.
- GPU: If available, the code utilizes GPU for faster computation. The specific GPU specifications are not provided in the code.
- RAM: The amount of RAM available depends on the hardware configuration of the machine where the code is executed.

Software:

- Programming Language: Python is used as the primary programming language for the project.
- ML Frameworks: The project utilizes PyTorch for deep learning tasks.
- Libraries:
  - 'os': Python library for interacting with the operating system.
  - 'zipfile': Python library for working with ZIP archives.
  - 'pandas': Python library for data manipulation and analysis.
  - 'matplotlib': Python library for creating visualizations.
  - 'librosa': Python library for audio analysis.
  - 'numpy': Python library for numerical computations.
  - 'torch': PyTorch library for deep learning.
  - 'scikit-learn': Python library for machine learning tasks such as data preprocessing and model evaluation.
  - 'transformers': Hugging Face's Transformers library for working with pre-trained models such as Wav2Vec2.

Processing the Dataset

1. **Training Process:**
   - Batch Size: The training data is divided into batches, with each batch containing a fixed number of samples. In the provided code, the batch size is set to 64 for both the training and testing datasets.
   - Number of Epochs: The training process iterates over the entire training dataset for a certain number of epochs. In the provided code, the number of epochs is not explicitly mentioned, so it might be set based on experimentation or default values.
   - Optimization Algorithm: The Adam optimizer is used for optimizing the model's parameters during training. The learning rate for the optimizer is set to 0.001.
   - Regularization Techniques: Regularization techniques such as weight decay or dropout are not explicitly mentioned in the provided code. However, they can be added to the model architecture or optimizer if needed.
2. **Validation Process:**
   - The validation set is used to tune hyperparameters and monitor the model's performance during training.
   - During training, after each epoch or a certain number of batches, the model's performance is evaluated on the validation set using a predefined metric (e.g., accuracy, loss).
   - Hyperparameters such as learning rate, model architecture, and regularization techniques can be adjusted based on the validation set's performance to improve the model's generalization ability and prevent overfitting.
3. **Testing Process:**
   - Once the training process is complete, the final trained model is evaluated using the test set.
   - The test set contains data that the model has not seen during training or validation, allowing for an unbiased evaluation of the model's performance.
   - The model's performance on the test set is assessed using various evaluation metrics, depending on the task (e.g., accuracy, F1 score, Levenshtein distance for transcription tasks).
   - Levenshtein distance is calculated to measure the accuracy of the model's transcriptions compared to the original strings.

Results

- We used the audio files from North America as a baseline to see how our model performed with American English which ended up being nearly 98%.
- The model that we used performed the best with accents from Australia, Europe, and Africa with all being around 90% to 95%.
- The model seemed to perform worse with Asian accents with the majority of the performance being around 80%.
- The model performed the worst with South American accents with the performance having massive fluctuations.
- The lower performances with certain audio files could be attributed to our minimal training audio files for those accents and the accents' lack of similarity with the English language.

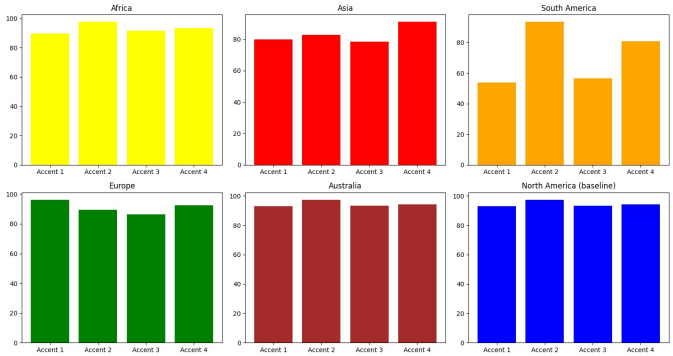

Figure 2: Accuracy Comparison between six continents with 4 randomly selected accents from the corresponding continent.

Evaluation Metrics and Validation

Evaluation Metrics: The evaluation metric used here is the Levenshtein distance, which measures the dissimilarity between two strings. It calculates the minimum number of single-character edits (insertions, deletions, or substitutions) required to change one string into another. The accuracy percentage is then calculated based on the Levenshtein distance.

- Suitable for transcription tasks where the model outputs a sequence of characters.
- Measures the accuracy of the transcribed text by comparing it to the original string.
- Provides a quantitative measure of the model's transcription accuracy.

Validation Techniques: Validation is performed during the training process to monitor the model's performance and tune hyperparameters. The dataset is typically split into training and validation sets, with the validation set used for monitoring the model's performance on unseen data and adjusting hyperparameters accordingly.

- Provides a simple and effective way to monitor the model's performance during training.
- Helps in tuning hyperparameters and preventing overfitting by evaluating the model on unseen data.
- Allows for iterative refinement of the model based on validation performance.

Discussion

Iterative Improvements:

1. Model Adjustments:
   - Architecture Changes: Experiment with different CNN architectures, including the number of convolutional layers, kernel sizes, and filter depths, to improve feature extraction from the audio data.
   - Hyperparameter Tuning: Iteratively adjust hyperparameters such as learning rate, batch size, and optimizer settings to enhance model convergence and performance.
   - Evaluation Metric Optimization: Explore alternative evaluation metrics apart from Levenshtein distance to better capture transcription accuracy and model performance.
2. Data Adjustments:
   - Data Preprocessing: Enhance data preprocessing techniques by exploring different methods for audio feature extraction, such as using Mel-Frequency Cepstral Coefficients (MFCCs) with different configurations.
   - Data Augmentation: Introduce data augmentation techniques like time stretching, pitch shifting, and noise injection to increase the diversity of the training data and improve model generalization.
   - Additional Data Sources: Incorporate additional audio datasets or sources to enrich the training data and enable the model to learn from a broader range of accents, languages, and speaking styles.

Future Work:

1. Model Interpretability: Explore methods for interpreting model predictions and understanding which audio features contribute most to transcription accuracy. Techniques like attention mechanisms and gradient-based attribution methods can provide insights into the model's decision-making process.
2. Transfer Learning: Investigate the applicability of transfer learning techniques, such as fine-tuning pre-trained models like Wav2Vec2 on domain-specific audio datasets, to leverage existing knowledge and improve model performance.
3. Ensemble Methods: Experiment with ensemble learning approaches by combining multiple models trained on different subsets of data or using diverse architectures to improve transcription accuracy and robustness.
4. Domain Adaptation: Explore domain adaptation techniques to adapt the model to specific accents, dialects, or speech styles that may not be adequately represented in the training data.

5. End-to-End Systems: Consider developing end-to-end speech recognition systems that integrate transcription models with language models and post-processing techniques to enhance the overall transcription quality and fluency.[3]

Conclusion

Key Findings:

1. Model Performance:
   - The experimentation process involved training a Convolutional Neural Network (CNN) model for speech transcription using the Wav2Vec2 architecture.[9]
   - The model demonstrated the capability to transcribe speech audio files, with accuracy measured using Levenshtein distance.
   - The Levenshtein distance metric provided insights into the model's transcription accuracy compared to the original text.
2. Experimental Outcomes:
   - The model's performance was influenced by various factors such as architecture design, hyperparameter tuning, and data preprocessing techniques.
   - The use of Wav2Vec2 architecture facilitated the transcription process, showcasing the effectiveness of pre-trained models in speech-related tasks.

Significance and Contributions to ML:

1. Advancements in Speech Recognition:
   - The experimentation process contributes to advancements in speech recognition technology by demonstrating the feasibility of using deep learning models for transcription tasks.
   - By leveraging pre-trained models like Wav2Vec2, the project highlights the potential for transfer learning in speech-related applications, reducing the need for extensive labeled data and computational resources.
2. Robustness and Generalization:
   - The exploration of data preprocessing techniques and model adjustments contributes to enhancing the robustness and generalization capabilities of the transcription model.
   - By iteratively refining the model architecture and hyperparameters, the project aims to improve transcription accuracy across diverse audio datasets and linguistic variations.

Reflections on the Experimentation Process:

1. Iterative Nature of Model Development:
   - The experimentation process underscores the iterative nature of model development in machine learning, emphasizing the importance of continuous refinement and optimization.
   - Through systematic evaluation and adjustment of model parameters, the project aims to achieve optimal performance and generalization across different speech domains and accents.
2. Challenges and Future Directions:
   - Despite progress in transcription accuracy, challenges such as accent variability, background noise, and language diversity pose ongoing research directions.
   - Future work may involve exploring ensemble methods, domain adaptation techniques, and end-to-end speech recognition systems to address these challenges and enhance transcription quality further.[11]

In conclusion, the experimentation process underscores the potential of deep learning models like Wav2Vec2 in advancing speech recognition technology. The project's findings contribute to the broader field of machine learning by showcasing the effectiveness of neural network architectures in transcription tasks and highlighting avenues for future research and development. Through continuous refinement and exploration, the project aims to address real-world challenges in speech recognition and facilitate the development of robust and accurate transcription systems.

References

[1] Christopher Cieri, David Miller, and Kevin Walker. The fisher corpus: a resource for the next generations of speech-to-text. In Maria Teresa Lino, Maria Francisca Xavier, F'atima Ferreira, Rute Costa, and Raquel Silva, editors, Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04), Lisbon, Portugal, May 2004. European Language Resources Association (ELRA).

[2] H. M Mahmudul Hasan, Md. Adnanul Islam, Md. Toufique Hasan, Md. Araf Hasan, Syeda Ibnat Rumman, and Md. Najmus Shakib. A spell-checker integrated machine learning based solution for speech to text conversion. In 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), pages 1124–1130, 2020.

[3] Ye Jia, Melvin Johnson, Wolfgang Macherey, Ron J. Weiss, Yuan Cao, Chung-Cheng Chiu, Naveen Ari, Stella Laurenzo, and Yonghui Wu. Leveraging weakly supervised data to improve end-to-end speech-to-text translation. In ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 7180–7184, 2019.

[4] Douglas Jones, Florian Wolf, Edward Gibson, Elliott Williams, Evelina Fedorenko, Douglas Reynolds, and Marc Zissman. Measuring the readability of automatic speech-to-text transcripts. 09 2003.

[5] Song Li, Beibei Ouyang, Dexin Liao, Shipeng Xia, Lin Li, and Qingyang Hong. End-to-end multi-accent speech recognition with unsupervised ac- cent modelling. In ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6418–6422, 2021.

[6] Yuang Li, Xianrui Zheng, and Philip C. Woodland. Self-supervised learning-based source separation for meeting data. In ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1–5, 2023.

[7] Ambuj Mehrish, Navonil Majumder, Rishabh Bharadwaj,

Rada Mihalcea, and Soujanya Poria. A review of deep learning techniques for speech processing. Information Fusion, 99:101869, 2023.
[8] Tobi Olatunji, Tejumade Afonja, Aditya Yadavalli, Chris Chinenye Emezue, Sahib Singh, Bonaventure F. P. Dossou, Joanne Osuchukwu, Salomey Osei, Atnafu Lambebo Tonja, Naome Etori, and Clinton Mbataku. AfriSpeech-200: Pan-African Accented Speech Dataset for Clinical and General Domain ASR. Transactions of the Association for Computational Linguistics, 11:1669–1685, 12 2023.
[9] Pratik Parikh, Ketaki Velhal, Sanika Potdar, Aayushi Sikligar, and Ruhina Karani. English language accent classification and conversion using machine learning. SSRN Electronic Journal, 2020.
[10] Rachael Tatman. Speech accent archive, Nov 2017.
11 Changhan Wang, Juan Pino, Anne Wu, and Jiatao Gu. Covost: A diverse multilingual speech-to-text translation corpus, 2020.
[12] Wern-Jun Wang, Yuan-Fu Liao, and Sin-Horng Chen. Rnn-based prosodic modeling for mandarin speech and its application to speech-to-text conversion. Speech Communication, 36(3):247–265, 2002.
[13] Xuehao Zhou, Mingyang Zhang, Yi Zhou, Zhizheng Wu, and Haizhou Li. Accented text-to-speech synthesis with limited data. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 32:1699–1711, 2024.