A Project report on

Voice Accent Detection Model

A Dissertation submitted to JNTU Hyderabad in partial fulfillment of the academic requirements for the award of the degree.

Bachelor of Technology

in

Computer Science and Engineering

Submitted by

A.SHARON
(20H51A0502)
E.NISHANTH REDDY
(20H51A0590)
C.S.K. SANKEERTH
(20H51A05N4)

Under the esteemed guidance of

Ms.E.Krishnaveni
(Assistant Professor)

Department of Computer Science and Engineering

CMR COLLEGE OF ENGINEERING & TECHNOLOGY
(UGC Autonomous)
*Approved by AICTE *Affiliated to JNTUH *NAAC Accredited with A^{+} Grade
KANDLAKOYA, MEDCHAL ROAD, HYDERABAD - 501401.

2020- 2024

CMR COLLEGE OF ENGINEERING & TECHNOLOGY

KANDLAKOYA, MEDCHAL ROAD, HYDERABAD – 501401

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the Major Project Phase I report entitled **"Voice Accent Detection Model"** being submitted by A.Sharon (20H51A0502), E.Nishanth Reddy (20H51A0590), C.S.K.Sankeerth(20H51A05N4) in partial fulfillment for the award of **Bachelor of Technology in Computer Science and Engineering** is a record of bonafide work carried out his/her under my guidance and supervision.

The results embodies in this project report have not been submitted to any other University or Institute for the award of any Degree.

Ms.E.Krishnaveni
Assistant Professor
Dept. of CSE

Dr. Siva Skandha Sanagala
Associate Professor and HOD
Dept. of CSE

ACKNOWLEDGEMENT

With great pleasure we want to take this opportunity to express my heartfelt gratitude to all the people who helped in making this project work a grand success.

We are grateful to **Ms.E.Krishnaveni, Assistant Professor** , Department of Computer Science and Engineering for his valuable technical suggestions and guidance during the execution of this project work.

We would like to thank **Dr. Siva Skandha Sanagala,** Head of the Department of Computer Science and Engineering, CMR College of Engineering and Technology, who is the major driving forces to complete my project work successfully.

We are very grateful to **Dr. Vijaya Kumar Koppula**, Dean-Academics, CMR College of Engineering and Technology, for his constant support and motivation in carrying out the project work successfully.

We are highly indebted to **Major Dr. V A Narayana,** Principal, CMR College of Engineering and Technology, for giving permission to carry out this project in a successful and fruitful way.

We would like to thank the **Teaching & Non- teaching** staff of Department of Computer Science and Engineering for their co-operation

We express our sincere thanks to **Shri. Ch. Gopal Reddy**, Secretary, CMR Group of Institutions, for his continuous care.

Finally, We extend thanks to our parents who stood behind us at different stages of this Project. We sincerely acknowledge and thank all those who gave support directly and indirectly in completion of this project work.

A.Sharon	20H51A0502
E.Nishanth Reddy	20H51A0590
C.S.K.Sankeerth	20H51A05N4

TABLE OF CONTENTS

ABSTRACT

Accurate identification of an individual's mother tongue from their English speech is a challenging task due to the presence of subtle linguistic influences. In this study, we propose a novel voice accent detection model aimed at predicting the speaker's mother tongue based on their English speech patterns. The model leverages advanced deep learning techniques to capture intricate phonetic variations and linguistic characteristics unique to each mother tongue. By utilizing a hybrid architecture that combines Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNNs), the model effectively extracts temporal and spectral features from the audio data. Our diverse dataset includes multilingual speakers with various accents and regional speech patterns, enabling the model to discern and accurately predict the speaker's mother tongue, even when faced with varying degrees of English fluency. This research holds significant potential for applications in language assessment, speech recognition systems, and personalized language learning tools, with implications for cross-cultural communication, linguistic research, and multilingual education.

CHAPTER 1
INTRODUCTION

CHAPTER 1
INTRODUCTION

1.1.Problem Statement

Accurately determining someone’s language based on their speech poses a significant hurdle as there are subtle linguistic influences that contribute to diverse accents. This challenge arises from the importance of promoting cultural communication and inclusivity in an interconnected world. The main objective is to create a voice accent detection model that can accurately recognize these accents even when individuals have varying levels of fluency, in English. To tackle this challenge we utilize cutting edge learning techniques like LSTM and CNN architectures training the model on a diverse dataset consisting of multilingual speakers with different accent nuances. Solving this problem is crucial, for advancing language assessment methods improving speech recognition systems and enabling personalized language education while embracing the richness of diversity.

1.2.Research Objective

The primary objective of this research is to develop a state-of-the-art voice accent detection model that transcends traditional language boundaries, enabling accurate prediction of an individual's mother tongue based on their English speech patterns. In an era of global communication and cultural diversity, the need for such a model has never been more pronounced. Leveraging the power of advanced deep learning techniques, including hybrid LSTM and CNN architectures, our research seeks to unravel the intricate phonetic variations and linguistic nuances that characterize diverse accents. Our mission extends beyond technical achievement; it aims to foster cross-cultural understanding and inclusivity. By training our model on a diverse dataset that spans multilingual speakers with varying English proficiency and regional speech intricacies.

1.3. Project Scope and Limitattions

Project Scope:

1. Accent Recognition: The model aims to accurately identify a diverse range of accents and regional speech patterns, enhancing cross-cultural communication and inclusivity.
2. Linguistic Inclusivity: By distinguishing accents and mother tongues, the model promotes effective understanding and communication, fostering linguistic inclusivity in an interconnected world.
3. Deep Learning Techniques: Leveraging advanced deep learning methods, including LSTM and CNN architectures, to capture intricate phonetic and linguistic nuances unique to each mother tongue.
4. Diverse Dataset: The model is trained on a diverse dataset that includes multilingual speakers with varying degrees of English fluency, enabling it to adapt to real-world scenarios effectively.

Limitations:

1. English Language Dependency: The model focuses on detecting accents and mother tongues in English speech; its accuracy may vary when applied to other languages.
2. Data Availability: The model's performance depends on the quality and diversity of the training dataset. Limited or biased data may affect its accuracy.
3. Continuous Learning: As language evolves, the model may require periodic updates to adapt to new accents and linguistic changes.
4. Regional Variations: Some regional accents can be highly nuanced, posing a challenge for accurate recognition. The model may not distinguish very subtle variations.
5. English Fluency: The model's performance can be impacted when English fluency is exceptionally high, as the mother tongue's influence may become less pronounced.

CHAPTER 2
BACKGROUND WORK

CHAPTER 2

BACKGROUND WORK

2.1 Google Cloud Speech-to-Text

2.1.1 Introduction

Google Cloud Speech-to-Text is a cloud-based service that leverages cutting-edge machine learning and deep learning models to convert spoken language into written text. It's a versatile and powerful tool for applications that require speech recognition and transcription capabilities. Whether for transcription services, voice assistants, voice commands, or any application that needs to convert spoken words into text, Google Cloud Speech-to-Text provides a robust solution.

2.1.2 Merits, Demerits and Challenges

Merits:

1. Accuracy: Google Cloud Speech-to-Text offers high accuracy in recognizing spoken language, making it suitable for critical applications like transcription services and voice assistants.
2. Multilingual Support: It supports multiple languages and dialects, which is essential for global applications.
3. Real-Time Processing: The service can process audio streams in real-time, enabling live captioning and other real-time applications.
4. Customization: Users can train custom models to improve accuracy for domain-specific content, making it adaptable for various industries.
5. Speaker Diarization: It can distinguish between different speakers in an audio recording, which is valuable in applications like meeting transcription.
6. Integration: It seamlessly integrates with other Google Cloud services, facilitating the development of comprehensive, cloud-based applications.

Demerits:

1. Cost: Google Cloud services are not free, and the cost can add up, especially for applications with high usage.
2. Internet Dependency: It requires a reliable internet connection to access the cloud service, which may not be suitable for offline or edge computing applications.

Challenges:

1. Speaker Variability: Accurate recognition can be challenging when dealing with multiple speakers or strong accents.
2. Background Noise: Background noise can interfere with accurate speech recognition, requiring additional audio preprocessing.
3. Security and Privacy: In applications involving sensitive data, ensuring the security and privacy of transcribed content is a challenge.

2.1.3 Implementation of Google Cloud Speech-to-Text:

1. Access the Service: Sign up for Google Cloud and enable the Speech-to-Text API.
2. Configure Authentication: Set up authentication to access the API securely.
3. Select a Recognition Method: Choose between asynchronous and synchronous recognition based on your application's needs.
4. Send Audio Data: Send the audio data you want to transcribe to the API.
5. Receive Transcriptions: Collect and process the transcribed text returned by the API.
6. Handle Results: Depending on your application, you can store, display, or further process the transcribed text.

2.2 iSpeech

2.2.1 Introduction

iSpeech stands as a prominent innovator in the realm of speech technology, specializing in both text-to-speech (TTS) and speech recognition solutions. With a core commitment to bridging the gap between human speech and technology, iSpeech's portfolio of products and services caters to a diverse array of industries and applications. Its text-to-speech technology converts written text into remarkably lifelike spoken language, making content accessible and interactive, particularly in the realms of accessibility services and e-learning platforms. Meanwhile, iSpeech's speech recognition expertise empowers applications to transcribe spoken words into text, improving efficiency and accuracy in transcription services, voice-controlled applications, and voice assistants. Multilingual support and the flexibility for users to personalize their interactions set iSpeech apart, even though there are challenges related to complexity in integration and associated costs. iSpeech's commitment to enhancing language interaction and accessibility marks it as a pivotal player in the evolving landscape of speech technology.

2.2.2 Merits, Demerits and Challenges

Merits:

1. Text-to-Speech Solutions: iSpeech offers text-to-speech conversion services that can be used to enhance accessibility in applications like screen readers and e-learning platforms.
2. Speech Recognition: The company provides speech recognition technology, which can be used for transcription services and voice-controlled applications.
3. Multilingual Support: iSpeech solutions often support multiple languages and dialects, making them versatile for global applications.
4. Customization: Users can often customize the voice and speaking style in text-to-speech applications, which can enhance the user experience.

Demerits:

1. Integration Complexity: Depending on the specific iSpeech product, integrating their solutions into existing applications can be complex and may require expertise in speech technology.
2. Cost: iSpeech services may involve licensing or subscription fees, which can be costly for businesses, particularly those with high usage.

Challenges:

1. Accuracy**:** Like any speech recognition technology, achieving high accuracy, especially in the presence of accents, background noise, and multiple speakers, can be a significant challenge.
2. Customization Limits**:** While customization is a merit, there may be limitations in how much the voice and speaking style can be tailored.

2.2.3 Implementation of iSpeech

1. Select the iSpeech Solution: Choose the specific iSpeech product or service that aligns with your application's needs, whether it's text-to-speech, speech recognition, or another solution.
2. Licensing or Subscription: Depending on the chosen solution, subscribe to or license the required iSpeech services.
3. Integration: Integrate the iSpeech technology into your application, which may involve using provided APIs and libraries.
4. Configuration: Configure the technology to meet your specific requirements, such as choosing the desired voice for text-to-speech.
5. Testing and Optimization: Test the implementation and optimize it for accuracy and user experience.
6. Deployment: Deploy the application with integrated iSpeech technology to serve your users.

2.3 OpenSMILE

2.3.1 Introduction

OpenSMILE, which stands for "Open-Source Speech and Music Interpretation by Large Space Extraction," is a robust and open-source audio feature extraction tool developed by the Multimedia Computing Group at the Technical University of Munich. This versatile software is a cornerstone in the realm of speech and audio signal processing, enabling the extraction of a wide array of acoustic and prosodic features from audio data. OpenSMILE's capabilities are employed across a spectrum of applications, including emotion recognition, voice authentication, and natural language processing. As an open-source resource, it is freely accessible, fostering collaboration and innovation within the global community. OpenSMILE's adaptability allows integration with various programming languages and libraries, making it indispensable for researchers, engineers, and developers seeking to delve into the intricacies of audio data, revealing the nuances of speech, and advancing numerous fields where sound analysis plays a crucial role.

2.3.2 Merits, Demerits and Challenges

Merits:

1. Extensive Feature Extraction**:** OpenSMILE provides a wide range of acoustic feature extraction capabilities, including prosodic features like pitch, intensity, and rhythm, as well as spectral and cepstral features.
2. Open-Source**:** It is open-source software, meaning that the source code is freely available, and the community can contribute to its development, making it accessible to researchers, developers, and practitioners.
3. Research and Education**:** OpenSMILE serves as a valuable tool for research and educational purposes, allowing individuals to experiment with and gain insights into acoustic feature extraction and analysis.
4. Integration**:** It can be integrated with various programming languages, libraries, and applications, providing flexibility for researchers and developers.

Demerits:

1. Complexity: While OpenSMILE offers powerful feature extraction capabilities, it can be complex to configure and utilize fully, particularly for those who are new to the field of speech and audio signal processing.
2. Resource-Intensive: Some of the feature extraction processes may be resource-intensive, which could impact performance when dealing with large datasets.

Challenges:

1. Expertise Requirement: Effective utilization of OpenSMILE often requires expertise in audio signal processing and feature extraction techniques.
2. Interpreting Features: Understanding the significance of the extracted features and how they relate to specific tasks, such as emotion recognition or speaker identification, can be challenging.

2.2.3 Implentation of OpenSMILE

1. Installation: Begin by installing the OpenSMILE software on your system. It is compatible with various operating systems.
2. Configuration: Configure OpenSMILE for the specific feature extraction tasks you require. This may involve selecting the appropriate configuration files and specifying the desired features.
3. Data Input: Provide the audio or speech data you want to analyze and extract features from. OpenSMILE supports various audio formats.
4. Feature Extraction: Execute the feature extraction process, which will generate feature vectors based on the provided audio data.
5. Analysis and Integration: Analyze the extracted features and integrate them into your applications or research projects. This step often involves using additional tools or libraries for further analysis and machine learning tasks.

CHAPTER 3

RESULTS AND DISCUSSION

CHAPTER 3

RESULTS AND DISCUSSION

The Voice Accent Detection Model is currently in the research phase, and our preliminary findings and discussions revolve around the examination of existing solutions addressing the same problem statement. Our research efforts have been dedicated to understanding the landscape of solutions available for identifying an individual's mother tongue from their English speech.

A central focus of our research has been the analysis of existing solutions and methodologies employed in similar domains. These solutions span a spectrum of techniques, ranging from traditional machine learning approaches to state-of-the-art deep learning models. This comprehensive analysis has provided valuable insights into the strengths and weaknesses of these existing solutions.

One key observation that has emerged from our research is the delicate balance between model complexity and accuracy. Many existing solutions leverage intricate deep learning architectures to achieve high levels of accuracy. However, this high accuracy often comes at the cost of increased model complexity, making deployment in real-world applications a resource-intensive task.

Another notable challenge evident in the landscape of existing solutions is their capacity to perform accurately in multilingual environments. Given the diverse linguistic backgrounds and varying levels of English fluency among speakers, accurately identifying the speaker's mother tongue becomes a multifaceted challenge.

Our discussions have also encompassed the practicality and ease of integration of existing solutions. Some solutions are tailored for specific applications and may lack versatility, while others offer flexibility for integration into a range of contexts.

CHAPTER 4
CONCLUSION

CHAPTER 4

CONCLUSION

In conclusion, the Voice Accent Detection Model, currently in the research phase, represents a pioneering effort to tackle the intricate task of identifying an individual's mother tongue from their English speech patterns. Our ongoing research, findings, and discussions have shed light on the model's significance and potential applications. Through a comprehensive analysis of existing solutions, we have observed the delicate balance between model complexity and accuracy, recognizing the challenges associated with multilingual environments. The model has shown initial promise with commendable accuracy, albeit with challenges when dealing with highly fluent English speakers. As we look forward, our commitment to enhancing accuracy, adaptability, and practicality remains steadfast. The model's impact extends to language learning tools, cross-cultural communication, and speech analysis, offering the potential to bridge linguistic gaps and provide valuable insights into the subtleties of the human voice. It is not merely a research project; it represents a path toward a more inclusive and interconnected global society. As we refine and expand its capabilities, we envision a future where the Voice Accent Detection Model plays a pivotal role in language education, cross-cultural understanding, and unraveling the intricate tapestry of accents and languages that define our world.

REFERENCES

REFERENCES

[1].https://scholarworks.calstate.edu/downloads/qz20sv442

[2].https://medium.com/analytics-vidhya/using-machine-learning-to-identify-accents-in-spectrograms-of-speech-5db91c191b6b

[3].https://www.mdpi.com/2227-7390/10/16/2913

[4].https://www.audeering.com/research/opensmile/

[5].https://www.ispeech.org/

[6].https://cloud.google.com/speech-to-text