MANIPAL INSTITUTE OF TECHNOLOGY
MANIPAL
(A constituent unit of MAHE, Manipal)

MACHINE LEARNING LAB MINI PROJECT SYNOPSIS ON

Speaker Diarization and Language Classification on Audio

SUBMITTED TO

Department of Computer Science & Engineering

by

Name	**Registration Number**	**Roll Number**	**Semester**
Sriram Sunderrajan	220962444	77	V
Anuj Kamath	220962446	78	V
Balaji Ramadhurai	220962448	79	V
Ankur Monga	220962137	81	V

Introduction

Vast amounts of audio data are generated daily, ranging from phone conversations and meetings to multimedia broadcasts and social media content. Extracting meaningful information from these audio streams is essential for various applications, such as transcription services, multilingual speech recognition, and content indexing. This project focuses on two crucial tasks in the audio processing domain: speaker diarization and language classification.

Speaker diarization is the process of identifying and segmenting an audio file to distinguish which portions are spoken by different speakers. This technique is essential for tasks like Automatic Speech Recognition (ASR), as time stamped speaker information helps provide context to the generated text. We can also leverage this information to better serve certain use cases, such as customer service and legal proceedings. In customer service, diarization enables detailed sentiment analysis of interactions between customers and representatives, improving service quality and customer satisfaction. In legal contexts, it ensures accurate transcription of court proceedings and interviews by attributing statements to the correct individuals, which is crucial for maintaining legal accuracy and record-keeping.

Language classification, on the other hand, seeks to identify the language spoken in a given audio segment. This information is crucial for ASR models, as it allows them to handle audio files where speakers might switch between languages or face other challenging scenarios.

Overall, this project can enable more performant and flexible ASR systems, leading to a more inclusive and language-agnostic internet. This advancement is also significant for improving accessibility and ensuring that diverse linguistic communities can engage more fully with digital content.

Literature Survey

Recent literature on speaker diarization has focused on comparing the effectiveness of i-vector and d-vector techniques, with an emphasis on clustering algorithms. This survey integrates foundational methods like Gaussian Mixture Models (GMMs) and Hidden Markov Models (HMMs) with more advanced approaches.

1. GMM-Based Methods: GMMs model the distribution of acoustic features within a speech segment through a mixture of Gaussian distributions, each capturing different speaker characteristics. By calculating the likelihood of a speech segment belonging to a particular speaker model, GMMs help group segments with similar characteristics. However, GMMs struggle with temporal dynamics, treating each segment independently, which can lead to inaccuracies in segment boundaries and speaker assignments, especially under challenging acoustic conditions.

2. HMM-Based Methods: HMMs address the temporal limitations of GMMs by introducing a sequence of hidden states to model the underlying temporal structure of speech. HMMs can capture transitions between different acoustic states within a speaker's speech, integrating with GMMs to provide a more comprehensive model of both spectral and temporal information. This

integration improves diarization accuracy, particularly in noisy environments or when speaker characteristics vary over time.

3. i-Vector Based Clustering: i-Vector techniques involve extracting compact speaker representations known as i-vectors from GMM-based models. These i-vectors encapsulate speaker-specific information in a low-dimensional space. i-vectors effectively simplify the clustering process while retaining crucial speaker information.

4. d-Vector Based Clustering: Like i-vectors, d-vectors are clustered to identify speaker segments. d-Vectors generally offer higher accuracy due to their ability to model intricate patterns in audio data through deep learning techniques.

5. End-to-End Diarization Systems: We can compare our performance with end-to-end systems that utilize a single deep learning model to handle the entire diarization process, including segmentation and speaker assignment.

Evaluation Challenges:
Comparing diarization systems remains challenging due to variations in datasets, preprocessing techniques, and evaluation metrics. Factors like voice activity detection, training data quality, and handling of overlapping speech can significantly influence diarization error rates (DER), necessitating careful consideration of these factors in result interpretation.

Research Gaps & Objectives

Despite advancements in speaker diarization and language classification, several challenges remain:

1. Speaker Overlap: Handling overlapping speech, where multiple speakers talk simultaneously, is still challenging. Current models often struggle in these scenarios, leading to decreased accuracy . Research specifically targeting this issue is limited.

2. Diverse Accents and Languages: Variability in accents, dialects, and languages impacts both diarization and language classification accuracy. There is a gap in developing models that generalize well across different linguistic features, particularly for underrepresented languages and accents.

3. Scalability: Efficiently processing large volumes of audio data remains difficult. Existing systems for both tasks struggle to scale without compromising accuracy, indicating a need for more research in creating scalable frameworks.

To address these gaps, the project aims to:

1. Develop a Diarization System for Overlapping Speech: Create and evaluate a model that can effectively segment speakers in overlapping speech scenarios.
2. Improve Performance Across Languages and Accents: Enhance the models' ability to generalize across different languages and accents for both diarization and language classification.
3. Expand Resources: Curate or create comprehensive datasets that address the diversity and complexity of real-world audio data, contributing to the research community.

These objectives will address key challenges in speaker diarization and language classification, advancing the field with more robust, scalable, and versatile models.

Dataset Description/Method of Creation

Since speaker diarization will be approached using unsupervised methods, we will not require labeled training data for this task. Instead, our focus will be on collecting and curating a dataset for the language classification component of the project. This dataset will include:

1. Diverse Language Samples: A comprehensive collection of audio recordings in multiple languages will be sourced from publicly available datasets such as the Mozilla Common Voice project, which offers a large-scale, high-quality dataset of speech in various languages. Another valuable resource is the Linguistic Data Consortium (LDC) catalog, which provides a wide range of multilingual speech datasets that have been rigorously annotated and vetted for research purposes.
2. Noisy and Clean Audio Conditions: To ensure the robustness of the language classification model, recordings will be curated to include both clean audio and audio with various levels of background noise. This diversity will help in training the model to perform well in real-world scenarios where audio quality can vary significantly.
3. Custom Recordings: To introduce further variability and to address specific project needs, additional recordings will be created. These will include scenarios with multilingual speakers and overlapping speech, reflecting more complex and realistic audio environments. This will help in making the language classification model more resilient to challenging conditions.

By leveraging these datasets, the project aims to create a robust and diverse language classification model that can perform well across different audio conditions and linguistic contexts.

Expected Output

The expected outputs of this project include:

1. A speaker diarization model capable of accurately segmenting and identifying speakers in multi-speaker audio recordings, including those with overlapping speech. Its functions would be
    a) Accurately Segments Speech: Capable of distinguishing and segmenting different speakers even in overlapping scenarios.
    b) Generalizes Across Languages: Shows improved performance across various languages and accents.
    c) Handles Large Datasets: Demonstrates scalability by efficiently processing large volumes of audio data.

2. A language classification model that can correctly identify the language spoken in a given audio segment, with high accuracy even in noisy conditions.

3. A combined framework that integrates both speaker diarization and language classification, providing a comprehensive tool for analyzing multilingual audio data.

4. Confidence Score Parameter: A confidence score parameter will be implemented to assign proper weightage to the correctness of the results, providing users with an indication of the reliability of the speaker and language identification outputs. This feature will enhance the interpretability and usability of the system by offering a quantitative measure of certainty in its predictions.

References

- D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," in *IEEE Transactions on Speech and Audio Processing.*

- H. Meinedo and J. Neto, "Audio segmentation, classification and clustering in a broadcast news task," *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing.*

- Singh, Prachi and Sriram Ganapathy. "Overlap-aware End-to-End Supervised Hierarchical Graph Clustering for Speaker Diarization," *ArXiv*

- Gupta, A., Purwar, A. "Speech refinement using Bi-LSTM and improved spectral clustering in speaker diarization," *Multimed Tools Appl*

- Moattar, M. H., & Homayounpour, M. M. (2012). A review on speaker diarization systems and approaches. *Elsevier BV*