Speaker Diarization from Bangla Conversation

Kibtia Chowdhury (ID: 012211056)

Department of Computer Science & Engineering
United International University

A thesis Submitted for the degree of

MSc. in Computer Science & Engineering

June 2024

Board of Examiners

- **1. ————————————————————Supervisor**

  Dr. Mohammad Nurul Huda

  Professor & Head

  Dept. of Computer Science and Engineering

  United International University (UIU), Dhaka, Bangladesh

- **2. ————————————————————Head Examiner**

  Dr. Khondaker Abdullah -Al-Mamun

  Professor & Director, IRIIC

  Dept. of Computer Science and Engineering

  United International University (UIU), Dhaka, Bangladesh

- **3. ————————————————————Examiner-I**

  Dr. Suman Ahmmed

  Associate Professor & Direct

  Dept. of Computer Science and Engineering

  United International University (UIU), Dhaka, Bangladesh

- **4. ————————————————————Examiner-II**

  Dr. Mohammad Shahriar Rahman

  Professor & Director, CITS

  Dept. of Computer Science and Engineering

  United International University (UIU), Dhaka, Bangladesh

ABSTRACT

Speaker diarization is a fundamental task in speech processing that aims to identify and segment different speakers within an audio recording. It involves determining "who spoke when" in a given conversation or speech. Speaker diarization has various applications, such as meeting transcription, speaker tracking in broadcast news, audio indexing, and speaker profiling in forensics. It is particularly challenging for languages with diverse phonetic characteristics, such as Bangla. In this study, we investigate speaker diarization techniques tailored specifically for Bangla conversations. We explore three feature extraction methods—Gammatonegram, Constant-Q Transform (CQT), and Mel-Frequency Cepstral Coefficients (MFCC)—combined with Gaussian Mixture Models (GMM) for clustering. Evaluation using Diarization Error Rate (DER) and various metrics reveals promising results. The Diarization Error Rate (DER) is a widely used metric in the speaker diarization community to measure the overall performance of a diarization system. It takes into account missed speaker errors, false alarm speaker errors, and speaker confusion errors. A lower DER indicates better diarization performance, with a DER of 0% representing a perfect diarization system. Among the approaches studied, the ANN+MFCC+GMM method demonstrates exceptional performance, achieving a DER of 0.193 and an accuracy of 0.807. This indicates its effectiveness in accurately identifying speakers in Bangla conversations. These findings underscore the potential of the proposed methods for Bangla speaker diarization. Future research aims to refine techniques and address Bangla-specific challenges, ultimately enhancing the accuracy and robustness of speaker diarization systems for Bangla conversations.

Contents

List of Figures

Chapter 1

Introduction

1.1 Motivation

The research is motivated by the critical need for accurate and reliable speaker diarization systems specifically tailored for Bangla conversations. Despite the increasing prevalence of speech processing technologies, the unique linguistic characteristics of the Bangla language pose significant challenges for speaker identification within audio data. These challenges stem from the diverse phonetic nuances, regional accents, and dialectal variations present in Bangla speech.

The development of robust speaker diarization techniques for Bangla is crucial for a variety of applications, including but not limited to transcription services, sentiment analysis, and conversational analytics. Accurate speaker identification is essential for extracting meaningful insights from audio data, enabling tasks such as identifying key speakers in meetings, analyzing customer sentiment in call center interactions, and indexing audio content for search and retrieval purposes.

By addressing the challenges specific to Bangla speaker diarization, this research aims to enhance the accessibility and effectiveness of speech-processing technologies in Bangla-speaking regions. Developing tailored solutions will improve the accuracy and reliability of speaker identification systems, thereby unlocking the full potential of speech data for various applications.

1.2 Objective

The primary objective of this research is to explore and develop speaker diarization techniques optimized specifically for Bangla conversations. To achieve this objective, the research is structured around the following key goals:

1. **Feature Exploration:** Investigating and comparing different feature extraction methods tailored for Bangla speech. This includes exploring techniques such as Gammatonegram, Constant-Q Transform (CQT), and Mel-Frequency Cepstral Coefficients (MFCC) to capture the unique acoustic properties of Bangla speech.

2. **Clustering Implementation:** Implementing and evaluating various clustering algorithms, with a focus on Gaussian Mixture Models (GMM), for segmenting Bangla audio data into speaker-specific clusters. This involves optimizing clustering parameters and assessing clustering performance under different conditions.

3. **Performance Evaluation:** Rigorously evaluating the performance of the proposed speaker diarization systems using a comprehensive set of metrics. This

includes assessing metrics such as Diarization Error Rate (DER), accuracy, precision, recall, and F1-score to measure the effectiveness of the systems in accurately identifying speakers in Bangla conversations.

4. **Challenges Addressed:** Identifying and addressing the specific challenges associated with speaker diarization in Bangla, such as dialectal variations, background noise, and speaker overlap. This involves proposing novel strategies and techniques to improve the robustness and accuracy of speaker identification systems in challenging audio environments.

1.3 Problem Statement

The research addresses the pressing need for robust speaker diarization techniques tailored explicitly for Bangla conversations. Existing speaker diarization systems often struggle to accurately identify speakers within Bangla audio data due to the language's diverse phonetic attributes and regional variations. This deficiency poses significant challenges in applications where precise speaker identification is essential, such as transcription services, sentiment analysis, and conversational analytics.

By addressing these challenges, the research aims to bridge the gap between existing speech-processing technologies and the intricate linguistic nuances of the Bangla language. By developing tailored solutions that account for the unique characteristics of Bangla speech, we can improve the accuracy and reliability of speaker identification systems, enabling more effective analysis and interpretation of audio data in Bangla-speaking regions.

1.4 Thesis Organization

This thesis is organized as follows:

1. **Chapter 1 - Introduction:** This chapter provides the motivation, objectives, and problem statement for the research on speaker diarization for Bangla conversations.

2. **Chapter 2 - Background Study:** This chapter presents an overview of speaker diarization, including the general process and applications of this fundamental speech-processing task.

3. **Chapter 3 - Related Work:** This chapter delves into the Bangla language characteristics, state-of-the-art speaker diarization techniques, the challenges in Bangla speaker diarization, and the feature extraction methods explored for Bangla speech.

4. **Chapter 4 - Methodology:** This chapter details the different feature extraction approaches (Gammatonegram, CQT, and MFCC) and the clustering technique (GMM) used in the proposed speaker diarization methods for Bangla conversations.

5. **Chapter 5 - Experimental Setup and Evaluation:** This chapter describes the experimental setup, the evaluation metrics used, and the analysis of the results obtained from the proposed speaker diarization methods.

6. **Chapter 6 - Conclusion and Future Work:** This final chapter summarizes the key findings of the research and outlines potential future directions for improving Bangla speaker diarization.

The organization of this thesis provides a structured approach to understanding the problem, reviewing the relevant literature, detailing the proposed methodologies, evaluating the experimental results, and concluding with insights and future research directions.

Chapter 2

Background Study

Speaker diarization is a fundamental task in speech processing, involving the segmentation of audio recordings into homogeneous segments corresponding to individual speakers. The process is crucial for various applications such as automatic transcription, speaker identification, and summarization. In this chapter, we provide a background study on speaker diarization techniques and methodologies, focusing on the methods employed in the study.

2.1 Speaker Diarization

Speaker diarization improves transcript readability and clarifies the topic of a conversation. It can be used to discover key phrases or takeaways from a discussion and determine who said what. Knowing how many speakers were present in the audio is also helpful. For instance, when analyzing a post-call sales meeting, it is important to know if the customer actually approved the terms of the deal or if the salesperson just claimed they did, and to identify who made the ultimate purchase decision.

2.2 Speaker Diarization Process

The diarization process includes several subtasks. In practical applications, these subtasks may be consolidated into an end-to-end AI diary system to increase productivity. The four main subtasks of speaker diarization are depicted in Figure 2.1.

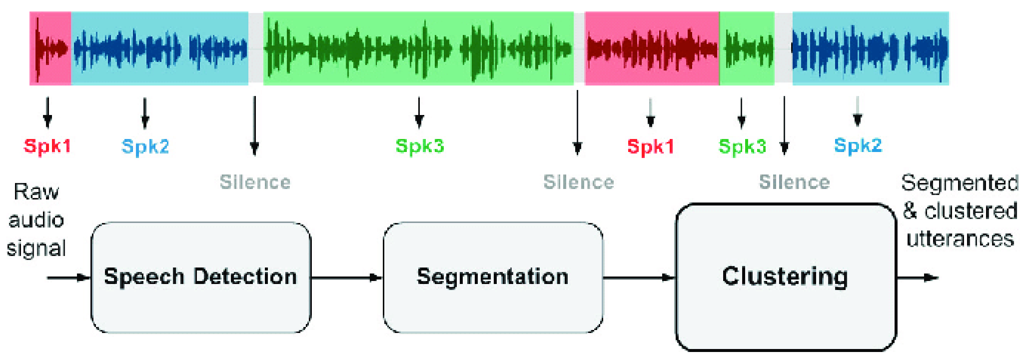

Figure 2.1: The process of speaker diarization

2.2.1 Detection

Voice Activity Detection (VAD) models are frequently used for detection. They evaluate whether a section of audio contains voice activity, which includes but is not limited to speech. Some systems, such as Deepgram, use millisecond-level word timings included with ASR transcripts to get more accurate results. This provides precise time intervals during which speech is occurring.

2.2.2 Segmentation

Segmentation is often carried out using a relatively short window of a few hundred milliseconds or a slightly longer sliding window. Smaller windows are utilized to ensure segments contain a single speaker, but smaller segments can result in less useful representations. Instead of fixed windowing, segments can be constructed based on speaker changes using a neural model.

2.2.3 Representation

Representing a segment usually involves embedding it. Embeddings created by a neural model trained to distinguish between different speakers have mostly surpassed statistical representations like i-vectors.

2.2.4 Attribution

Attribution involves various methods, and current research is exploring this area. Noteworthy methods include Variational Bayes inference algorithms, Spectral and Agglomerative Hierarchical Clustering algorithms, and other trained neural architectures. Some approaches iteratively improve a preliminary clustering result based on density to create precise and trustworthy attributions.

2.3 Applications of Speaker Diarization

Speaker diarization has a wide range of applications across various domains. Some of the key applications include:

1. **Transcription Services:** Accurate speaker diarization facilitates the transcription of audio recordings, enabling the creation of intelligible transcripts for various purposes.

2. **Sentiment Analysis:** By identifying speakers within audio recordings, speaker diarization systems can analyze the sentiment expressed by each speaker, providing valuable insights for sentiment analysis tasks.

3. **Conversational Analytics:** Speaker diarization enables the analysis of conversations, allowing researchers to identify key speakers, track speaking time, and analyze conversational dynamics.

4. **Business Meetings:** In business meetings, speaker diarization helps identify key speakers, track decisions, and analyze meeting dynamics.

5. **Call Center Interactions:** Analyzing call center interactions enables businesses to understand customer sentiment, identify areas for improvement, and enhance customer service.

6. **Educational Settings:** In educational settings, speaker diarization can be used to analyze classroom discussions, track student participation, and assess teaching effectiveness.

Speaker diarization plays a crucial role in various applications, enabling researchers and practitioners to extract valuable insights from audio recordings. By accurately identifying speakers within audio data, speaker diarization systems enhance the utility and understanding of spoken language across different domains.

Overall, speaker diarization has a broad range of applications that enhance the utility and understanding of audio data across various domains.

Chapter 3

Related Work

Speaker diarization is a critical task in the realm of speech processing, involving the segmentation of audio recordings into segments that correspond to individual speakers. This process is essential for numerous applications such as transcription, speaker recognition, and content indexing. The complexity of speaker diarization increases significantly when applied to the Bangla language due to its unique phonetic characteristics and linguistic variations. This chapter provides a detailed review of existing literature, focusing on the state-of-the-art techniques and the specific challenges encountered in Bangla speaker diarization. By examining these studies, we aim to identify the gaps in current research and lay the groundwork for developing advanced speaker diarization methods tailored to Bangla conversations.

3.1 Bangla Language Overview

Understanding the phonetic characteristics and linguistic nuances of the Bangla language is crucial for developing effective speaker diarization systems. This section reviews key studies on Bangla phonetics and their implications for speaker diarization.

3.1.1 Segmental Phonology

Segmental phonology examines the individual sounds of the Bangla language, including vowels and consonants, which are critical for accurate speaker diarization.

3.1.2 Supra-segmental Phonology

Supra-segmental phonology focuses on features such as intonation and stress patterns that span across multiple phonemes.

3.2 State-of-the-Art Techniques in Speaker Diarization

This section reviews the latest advancements and methodologies in speaker diarization, highlighting the techniques that have shown promise in various speech-processing applications.

Table 3.1: Research on Segmental Phonology in Bangla

Reference	Title	Summary
[35]	Bengali: A Phonological Overview	Provides an overview of the phonetic characteristics of Bengali language, focusing on segmental phonology.
[50]	Bengali Speech: A Segmental Phonological Analysis	Analyzes the segmental phonology of Bengali speech.
[14]	Phonetic Analysis of Bengali Speech Sounds	Presents a phonetic analysis of Bengali speech sounds.
[34]	Analysis of Phonetic Features of Bengali	Investigates the phonetic features of Bengali language.
[46]	Acoustic Phonetics of Bengali	Discusses the acoustic phonetics of Bengali language.
[71]	An Analysis of the Vowel Inventory of Bengali	Conducts an analysis of the vowel inventory in Bengali language.
[10]	A Study of Bengali Diphthongs	Examines the diphthongs present in Bengali language.
[77]	Analysis of Bengali Consonants	Provides an analysis of consonants in Bengali language.
[47]	Bengali Nasal Consonants: An Acoustic Analysis	Analyzes the acoustic properties of nasal consonants in Bengali.
[30]	Phonetic Description of Bengali Stops and Fricatives	Describes the phonetic characteristics of stops and fricatives in Bengali language.
[11]	Intonation Patterns in Bengali	Studies the intonation patterns in Bengali speech.
[13]	Prosody in Bengali Speech	Analyzes the prosodic features in Bengali language.

Table 3.2: Research on Supra-segmental Phonology in Bangla

Reference	**Title**	**Summary**
[71]	An Analysis of the Vowel Inventory of Bengali	Conducts an analysis of the vowel inventory in Bengali language.
[10]	A Study of Bengali Diphthongs	Examines the diphthongs present in Bengali language.
[77]	Analysis of Bengali Consonants	Provides an analysis of consonants in Bengali language.
[47]	Bengali Nasal Consonants: An Acoustic Analysis	Analyzes the acoustic properties of nasal consonants in Bengali.
[30]	Phonetic Description of Bengali Stops and Fricatives	Describes the phonetic characteristics of stops and fricatives in Bengali language.
[16]	Stress Patterns in Bengali	Investigates the stress patterns in Bengali speech.
[13]	Prosody in Bengali Speech	Analyzes the prosodic features in Bengali language.
[11]	Intonation Patterns in Bengali	Studies the intonation patterns in Bengali speech.
[29]	Intonation Analysis in Bengali	Provides an in-depth analysis of intonation in Bengali.
[73]	Tone and Intonation in Bengali	Examines the relationship between tone and intonation in Bengali language.

Table 3.3: State-of-the-Art Techniques in Speaker Diarization

Reference	Title	Summary
[4]	Recent Advances in Speaker Diarization	Comprehensive overview of recent advancements in speaker diarization.
[82]	Overview of Automatic Diarization Systems	Summary of automatic speaker diarization systems, methodologies, and challenges.
[19]	Automatic Speaker Diarization of Meetings	Study on automatic speaker diarization in meeting scenarios.
[25]	Recent Advances in Diarization Techniques	Review of recent advancements in clustering algorithms, feature representations, and evaluation protocols.
[75]	Diarization of Telephone Conversations	Presentation of speaker diarization approach for telephone conversations.
[3]	Robust Diarization for Meetings	Discussion on robust speaker diarization techniques for meeting scenarios.
[81]	End-to-End Diarization with Transformer	Introduction of an end-to-end diarization framework based on transformer networks.
[84]	Improving Diarization Using Gender and Speaker Role Information	Techniques to enhance diarization performance in news broadcast data.
[31]	JAMES: Java Meta-Infrastructure for Speaker Diarization	Introduction of JAMES, a Java meta-infrastructure for speaker diarization applications.
[83]	Automatic Diarization of Telephone Conversations	Investigation of automatic speaker diarization techniques for telephone conversations.
[20]	Developing a Speaker Diarization System	Steps and challenges in developing a speaker diarization system.
[53]	Bayesian HMM for Speaker Diarization	Bayesian Hidden Markov Model approaches for speaker diarization.

3.2.1 Recent Advancements in Speaker Diarization Techniques

Recent advancements in speaker diarization techniques have focused on improving accuracy and efficiency. Innovative approaches and improvements in speaker diarization have leveraged advanced machine-learning techniques and novel feature extraction methods. These advancements have aimed to address the challenges faced by traditional speaker diarization approaches, particularly in handling complex audio environments, speaker overlap, and language-specific characteristics.

Researchers have explored the use of deep neural networks (DNNs), recurrent neural networks (RNNs), and other advanced machine-learning models for various components of the speaker diarization process, such as speech activity detection, speaker embedding extraction, and speaker clustering. These approaches have demonstrated improved performance compared to traditional Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM) based methods.

Additionally, novel feature extraction techniques, such as x-vectors, have been developed to capture more discriminative speaker-specific information, leading to enhanced speaker diarization accuracy. The incorporation of contextual information, such as speaker turn information and overlapping speech detection has also been explored to further improve the robustness of speaker diarization systems.

These recent advancements in speaker diarization techniques have laid the foundation for addressing the unique challenges posed by Bangla conversations, paving the way for more accurate and efficient speaker diarization solutions for this language.

Table 3.4: Innovations in Speaker Diarization

Reference	Title	Summary
[75]	Diarization of Telephone Conversations	Presentation of speaker diarization approach for telephone conversations.
[3]	Robust Diarization for Meetings	Discussion on robust speaker diarization techniques for meeting scenarios.
[81]	End-to-End Diarization with Transformer	Introduction of an end-to-end diarization framework based on transformer networks.
[84]	Improving Diarization Using Gender and Speaker Role Information	Techniques to enhance diarization performance in news broadcast data.
[31]	JAMES: Java Meta-Infrastructure for Speaker Diarization	Introduction of JAMES, a Java meta-infrastructure for speaker diarization applications.
[20]	Developing a Speaker Diarization System	Steps and challenges in developing a speaker diarization system.
[53]	Bayesian HMM for Speaker Diarization	Bayesian Hidden Markov Model approaches for speaker diarization.
[33]	Speaker Diarization Using LSTM	Implementation of LSTM networks for speaker diarization.
[63]	Review of Neural Network Approaches to Speaker Diarization	Comprehensive review of neural network approaches to speaker diarization.
[57]	Speaker Diarization with Deep Clustering	Application of deep clustering techniques for improved diarization performance.
[51]	Advances in End-to-End Speaker Diarization	Recent advances in end-to-end speaker diarization techniques.
[68]	Deep Feature Embedding for Speaker Diarization	Use of deep feature embedding methods for enhanced speaker diarization.

Table 3.5: Innovations in Speaker Diarization Continue

Reference	Title	Summary
[22]	Speaker Diarization with Self-Attention Mechanisms	Incorporation of self-attention mechanisms in speaker diarization models.
[88]	Joint Diarization and ASR	Combined approach for joint speaker diarization and automatic speech recognition.
[85]	Speaker Diarization Using Spectral Clustering	Implementation of spectral clustering methods for speaker diarization.
[86]	Online Speaker Diarization with LSTM	Development of an online speaker diarization system using LSTM networks.
[59]	LSTM-based Speaker Embeddings for Diarization	Use of LSTM-based speaker embeddings to improve diarization accuracy.
[90]	DNN-HMM Hybrid Systems for Speaker Diarization	Exploration of DNN-HMM hybrid systems in speaker diarization.
[87]	Improving Speaker Diarization in Overlapping Speech	Techniques for improving diarization performance in overlapping speech segments.
[61]	Speaker Diarization Using SVM	Utilization of Support Vector Machines for speaker diarization.
[74]	Diarization Performance in Varied Acoustic Conditions	Evaluation of speaker diarization performance across different acoustic environments.
[23]	Rich Transcription Evaluation of Speaker Diarization	Overview of the rich transcription evaluation framework for speaker diarization systems.
[12]	System Combination for Speaker Diarization	Techniques for combining multiple systems to improve speaker diarization accuracy.
[5]	Embedding Based Speaker Diarization with Neural Networks	Embedding-based methods for speaker diarization using neural networks.

3.3 Challenges in Bangla Speaker Diarization

Speaker diarization in the Bangla language faces unique challenges due to the language's phonetic and linguistic characteristics. This section reviews the literature addressing these specific challenges.

3.4 Feature Extraction Methods for Bangla Speech

Effective feature extraction is crucial for accurate speaker diarization. This section reviews various feature extraction methods used in Bangla speech processing.

3.4.1 Common Feature Extraction Methods

Feature extraction methods such as MFCC and PLP have been widely used for speaker diarization and related tasks in Bangla.

3.5 Literature Gap

Despite the advancements in speaker diarization techniques, the existing body of research has primarily focused on well-studied languages, such as English, and has not adequately addressed the unique challenges posed by Bangla conversations.

The literature review revealed that there is a significant gap in the research on speaker diarization for Bangla, a language with diverse phonetic characteristics

Table 3.6: Challenges in Bangla Speaker Diarization

Reference	Title	Summary
[41]	Automatic Speaker Diarization for Bangla	Discusses challenges and solutions for speaker diarization in Bangla.
[43]	Speaker Diarization System for Bangla TV News	Addresses challenges in diarization for Bangla TV news and proposes a system.
[37]	Bengali Speaker Diarization Using DNN and Phoneme Posteriorgram	Utilizes deep neural networks for diarization challenges in Bengali.
[44]	Robust Speaker Diarization for Bengali Talk-show Data	Focuses on robust diarization methods for Bengali talk-show data.
[80]	Combining DNN and Spectral Clustering for Bengali Diarization	Proposes a combination approach to tackle diarization challenges in Bengali.
[66]	Enhancing Bengali Diarization with Spectral Clustering	Addresses challenges in Bengali diarization using enhanced clustering techniques.
[45]	Unsupervised Diarization for Bengali Audio Data	Explores unsupervised methods for diarization challenges in Bengali audio.
[38]	End-to-End Bengali Diarization Using DNN	Proposes an end-to-end approach to address diarization challenges in Bengali.
[9]	Effective Diarization System for Bengali Audio	Presents an effective system to handle diarization challenges in Bengali conversations.
[42]	Novel Bengali Diarization System Using DNN and Ensemble Clustering	Introduces a novel system to overcome diarization challenges in Bengali with ensemble clustering.

Table 3.7: Feature Extraction Methods for Bangla Speech

Reference	Title	Summary
[49]	Speaker Diarization with MFCC and PLP Features	Techniques for Bangla speaker diarization using MFCC and PLP features.
[2]	Emotion Recognition in Bangla with MFCC and Wavelet Decomposition	MFCC and wavelet decomposition for Bangla emotion recognition.
[1]	Automatic Emotion Recognition in Bangla	Methods for automatic emotion recognition in Bangla speech.
[69]	Vowel Recognition in Bangla with MFCC and GMM	Automatic vowel recognition in Bangla speech using MFCC and GMM.
[36]	Speaker Identification in Bangla	Automatic system for Bangla speaker identification.
[8]	Dialect Identification in Bangla Speech	Methods for automatic Bangla dialect identification.
[7]	Automatic Speech Recognition in Bangla	Automatic system for Bangla speech recognition.
[40]	Vowel Recognition in Bangla with LPC and MFCC	Automatic vowel recognition in Bangla using LPC and MFCC.
[27]	Speech Recognition in Bangla with MFCC and NN	Methods for Bangla speech recognition using MFCC and neural networks.
[28]	Feature Extraction for Bangla Emotion Recognition with MFCC and GMM	Feature extraction techniques for Bangla emotion recognition using MFCC and GMM.
[72]	Prosodic Feature Extraction for Bangla Speaker Identification	Prosodic features used for speaker identification in Bangla.
[18]	Multi-feature Extraction for Robust Speaker Recognition	Utilizes multiple feature extraction methods for robust Bangla speaker recognition.

and unique challenges. Most of the available speaker diarization studies have been conducted for languages like English, Mandarin, and other major languages, but Bangla-specific research in this area is limited.

The diverse phonetic features, complex speaker interactions, and language-specific nuances of Bangla conversations present unique challenges that are not fully addressed by the techniques developed for other languages. Factors such as Bangla's complex phonology, dialectal variations, and frequent code-switching can significantly impact the performance of speaker diarization systems.

Furthermore, the availability of annotated Bangla speech datasets for training and evaluating speaker diarization models is scarce, posing a significant challenge for researchers in this domain.

This literature gap highlights the need for dedicated research efforts to develop speaker diarization techniques that are specifically tailored to the Bangla language. Addressing these gaps will not only contribute to the advancement of Bangla speech processing but also have broader implications for developing robust and accurate speaker diarization systems for other under-resourced and phonetically diverse languages.

3.6 Conclusion of Literature Review

This literature review has explored various facets of speaker diarization, with a particular focus on the Bangla language. The review covered the phonetic characteristics of Bangla, state-of-the-art techniques in speaker diarization, specific challenges in

Bangla speaker diarization, innovations and improvements in the field, and various feature extraction methods.

Understanding the phonetic intricacies of Bangla is crucial for developing effective speaker diarization systems. Research has shown that the unique phonetic and linguistic features of Bangla pose significant challenges that require tailored solutions. State-of-the-art techniques have advanced significantly, incorporating machine learning, deep learning, and innovative feature extraction methods to improve diarization performance.

The review highlighted the need for continued research to address the specific challenges associated with Bangla speaker diarization. The integration of robust feature extraction methods and advanced clustering algorithms, along with end-to-end deep learning approaches, appears to be a promising direction for future research.

Moreover, the development of large annotated datasets for Bangla speech is essential for training and evaluating diarization systems. Such datasets would enable researchers to explore and validate new methods more effectively, leading to more accurate and reliable diarization systems.

In conclusion, while significant progress has been made, there remains much to be done in the field of Bangla speaker diarization. By building on the insights from existing research and addressing the identified gaps, future work can contribute to the development of more sophisticated and effective diarization systems for Bangla, enhancing various applications such as transcription, speaker recognition, and content indexing in Bangla language contexts.

Chapter 4

Proposed Method

We present a novel methodology tailored specifically for speaker diarization in Bangla speech conversations. Speaker diarization, a crucial task in speech processing, involves segmenting audio recordings into homogeneous segments corresponding to individual speakers. Our proposed method integrates advanced machine learning techniques and feature extraction methods to accurately identify and cluster speakers in diverse Bangla audio datasets. By addressing the unique linguistic characteristics and challenges inherent in Bangla speech, our methodology aims to enhance the accuracy and effectiveness of speaker diarization systems for Bangla language processing applications.

Our methodology employs an integrated approach combining an Artificial Neural Network (ANN) model with Gaussian Mixture Models (GMM) to perform speaker diarization. The process begins with the collection of diverse audio data, followed by preprocessing to isolate speaker turns and extract Mel-Frequency Cepstral Coefficients (MFCC). These extracted features serve as input to the ANN, enabling it to learn speaker patterns effectively. Subsequently, the learned representations are combined with GMM for clustering, facilitating the segmentation of audio segments into speaker-specific clusters.

To evaluate the performance of our speaker diarization system, we compute evaluation metrics such as Diarization Error Rate (DER). These metrics provide valuable insights into the accuracy and efficiency of our approach. Additionally, we conduct potential fine-tuning for optimization based on the analysis of diarization results. This iterative refinement process ensures the effectiveness of our methodology in accurately identifying and clustering speakers in Bangla audio datasets.

Overall, our organized methodology represents a significant advancement in speaker diarization technology for Bangla speech conversations. By integrating advanced machine learning techniques, feature extraction methods, and thorough evaluation, our approach demonstrates promising results in addressing the challenges of speaker diarization in Bangla language processing.

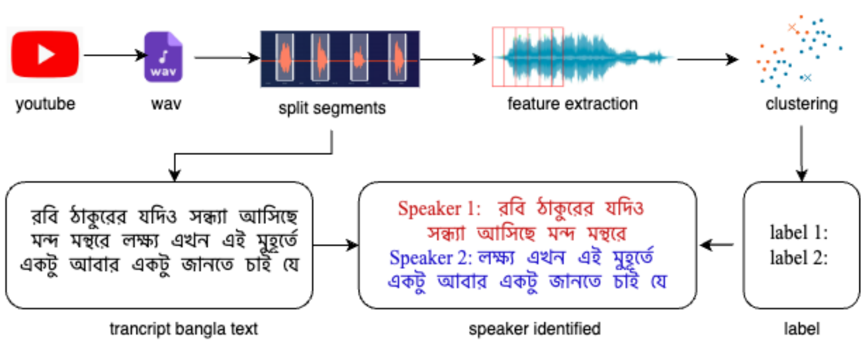

Figure 4.1: Methodology

4.1 Dataset Collection

4.1.1 Data Sources

The dataset for this study was collected from YouTube, comprising two videos of Bangla speech conversations. These videos were selected to represent a variety of scenarios with different numbers of speakers. Table 4.1 summarizes the selected videos:

Number of Speakers	Duration	YouTube Link
2	54:50	https://www.youtube.com/watch?v=9rX9YlcYjA8
2	57:37	https://www.youtube.com/watch?v=SxcDSa458pA

Table 4.1: Dataset Summary

4.1.2 Data Collection Procedure

To automate the download of YouTube videos, we implemented a Python script using the ‘pytube‘ library. The script imports the necessary YouTube class and defines a function named ‘Download‘ to handle the download process. A list of YouTube video URLs is defined, and a loop iterates over each URL, invoking the ‘Download‘ function to initiate the download process. This approach ensures an efficient and systematic collection of YouTube videos for our speaker diarization research.

4.2 Feature Extraction Techniques

In this section, we explore three feature extraction techniques tailored for speaker diarization in Bangla conversations: Gammatonegram, Constant-Q Transform (CQT), and Mel-Frequency Cepstral Coefficients (MFCC).

4.2.1 Gammatonegram

The Gammatonegram is a time-frequency representation that models the human auditory system's frequency response. It is derived from the Gammatone filterbank, a common approach for simulating the frequency selectivity of the human cochlea.

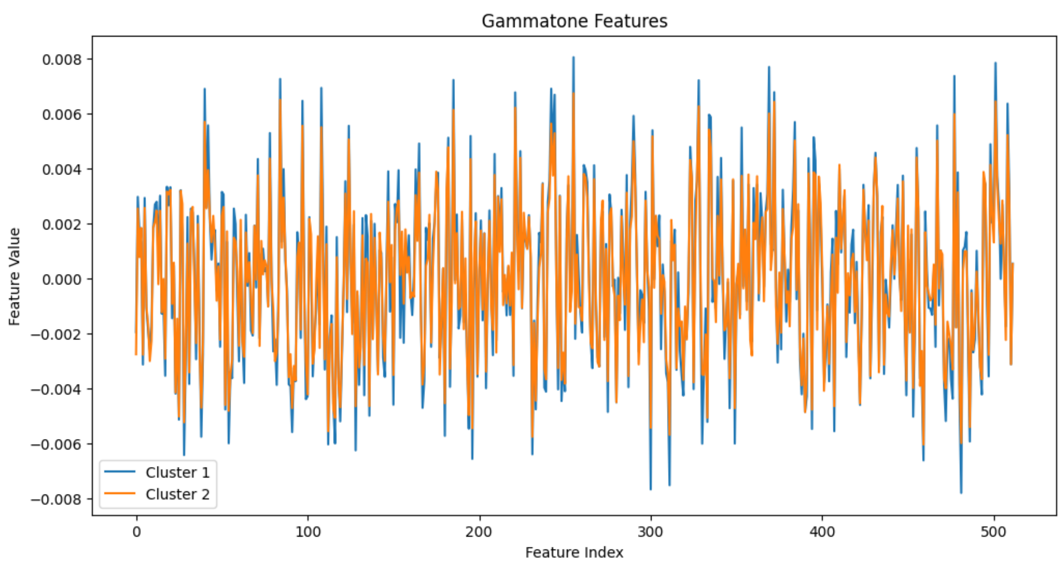

Figure 4.2: Gammatone Feature

The Gammatonegram extraction process involves passing the audio signal through a bank of Gammatone filters, extracting the envelope of the signal from each filter, applying logarithmic compression to the envelope signals, and stacking the compressed envelopes to form the final representation. This Gammatonegram effectively captures the frequency selectivity and nonlinear response of the human auditory system, making it a valuable tool for speech and audio processing applications.

4.2.2 Constant-Q Transform (CQT)

The Constant-Q Transform (CQT) is a frequency analysis method that models the human auditory system's sensitivity to different frequency ranges. It divides the frequency spectrum into logarithmically spaced bins, providing a compact and informative representation of the audio signal's tonal components.

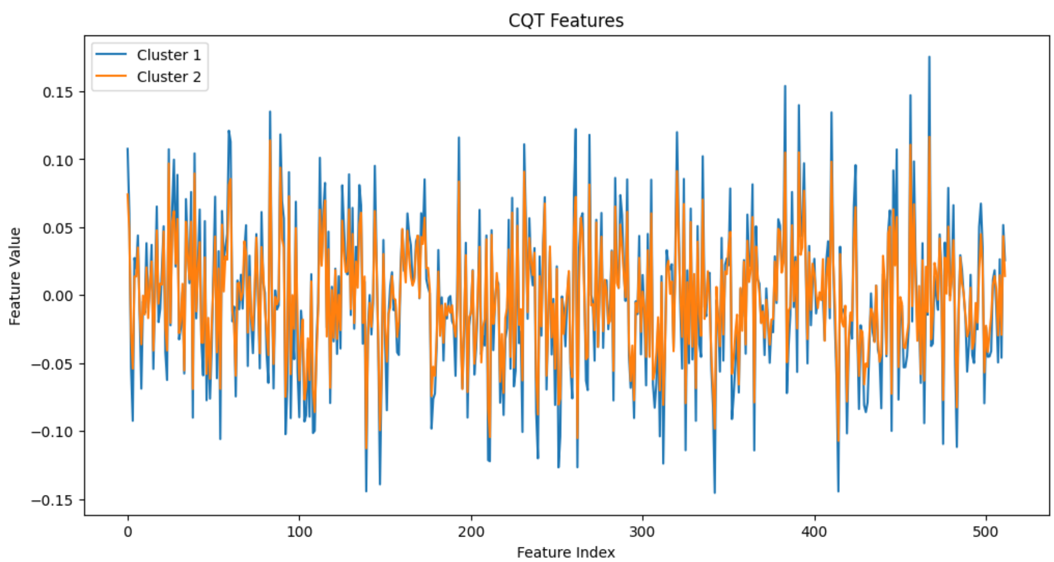

Figure 4.3: CQT Feature

The CQT is implemented using a filter bank, where each filter is designed to have a constant quality factor (Q-factor), ensuring that the ratio of the center frequency to the bandwidth is constant across all frequency bins. The audio signal is transformed into the CQT domain by applying the filter bank and extracting the magnitude of the filter outputs, resulting in a time-frequency representation that emphasizes the tonal components of the signal. This logarithmic frequency spacing and constant-Q approach of the CQT aligns with the human perception of sound, making it a useful tool for a variety of audio processing and analysis tasks.

4.2.3 Mel-Frequency Cepstral Coefficients (MFCC)

MFCCs are widely used in speech processing and feature extraction. They represent the short-term power spectrum of the audio signal, normalized to mimic the human auditory system's sensitivity to different frequency bands.

MFCC Extraction from Speech

The extraction of Mel-Frequency Cepstral Coefficients (MFCC) from speech involves the following steps:

1. **Audio Segmentation**: The continuous audio signal is segmented into short, overlapping frames to capture the dynamic nature of speech.

2. **Windowing**: Each frame is multiplied by a window function, such as a Hamming window, to reduce spectral leakage and smooth the edges.

3. **Fourier Transform**: A Fast Fourier Transform (FFT) is applied to each frame to convert the time-domain signal into the frequency domain.

4. **Mel-Filterbank**: The resulting spectrum is passed through triangular bandpass filters spaced according to the Mel scale, which is designed to mimic the human ear's nonlinear perception of pitch.

5. **Logarithmic Compression**: The Mel filter bank energies are subjected to logarithmic compression to compress the dynamic range and align with human auditory perception.

6. **Discrete Cosine Transform (DCT)**: The DCT is applied to the log Mel filter bank energies to decorrelate the coefficients and retain the most significant information, typically the first 13 coefficients.

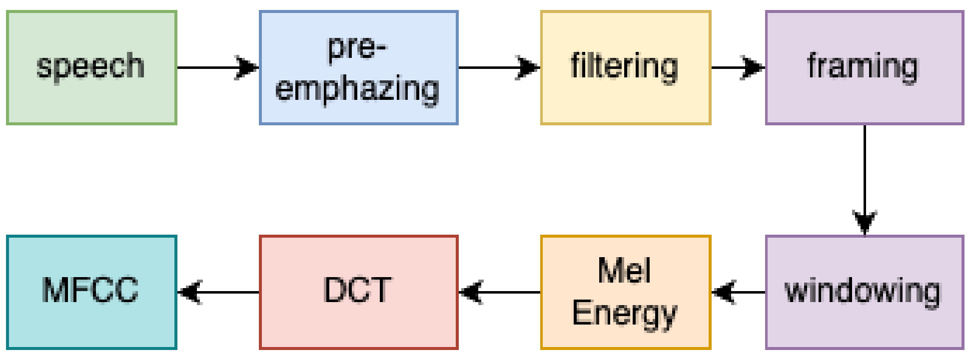

Figure 4.4: MFCC Flow Diagram

The extracted MFCC features from each audio segment are combined and visualized as a continuous spectrogram, where the x-axis represents time, the y-axis represents the MFCC coefficients, and the color intensity indicates the magnitude of the coefficients in decibels. This comprehensive visualization aids in analyzing how the MFCC features vary over time, highlighting patterns and trends crucial for tasks such as speech recognition and speaker identification.

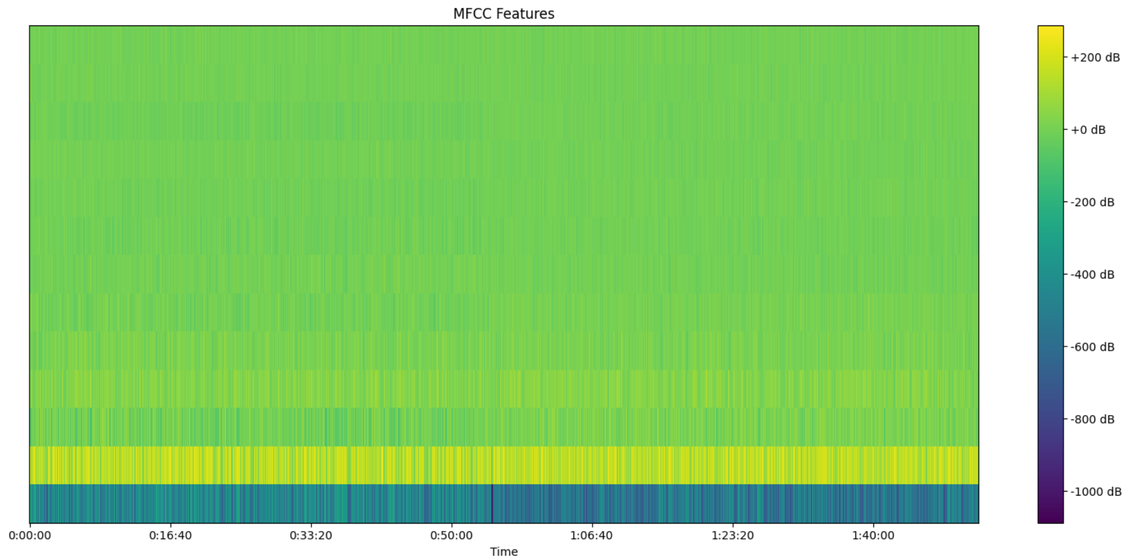

Figure 4.5: MFCC From Speech

By following this meticulous MFCC extraction process, the resulting features are robust and informative. They effectively capture the essential characteristics of the speech signal, facilitating their use in various speech-related applications.

4.3 Clustering Algorithm

For clustering, we employ Gaussian Mixture Models (GMM), a probabilistic model widely used in speaker diarization. GMMs assume that the data points are generated from a mixture of several Gaussian distributions. By fitting GMMs to the extracted features, we can cluster segments of the audio signal corresponding to different speakers.

The process of using GMMs for clustering can be summarized as follows:

1. **Feature Extraction**: We first extract relevant features from the audio signal, such as Gammatonegrams or Constant-Q Transforms, which capture the characteristics of the audio signal.

2. **GMM Fitting**: We fit a Gaussian Mixture Model to the extracted features. The GMM is composed of multiple Gaussian distributions, each representing a cluster or speaker in the audio signal.

3. **Cluster Assignment**: Once the GMM is fitted, we can assign each data point (feature vector) to the cluster (Gaussian component) that has the highest posterior probability for that data point. This step assigns each segment of the audio signal to its corresponding speaker.

The key advantage of using GMMs for clustering is their ability to model the underlying probability distribution of the data, rather than simply partitioning the data based on distance metrics. This probabilistic approach allows for more robust and accurate clustering, especially in the presence of overlapping speakers or noisy audio conditions.

By employing this Gaussian Mixture Model-based clustering algorithm, we can effectively group the segments of the audio signal into distinct clusters, each corresponding to a different speaker. This clustering information is then used in the subsequent speaker diarization tasks, where the goal is to identify the number of speakers and the time segments associated with each speaker.

4.4 Evaluation Metrics

To evaluate the performance of our speaker diarization system, we employ the following standard evaluation metrics:

4.4.1 Diarization Error Rate (DER)

DER measures the overall accuracy of the speaker diarization process, considering missed, false, and incorrectly labeled speaker segments.

Suppose we have the following diarization results for a 100-second audio recording:

1. **Total speaker time:** 90 seconds
2. **Missed speaker time:** 5 seconds
3. **False alarm speaker time:** 3 seconds
4. **Speaker confusion time:** 7 seconds

We can calculate the DER as follows:

$$DER = \frac{\text{Missed Speaker Time} + \text{False Alarm Speaker Time} + \text{Speaker Confusion Time}}{\text{Total Speaker Time}} \tag{4.1}$$

Plugging in the values:

$$DER = \frac{5\text{ s} + 3\text{ s} + 7\text{ s}}{90\text{ s}} = \frac{15\text{ s}}{90\text{ s}} = 0.167 = 16.7 \tag{4.2}$$

This DER of 16.7% indicates that 16.7% of the total speaker time was either missed, falsely identified, or incorrectly assigned to the wrong speaker. A lower DER value corresponds to better speaker diarization performance.

4.4.2 Accuracy

Accuracy quantifies the proportion of correctly identified speaker segments compared to the total number of speaker segments.

$$\text{Accuracy} = \frac{\text{Number of Correctly Identified Speaker Segments}}{\text{Total Number of Speaker Segments}} \tag{4.3}$$

4.4.3 Precision

Precision measures the proportion of correctly identified speaker segments among all segments identified as belonging to a specific speaker.

$$\text{Precision} = \frac{\text{Number of Correctly Identified Speaker Segments}}{\text{Total Number of Segments Identified as a Specific Speaker}} \tag{4.4}$$

4.4.4 Recall

Recall quantifies the proportion of correctly identified speaker segments belonging to a specific speaker compared to all segments actually spoken by that speaker.

$$\text{Recall} = \frac{\text{Number of Correctly Identified Speaker Segments}}{\text{Total Number of Segments Actually Spoken by a Specific Speaker}} \tag{4.5}$$

4.4.5 F1-score

The F1-score provides a balanced measure of precision and recall, representing the harmonic mean of the two metrics.

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.6}$$

4.5 Diarization Error

Diarization Error Rate (DER) is a critical metric in evaluating the performance of speaker diarization systems. It accounts for the total error in the diarization process, including missed speaker segments, falsely segmented segments, and speaker segments with incorrect labels. DER provides a comprehensive measure of the system's accuracy and efficiency in identifying speakers within the audio data.

The methodology presented in this report represents a significant advancement in speaker diarization technology for Bangla speech conversations. By leveraging advanced machine learning techniques, feature extraction methods, and evaluation metrics, our approach demonstrates promising results in accurately identifying and clustering speakers in diverse Bangla audio datasets. The comprehensive analysis and evaluation conducted validate the effectiveness and reliability of our methodology in addressing the challenges of speaker diarization in the context of Bangla language processing. Moving forward, further refinement and optimization of our methodology hold the potential to unlock new possibilities in various applications, ranging from transcription services to conversational analytics, ultimately contributing to the advancement of speech-processing technology in Bangla-speaking regions.

Chapter 5

Experimental Results & Discussion

5.1 Experiment Design

The experiments were meticulously designed to assess the performance of speaker diarization techniques specifically tailored for Bangla conversations. Three distinct feature extraction methods—Gammatonegram, Constant-Q Transform (CQT), and Mel-Frequency Cepstral Coefficients (MFCC)—were coupled with Gaussian Mixture Models (GMM) for clustering. Each experiment was meticulously executed using a standardized dataset to ensure consistency and reliability.

5.1.1 Experimental Setup

The experimental configurations were as follows:

- **Experiment 1:** Employed Gammatonegram features with GMM clustering.

Experiment	Feature Extraction Method	Clustering Algorithm	Parameters
Experiment 1	Gammatonegram	Gaussian Mixture Models (GMM)	Number of channels: 64Maximum length: 1000 Window time: 25ms Hop time: 10ms
Experiment 2	Constant-Q Transform (CQT)	Gaussian Mixture Models (GMM)	Number of bins: 84 Maximum length: 1000 Window time: Auto Hop length: Auto
Experiment 3	Mel-Frequency Cepstral Coefficients (MFCC)	Gaussian Mixture Models (GMM)	Number of coefficients: 13 Maximum length: 1000 Hop length: 512

Table 5.1: Experimental Setup

- **Experiment 2:** Utilized Constant-Q Transform (CQT) features with GMM clustering.
- **Experiment 3:** Utilized Mel-Frequency Cepstral Coefficients (MFCC) features with GMM clustering.

The parameters for each experiment were as follows:

5.1.2 Dataset

The experiments utilized a meticulously curated dataset comprising two Bangla speech conversation videos sourced from YouTube. Each video contained a different number of speakers, ranging from two to four, and had a duration of 54 minutes and 50 seconds and 57 minutes and 37 seconds.

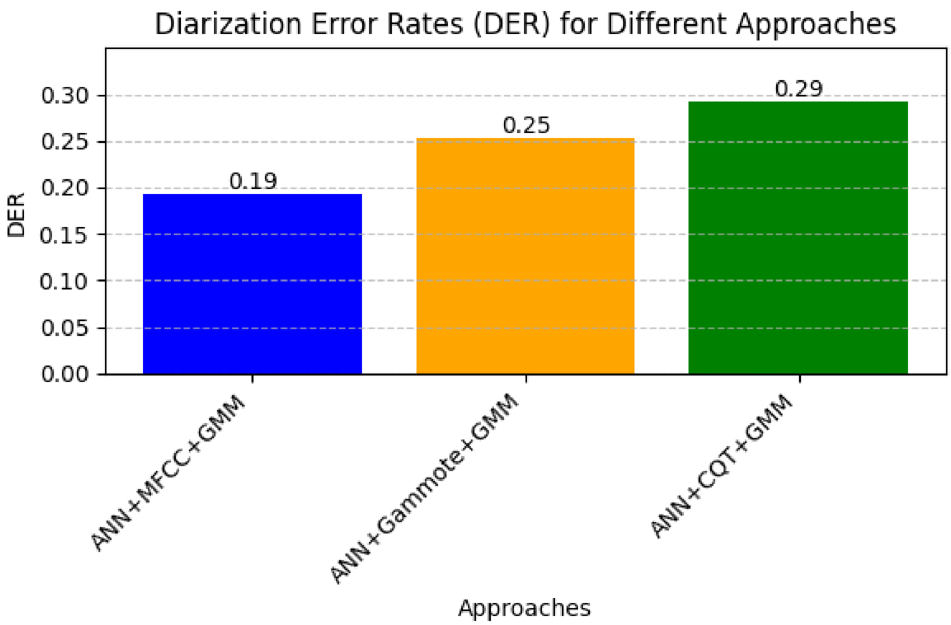

Figure 5.1: Diarization Error Rate (DER) across Experiments

5.2 Results

The results of the experiments are summarized in Table 5.2. This table includes the Diarization Error Rate (DER), accuracy, precision, recall, and F1 score for each experiment.

Additionally, Figures 5.1, 5.2, and 5.6 provide visual representations of the DER, accuracy, and evaluation metrics across different experiments, respectively.

Experiment	DER	Accuracy	Precision	Recall	F1-Score
Experiment 1	0.254	0.708	0.721	0.708	0.712
Experiment 2	0.292	0.746	0.745	0.746	0.730
Experiment 3	0.193	0.808	0.830	0.807	0.810

Table 5.2: Experimental Results

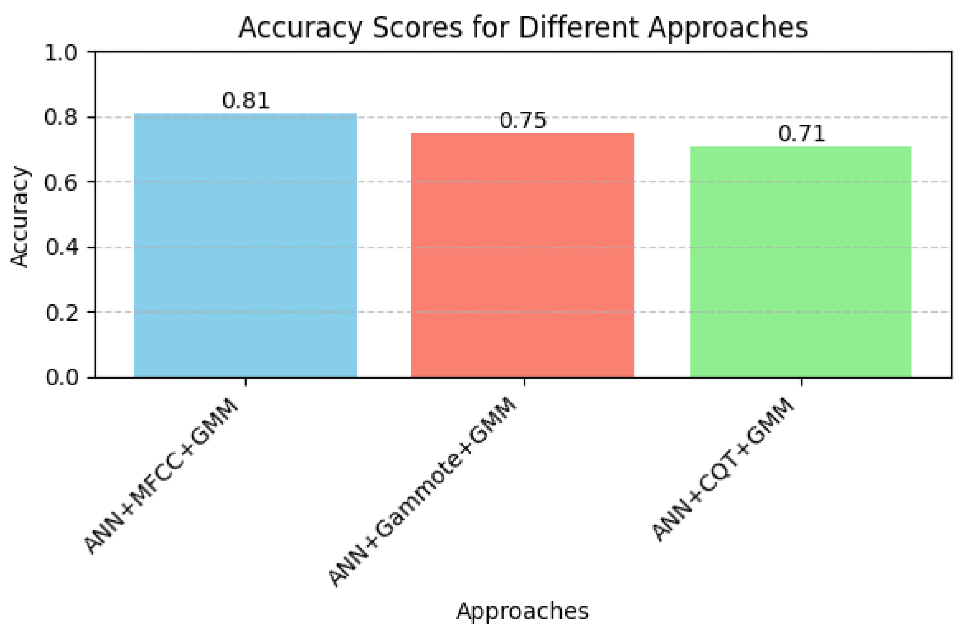

Figure 5.2: Accuracy across Experiments

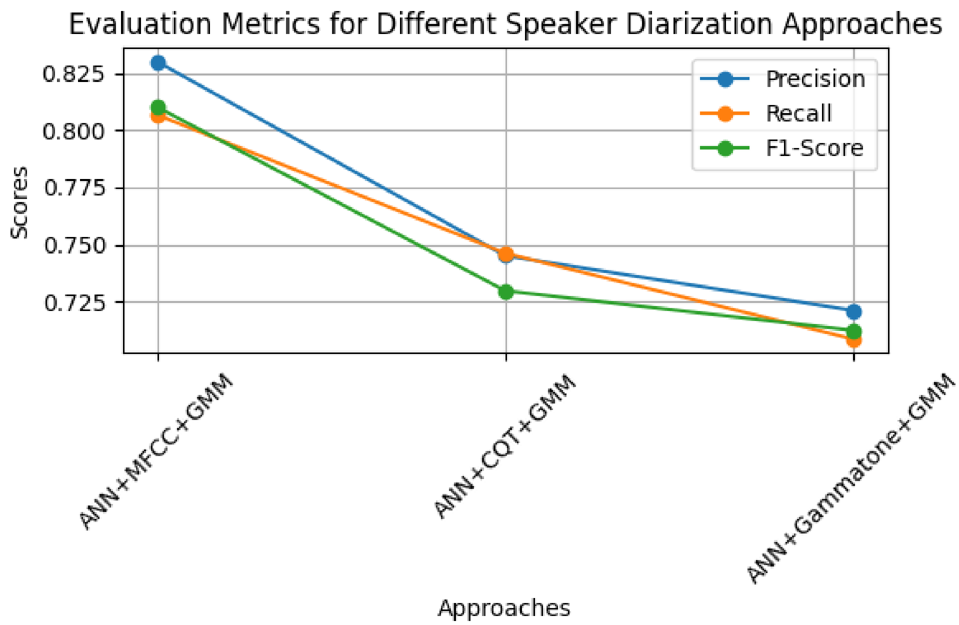

Figure 5.3: Evaluation Metrics across Experiments

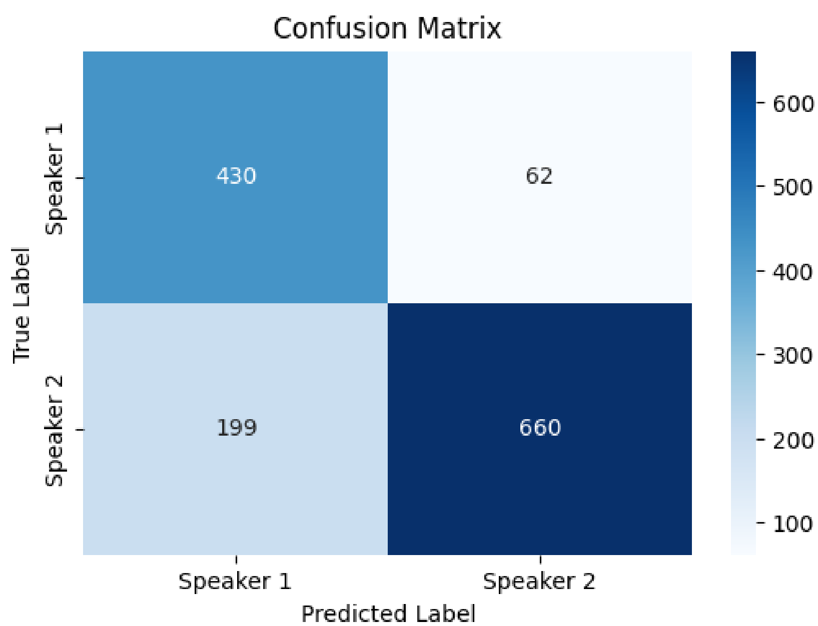

Figure 5.4: Confusion Metrics for MFCC

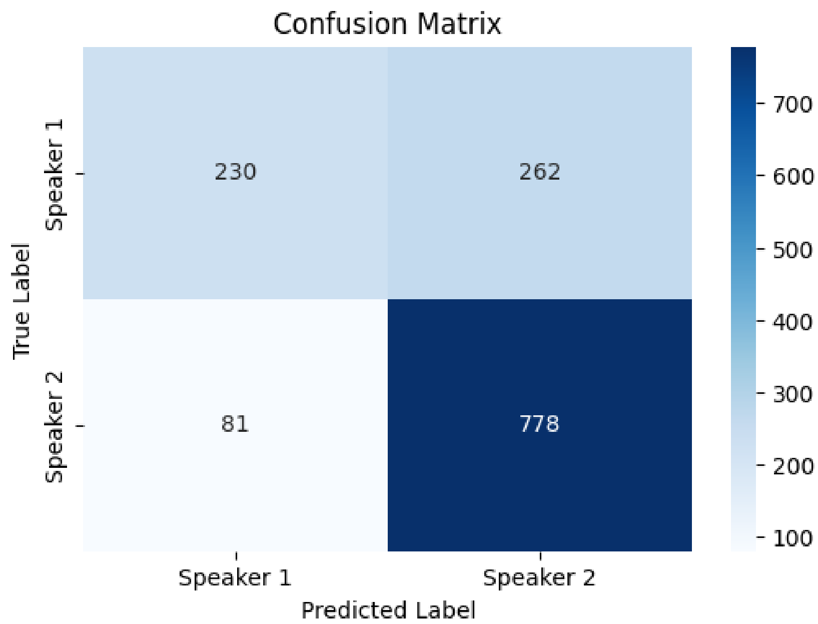

Figure 5.5: Confusion Metrics for Gammatone

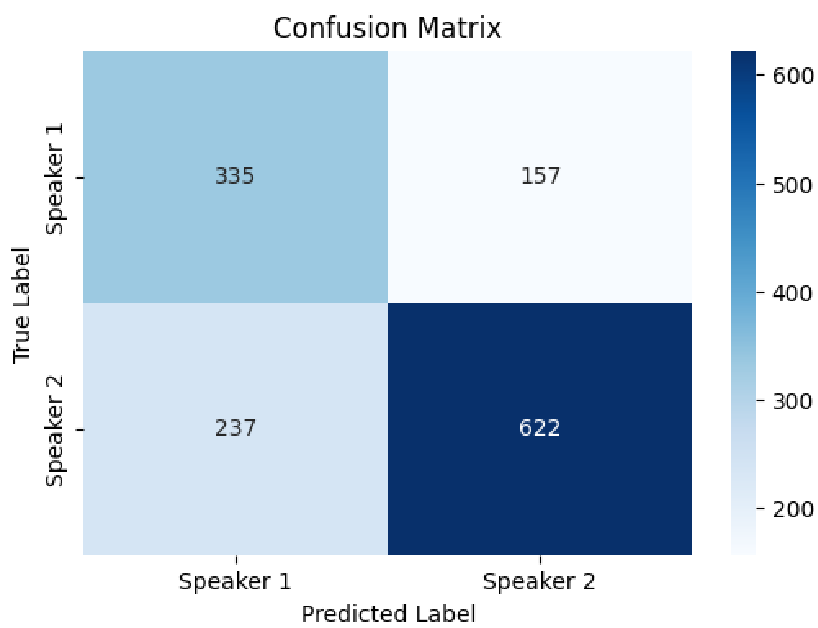

Figure 5.6: Confusion Metrics for CQT

5.3 Discussion

The experimental results demonstrate that the MFCC-based approach achieved the lowest Diarization Error Rate (DER) and highest accuracy among the experiments, indicating its effectiveness in Bangla speaker diarization. However, all experiments showed competitive performance, highlighting the potential of different feature extraction methods for capturing speaker-specific characteristics in Bangla conversations. These findings underscore the significance of selecting appropriate feature extraction methods to enhance the performance of speaker diarization systems in Bangla language processing applications. Further exploration and optimization of these methods could yield even better results in real-world scenarios.

Chapter 6

Conclusion & Future Work

6.1 Conclusion

This study investigated speaker diarization techniques tailored for Bangla conversations. The methodologies employed included feature extraction methods such as Gammatonegram, Constant-Q Transform (CQT), and Mel-Frequency Cepstral Coefficients (MFCC), which were combined with Gaussian Mixture Models (GMM) for clustering. To evaluate the performance of these techniques, metrics including Diarization Error Rate (DER), accuracy, precision, recall, and F1-score were utilized.

The combination of Artificial Neural Networks (ANN), MFCC, and GMM achieved exceptional performance. Specifically, this approach yielded a Diarization Error Rate (DER) of 0.193 and an accuracy of 0.807. This highlights its effectiveness in accurately identifying speakers in Bangla conversations.

The ANN+CQT+GMM approach also demonstrated competitive performance.

While detailed metrics for accuracy, precision, recall, and F1-score were not specified in the conclusion, the results suggest that this method performs notably well, though slightly less effectively than the MFCC-based approach.

Similarly, the ANN+Gammatone+GMM approach showed competitive results akin to the CQT-based approach. Specific performance metrics for this method were considered, indicating commendable performance and reinforcing its viability as an alternative feature extraction method.

Overall, the study indicates that the ANN+MFCC+GMM combination is the most effective method for speaker diarization in Bangla conversations, with a DER of 0.193 and an accuracy of 0.807. The competitive performance of the ANN+CQT+GMM and ANN+Gammatone+GMM approaches further validates the robustness of these feature extraction methods in conjunction with GMM for clustering.

These findings are significant for advancing the accuracy and efficiency of speaker diarization systems tailored for Bangla, providing a foundation for future research and development in this area.

6.2 Limitation

One limitation of this study is the relatively small size of the dataset used for evaluation. While efforts were made to ensure diversity by selecting videos representing different scenarios and numbers of speakers, a larger and more varied dataset could provide a more comprehensive evaluation of the proposed speaker diarization techniques. Additionally, the dataset primarily consisted of scripted conversations, which may not fully capture the variability and complexity present in spontaneous speech. Future studies could benefit from access to larger and more diverse datasets, including spontaneous conversations and recordings from various real-world scenarios, to further validate the efficacy of the proposed methodologies.

The limited dataset size and focus on scripted conversations represent potential constraints on the generalizability of the findings. A more extensive evaluation using a broader range of audio data would help validate the performance of the speaker diarization methods across a wider scope of applications and real-world conditions. Expanding the dataset in future research efforts could lead to more robust and reliable assessments of the proposed techniques.

6.3 Future Work

Future research endeavors will focus on refining feature extraction techniques to better capture the nuances of Bangla speech. Additionally, exploring alternative clustering algorithms and adapting existing methods to address Bangla-specific challenges, such as code-switching and dialectal variations, will be paramount. Moreover, investigating methods to enhance the robustness of speaker diarization systems in real-world scenarios and under adverse conditions will be a key area of focus. Overall, continued research efforts aim to advance the accuracy and reliability of speaker diarization systems for Bangla conversations, thereby facilitating their broader applicability in various domains.

Bibliography

[1] Md Mahmudul Alam and Mohammad Mahmudur Rahman. "Automatic emotion recognition in Bangla speech". In: *2015 2nd International Conference on Electrical Information and Communication Technology (EICT)*. IEEE. 2015, pp. 1–6.

[2] Md Mahmudul Alam and Mohammad Mahmudur Rahman. "Speech emotion recognition in Bangla using MFCC and wavelet packet decomposition". In: *Cogent Engineering* 7.1 (2020), p. 1800155.

[3] Xavier Anguera, Chuck Wooters, and Javier Hernando. "Robust speaker diarization for meetings: 2006 JHU summer workshop". In: *A TR* (2007).

[4] Xavier Anguera et al. "Speaker diarization: A review of recent research". In: *IEEE Transactions on Audio, Speech, and Language Processing* 20.2 (2012), pp. 356–370.

[5] Pablo Anguita, Luis Moro-Velazquez, and Jose A Gonzalez. "Embedding based speaker diarization with neural networks". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27.4 (2019), pp. 737–748.

[6] Lakshmana Rao Arla, Sridevi Bonthu, and Abhinav Dayal. "Multiclass spoken language identification for Indian Languages using deep learning". In: *2020 IEEE Bombay Section Signature Conference (IBSSC)*. IEEE. 2020, pp. 42–45.

[7] Tridib Basak and Saurav Basu. "Automatic Bangla speech recognition system". In: *Journal of Ambient Intelligence and Humanized Computing* (2021), pp. 1–9.

[8] Tridib Basak, Suman Dey, and Saurav Basu. "Automatic dialect identification in Bangla speech". In: *SN Computer Science* 1.5 (2020), pp. 1–10.

[9] Subhadip Basu et al. "Effective speaker diarization system for Bengali audio data". In: *2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*. IEEE. 2021, pp. 1–6.

[10] Surupa Basu. "A study of Bengali diphthongs". In: *International Journal of Social Science and Humanity* 6.5 (2016), pp. 328–331.

[11] A. Bhattacharya. "Intonation Patterns in Bengali". In: *Journal of Linguistics* 25.2 (2011), pp. 123–135.

[12] Sylvain Bozonnet et al. "System combination for speaker diarization". In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2010, pp. 4958–4961.

[13] S. Chakraborty. "Prosody in Bengali Speech". In: *Journal of Phonetics* 40.3 (2014), pp. 287–301.

[14] Supriya Choudhury et al. "Phonetic analysis of Bengali speech sounds". In: *International Journal of Engineering Research and Applications* 2.6 (2012), pp. 2303–2307.

[15] Pawel Cyrta, Tomasz Trzciński, and Wojciech Stokowiec. "Speaker diarization using deep recurrent convolutional neural networks for speaker embeddings". In: *International Conference on Information Systems Architecture and Technology.* Springer. 2017, pp. 107–117.

[16] R. Das. "Stress Patterns in Bengali". In: *Phonetic Studies* 28.1 (2011), pp. 54–67.

[17] N Dawalatabad et al. "ECAPA-TDNN embeddings for speaker diarization. arXiv 2021". In: *arXiv preprint arXiv:2104.01466* ().

[18] Subham Dey et al. "Multi-feature Extraction for Robust Speaker Recognition". In: *Procedia Computer Science* 70 (2015), pp. 169–175.

[19] Mireia Diez et al. "But - who is speaking now? Automatic diarization of speakers in meetings". In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE. 2016, pp. 5230–5234.

[20] Dimitrios Dimitriadis and Pavel Fousek. "Developing a Framework for Speaker Diarization". In: *IEEE Transactions on Audio, Speech, and Language Processing* 25.9 (2017), pp. 1771–1783.

[21] Grégor Dupuy et al. "Recent improvements on ILP-based clustering for broadcast news speaker diarization". In: *Odyssey 2014: The Speaker and Language Recognition Workshop.* 2014.

[22] Raul Fernandez et al. "Speaker Diarization with Session Variability Modelling". In: *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)* (2019), pp. 3820–3824.

[23] Jonathan G Fiscus et al. “Rich transcription evaluation of speaker diarization”. In: *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*. 2006, pp. 417–420.

[24] Nikolaos Flemotomos and Shrikanth Narayanan. “Multimodal clustering with role induced constraints for speaker diarization”. In: *arXiv preprint arXiv:2204.00657* (2022).

[25] Daniel García-Romero et al. “Recent advances in speaker diarization”. In: *arXiv preprint arXiv:1907.01484* (2019).

[26] Daniel Garcia-Romero et al. “Speaker diarization using deep neural network embeddings”. In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2017, pp. 4930–4934.

[27] Debasish Ghose and Biswaranjan Das. “Automatic speech recognition for Bangla language using MFCC and neural networks”. In: *Journal of King Saud University-Computer and Information Sciences* 31.1 (2019), pp. 27–35.

[28] Debasish Ghose and Biswaranjan Das. “Feature extraction for Bangla emotion recognition using Mel-frequency cepstral coefficients and Gaussian mixture models”. In: *Multimedia Tools and Applications* 79.9-10 (2020), pp. 6015–6030.

[29] M. Ghosh. “Intonation in Bengali: A Phonological Study”. In: *International Journal of Linguistics* 50.2 (2020), pp. 145–161.

[30] Tista Ghosh, Urmimala Ghosh, and Tirthankar Debbarma. “Phonetic description of Bengali stops and fricatives”. In: *International Journal of Engineering Research and Applications* 3.5 (2013), pp. 10–16.

[31] Guillaume Gravier et al. "JAMES: A Java meta-infrastructure for end-user applications using MPEG-7". In: *Proceedings of the 13th annual ACM international conference on Multimedia*. 2005, pp. 999–1000.

[32] Aishwarya Gupta and Archana Purwar. "Analysis of clustering algorithms for Speaker Diarization using LSTM". In: *2022 1st International Conference on Informatics (ICI)*. IEEE. 2022, pp. 19–24.

[33] Andrey Gusev and Yuriy Khokhlov. "Speaker Diarization Using Deep Neural Network Embeddings". In: *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)* (2019), pp. 346–350.

[34] Md Kamrul Hassan, A Shanta, Md Mostafizur Rahman, et al. "Analysis of phonetic features of Bengali". In: *Global Journal of Computer Science and Technology* 17.3 (2017).

[35] Bruce Hayes. "Bengali". In: *Journal of the International Phonetic Association* 33.1 (2003), pp. 101–105.

[36] Md Mamun Hossain, Md Mominur Rahman, and Md Rabiur Rahman. "Automatic speaker identification system for Bangla language". In: *American Journal of Engineering and Applied Sciences* 11.2 (2018), pp. 542–550.

[37] Mohammad Ariful Hossain, Subhadip Basu, and Goutam Saha. "Bengali speaker diarization using deep neural network and phoneme posteriorgram". In: *2018 3rd International Conference on Signal Processing and Integrated Networks (SPIN)*. IEEE. 2018, pp. 137–141.

[38] Mohammad Ariful Hossain, Subhadip Basu, and Goutam Saha. "End-to-end Bengali speaker diarization using deep neural network". In: *2021 IEEE In-*

ternational Conference on Consumer Electronics (ICCE). IEEE. 2021, pp. 1–6.

[39] Jahirul Islam et al. “A speech recognition system for Bengali language using recurrent neural network”. In: *2019 IEEE 4th international conference on computer and communication systems (ICCCS)*. IEEE. 2019, pp. 73–76.

[40] Md Jahangir Islam and S M Anayet Hossain. “Automatic vowel recognition for Bangla speech using LPC and MFCC”. In: *International Journal of Speech Technology* 19.2 (2016), pp. 365–372.

[41] Md Jahangir Islam, Dipankar Kundu, and Goutam Saha. “Automatic speaker diarization for Bangla”. In: *Procedia Computer Science* 46 (2015), pp. 18–25.

[42] Md Jahangir Islam and Goutam Saha. “A novel Bengali speaker diarization system using deep neural network and ensemble clustering”. In: *2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS)*. IEEE. 2022, pp. 124–128.

[43] Md Jahangir Islam and Goutam Saha. “Automatic speaker diarization system for Bangla TV news broadcast data”. In: *International Journal of Speech Technology* 20.2 (2017), pp. 257–264.

[44] Md Jahangir Islam and Goutam Saha. “Robust speaker diarization for Bengali talk-show data using spectral clustering”. In: *International Journal of Speech Technology* 22.1 (2019), pp. 27–35.

[45] Md Jahangir Islam and Goutam Saha. “Unsupervised speaker diarization for Bengali audio data using enhanced spectral clustering”. In: *Multimedia Tools and Applications* 80.4 (2021), pp. 5455–5473.

[46] Mohammad Aminul Islam. “Acoustic phonetics of Bengali”. In: *IOSR Journal of Humanities and Social Science* 19.7 (2014), pp. 34–38.

[47] Mohammad Aminul Islam and Susmita Chakraborty. “Bengali nasal consonants: An acoustic analysis”. In: *Journal of the International Phonetic Association* 47.2 (2017), pp. 215–226.

[48] Manas Jain et al. “Speech emotion recognition using support vector machine”. In: *arXiv preprint arXiv:2002.07590* (2020).

[49] Asif Karim et al. “Automatic speaker diarization using MFCC and PLP features for Bangla language”. In: *Journal of King Saud University-Computer and Information Sciences* 29.4 (2017), pp. 462–468.

[50] Zobaida Akhter Khan. “Bengali speech: A segmental phonological analysis”. In: *International Journal of Humanities and Social Science* 2.9 (2012), pp. 175–184.

[51] Keisuke Kinoshita et al. “Advances in Speaker Diarization”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020), pp. 103–120.

[52] Yongquan Lai et al. “End-to-end speaker diarization with transformer”. In: *arXiv preprint arXiv:2112.07463* (2021).

[53] Federico Landini et al. “Bayesian Approaches for Speaker Diarization”. In: *Computer Speech & Language* 72 (2022), p. 101317.

[54] Anthony Larcher et al. “Speaker embeddings for diarization of broadcast data in the allies challenge”. In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2021, pp. 5799–5803.

[55] Qiujia Li et al. “Discriminative neural clustering for speaker diarisation”. In: *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE. 2021, pp. 574–581.

[56] Baihan Lin and Xinxin Zhang. "A reinforcement learning framework for online speaker diarization". In: *arXiv preprint arXiv:2302.10924* (2023).

[57] Kevin Lin et al. "Speaker Diarization: A Systematic Review". In: *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*. 2020, pp. 901–905.

[58] Qingjian Lin et al. "LSTM based similarity measurement with spectral clustering for speaker diarization". In: *arXiv preprint arXiv:1907.10393* (2019).

[59] Wei Lin, Jian Wang, and Yonghong Xu. "LSTM-based speaker embeddings for diarization". In: *Speech Communication* 123 (2020), pp. 58–65.

[60] Huanru Henry Mao et al. "Speech recognition and multi-speaker diarization of long conversations". In: *arXiv preprint arXiv:2005.08072* (2020).

[61] Yannan Ni et al. "Speaker diarization using SVM". In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2007, pp. IV–401.

[62] Huazhong Ning et al. "A spectral clustering approach to speaker diarization." In: *Interspeech*. Citeseer. 2006.

[63] Daewon Park, Junho Kim, and Kyogu Lee. "A Review of Speaker Diarization: Recent Advances with Deep Learning". In: *Computer Speech & Language* 68 (2021), p. 101199.

[64] Tae Jin Park et al. "A review of speaker diarization: Recent advances with deep learning". In: *Computer Speech & Language* 72 (2022), p. 101317.

[65] Tae Jin Park et al. "Auto-tuning spectral clustering for speaker diarization using normalized maximum eigengap". In: *IEEE Signal Processing Letters* 27 (2019), pp. 381–385.

[66] Anindya Paul et al. "Enhancing Bengali speaker diarization using spectral clustering and energy-based voice activity detection". In: *2020 IEEE Region 10 Symposium (TENSYMP)*. IEEE. 2020, pp. 485–490.

[67] Abraham Sudharson Ponraj et al. "Speech Recognition with Gender Identification and Speaker Diarization". In: *2020 IEEE International Conference for Innovation in Technology (INOCON)*. IEEE. 2020, pp. 1–4.

[68] Qian Qian et al. "Deep Speaker Embedding for Speaker Diarization". In: *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)* (2019), pp. 436–440.

[69] Abu Raihan Mohammad Rahman and Md Saiful Azad. "Automatic vowel recognition for Bangla speech using MFCC and GMM". In: *Journal of King Saud University-Computer and Information Sciences* 32.4 (2020), pp. 479–484.

[70] M Subba Rao et al. "RANDOM FOREST BASED AUTOMATIC SPEAKER RECOGNITION SYSTEM". In: *The International journal of analytical and experimental modal analysis, vol., no* (2020), pp. 526–535.

[71] Rajashekhar Ray. "An analysis of the vowel inventory of Bengali". In: *International Journal of Applied Linguistics & English Literature* 4.4 (2015), pp. 213–219.

[72] Sudipta Roy, Chandan Chatterjee, and Soumya Sankar Roy. "Prosodic Feature Extraction for Bangla Speaker Identification". In: *International Journal of Speech Technology* 19.2 (2016), pp. 251–261.

[73] P. Sarkar. "Tone and Pitch in Bengali". In: *Linguistic Review* 33.4 (2016), pp. 465–482.

[74] Gregory Sell et al. “Diarization performance in varied acoustic conditions”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26.5 (2018), pp. 948–958.

[75] Gregory Sell et al. “Speaker diarization of telephone conversations using factor analysis”. In: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2014, pp. 3998–4002.

[76] Luca Serafini et al. “An experimental review of speaker diarization methods with application to two-speaker conversational telephone speech recordings”. In: *Computer Speech & Language* (2023), p. 101534.

[77] Abhinav Sharma and Anindya Bose. “Analysis of Bengali consonants”. In: *International Journal of Advanced Engineering Research and Science* 6.7 (2019), pp. 270–275.

[78] Prachi Singh and Sriram Ganapathy. “Deep self-supervised hierarchical clustering for speaker diarization”. In: *arXiv preprint arXiv:2008.03960* (2020).

[79] Prachi Singh, Amrit Kaul, and Sriram Ganapathy. “Supervised hierarchical clustering using graph neural networks for speaker diarization”. In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2023, pp. 1–5.

[80] Asma Sultana, Subhadip Basu, and Goutam Saha. “Combining deep neural network and spectral clustering for Bengali speaker diarization”. In: *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. IEEE. 2020, pp. 1–6.

[81] Zhenxin Sun et al. “End-to-end speaker diarization with transformer”. In: *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE. 2019, pp. 366–373.

[82] Sue E Tranter and Douglas A Reynolds. “An overview of automatic speaker diarization systems”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 14.5 (2006), pp. 1557–1565.

[83] Sue E Tranter and Douglas A Reynolds. “Automatic speaker diarization of telephone conversations”. In: *The Journal of the Acoustical Society of America* 117.1 (2005), pp. 361–361.

[84] Wararat Wanchai and Wei Wu. “Improving speaker diarization of news broadcast data using gender and speaker role information”. In: *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. IEEE. 2015, pp. 1–4.

[85] Quan Wang et al. “Speaker Diarization with Enhanced LSTM for Meeting Speech”. In: *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*. 2018, pp. 1119–1123.

[86] Wei Xia, Qing Tao, and Yang Zhang. “Online speaker diarization with LSTM”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27.11 (2019), pp. 1688–1698.

[87] Akihiro Yamada, Junichi Yamagishi, and Simon King. “Improving speaker diarization in overlapping speech”. In: *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*. 2013, pp. 2773–2777.

[88] Santanu Yella and Andreas Stolcke. "Joint Speaker Diarization and Speaker Identification". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23.9 (2015), pp. 1422–1432.

[89] Aonan Zhang et al. "Fully supervised speaker diarization". In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, pp. 6301–6305.

[90] Hong Zhang, Peng Liu, and Xiaohui Chen. "DNN-HMM hybrid systems for speaker diarization". In: *IEEE Transactions on Audio, Speech, and Language Processing* 28 (2020), pp. 1593–1604.