Improvising Diarization for Multilingual Communication

Thesis submitted in partial fulfillment of the requirements for the degree of

Master of Technology in Computer Technology

Submitted By

Satyaki Das

Registration No: **160079** of **2021-2022**

Exam Roll No: **M6TCT24006B**

Class Roll No: **002110504008**

Session: **2021-2024**

Under the Guidance and Supervision of

Dr. Sudip Kumar Naskar

Department of Computer Science and Engineering

Jadavpur University

188, Raja S.C. Mallick Rd,

Kolkata - 700032, West Bengal, India

Department of Computer Science and Engineering Faculty of Engineering and Technology
Jadavpur University, Kolkata - 700032

DECLARATION OF ORIGINALITY AND COMPLIANCE OF ACADEMIC ETHICS

I, hereby declare that the thesis entitled **"Improvising Diarization for Multilingual Communication"** contains a literature survey and original research work by the undersigned candidate, as a part of his degree of **Master of Technology in Computer Technology**, **Jadavpur University**. All the information has been obtained and presented in accordance with academic rules and ethical conduct.

I also declare that, as required by these rules and conduct, I have fully cited and referenced all materials and results that are not original to this work.

Name: Satyaki das

Examination Roll No: M6TCT24006B

Registration No: 160079 of 2021-2022

Class Roll No: 002110504008

Session: 2021-2024

Thesis Title: Improvising Diarization for Multilingual Communication

Signature of the Candidate:

Department of Computer Science and Engineering Faculty of Engineering and Technology
Jadavpur University, Kolkata - 700032

CERTIFICATE OF RECOMMENDATION

This is to certify that the work embodied in this thesis entitled **"Improvising Diarization for Multilingual Communication"** has been satisfactorily completed by **Satyaki Das** (Registration Number **160079** Of **2021-2022**; Class Roll No. **002110504008**; Examination Roll No. **M6TCT24006B**). It is a bonafide piece of work carried out under my supervision and guidance at Jadavpur University, Kolkata, for partial fulfillment of the requirements for the awarding of the Master of Technology in Computer Technology degree of the Department of Computer Science and Engineering, Faculty of Engineering and Technology, Jadavpur University, during the academic year 2021 - 2024.

_____

Dr. Sudip Kumar Naskar,
Department of Computer Science and Engineering,
Jadavpur University.
(**Supervisor**)

Forwarded By:

_____

Prof. Dr. Nirmalya Chowdhury,
HOD, Department of Computer Science and Engineering,
Jadavpur University.

_____

Prof. Rajib Bandyopadhyay
DEAN, Faculty of Engineering and Technology,
Jadavpur University.

Department of Computer Science and Engineering Faculty of Engineering and Technology
Jadavpur University, Kolkata - 700032

CERTIFICATE OF APPROVAL

This is to certify that the thesis entitled **"Improvising Diarization for Multilingual Communication"** is a bonafide record of work carried out by **Satyaki Das** (Registration Number **160079** Of **2021-2022**; Class Roll No. **002110504008**; Examination Roll No. **M6TCT24006B**) in partial fulfillment of the requirements for the award of the degree of Master of Technology in Computer Technology in the Department of Computer Science and Engineering, from September 2021 to Nov 2024. It is understood that by this approval, the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein, but approve the thesis only for which it has been submitted.

_____

Signature of Examiner 1

Date:

_____

Signature of Examiner 2

Date:

Acknowledgments

I would like to express my sincere gratitude to my guide, **Dr. Sudip Kumar Naskar**, for his continuous motivation, guidance, and support throughout the research journey. His expertise and dedication have been instrumental in completing my thesis.

_____

Satyaki das

Examination Roll No: M6TCT24006B

Registration No: 160079 of 2021-2022

Department of Computer Science and Engineering,

Jadavpur University

Abstract

Nowadays, many people use virtual meeting platforms like Google Meet and Microsoft Teams to communicate. In these meetings, it's common for participants to switch between different languages, such as Hindi, English, and Bengali while communicating. However, the current automatic speech recognition (ASR) systems and summarization tools need to work better with these multilingual conversations. They often miss important words or produce inaccurate transcribes because they are not designed to handle multiple languages simultaneously.

The main problem is that most ASR systems are built to recognize only one language at a time or struggle when people switch between languages during a conversation. This makes it hard to get accurate transcripts from these meetings, leading to misunderstandings or loss of important information.

This thesis focuses on improving automatic speech recognition (ASR) and transcription systems in virtual meetings for Hindi, English, and Bengali languages. The goal is to create a system that can better understand and transcribe conversations where people use multiple languages, making virtual meetings more effective and accessible for everyone.

Contents

Abstract i

Table of Contents ii

List of Figures v

List of Tables vi

1 Introduction 1

1.1 Automatic Speech Recognition (ASR)..... 1

1.2 Audio Diarization……..... 2

1.3 Senetence Transformation…….. 3

2 Literature Survey 4

2.1 Approaches to ASR…… 4

2.1.1 Acoustic-Phonetic Approach……..... 4

2.1.2 Pattern Recognition Approach…….. 4

2.1.3 Knowledge-Based Approach…… 5

2.1.4 Connectionist Approach…… 5

2.1.5 Support Vector Machines…….. 6

2.2 Approaches to Diarization……. 6

2.2.1 Preprocessing Approaches……..…….. 8

2.2.2 Feature Extraction and Normalization Approaches…….. 9

2.2.3 Speaker-Based Segmentation and Change Point Detection Approaches……..……..... 10

2.2.4 Speaker Segmentation Approaches……..……. 12

2.3 Overview of Commercially Available ASR and Speaker Diarization Services……..…… 12

2.3.1. Google Speech Recognition (speech_recognition module)..... 12

2.3.2. Amazon Transcribe……..…… 12

2.3.3. Rev.ai…… 13

2.3.4. IBM Watson Speech-to-Text…… 13

2.3.5. pyannote-audio……. 13

3 Datasets 14

3.1 Casual Political Conversation Dataset..... 14

3.2 Casual Educational Conversation Dataset..... 15
3.3 Casual Sports Conversation Dataset..... 17
3.4 Casual Weather Conversation Dataset..... 18
3.5 Casual Innovation Dataset……. 19
4 Experiments and Results 4 22
4.1 Diarization Error Rate…… 22
4.1.1 DER % on Casual Political Conversation Dataset…… 24
4.1.2 DER % on Casual Educational Conversation Dataset…….. 25
4.1.3 DER % on Casual Sports Conversation Dataset……... 27
4.1.4 DER % on Casual Weather Conversation Dataset…… 28
4.1.5 DER % on Casual Innovation Conversation Dataset…… 30
4.2 Word Error Rate (WER)..... 31
4.3 Result Analysis…….. 33
5 Conclusions 5 35
5.1 Conclusion of the Present Work……. 35
5.1.1. Impact of Audio Length on Processing Time…….. 35
5.1.2. Gender Variability in Diarization Accuracy…… 35
5.1.3. False Alarm Impact on Diarization Error Rate (DER)..... 36
5.1.4. Miss Duration and Its Effect on DER…….. 36
5.1.5. Speaker Error and Diarization Error Rate (DER)..... 36
5.1.6. ASR and Sentence Transformer Model Selection……... 36
5.1.7. Impact of Miss Duration on Word Error Rate (WER)..... 36
5.1.8. False Alarm and Speaker Error Influence on Speaker-wise WER…….. 36
5.1.9. Effect of Noise on Performance…….. 36
5.1.10. Audio Overlapping and Performance…… 37
5.1.11. Multi-Channel Audio and Channel Quality…….. 37
5.2 Scope of the Future Work…….. 37
5.2.1. Advanced Diarization Models with Improved Speaker Discrimination…….. 37
5.2.2. Real-Time Transcription and Diarization Solutions…… 38
5.2.3. Enhancement of Overlapping Speech Handling…….. 38
5.2.4. Noise Robustness and Environmental Adaptation…….. 38
5.2.5. Multi-Lingual and Code-Switching Capabilities……. 38

5.2.6. Enhanced ASR Models with Contextual and Semantic Understanding…… 38
5.2.7. Exploring Multi-Channel and Spatial Audio Processing……. 39
5.2.8. Improvement of DER Metrics and Evaluation Methods……. 39
5.2.9. Personalization and Adaptation for Speaker-Specific Characteristics…….. 39
5.2.10. Ethical Considerations and Privacy-Preserving Transcription…… 39
5.2.11. User Feedback Integration for Continuous Model Improvement…….. 40
6 Bibliography 6 41

List of Figures

1.1 Key aspects of the Diarization process through visualization……...02

1.2 A general speaker diarization framework……..07

1.3 Structure of ICSI-SRI HMM-based speech/non-speech detector……..11

3.1 Snapshot of Casual Political Conversation Dataset……...15

3.2 Snapshot of Casual Educational Conversation Dataset……..16

3.3 Snapshot of Casual Political Sports Dataset……..…18

3.4 Snapshot of Casual Weather Dataset…….....19

3.5 Snapshot of Casual Innovation Dataset…….21

List of Tables

Table 3.1: Casual Political Conversation Dataset Summary…….....14

Table 3.2: Casual Educational Conversation Dataset Summary……16

Table 3.3: Casual Political Sports Dataset Summary…….17

Table 3.4: Casual Weather Coversation Dataset Summary……18

Table 3.5: Casual Innovation Conversation Dataset Summary……..19

Table 4.1: Ground Truth of Casual Political Conversation Dataset…….....24

Table 4.2: False Alarm of Casual Political Conversation Dataset……..24

Table 4.3: Miss of Casual Political Conversation Dataset…….....25

Table 4.4: Speaker Error of Casual Political Conversation Dataset …….….25

Table 4.5: Ground Truth of Casual Educational Conversation Dataset…….…..…25

Table 4.6: False Alarm of Casual Educational Conversation Dataset…….……...26

Table 4.7: Miss of Casual Educational Conversation Dataset……....26

Table 4.8: Speaker Error of Casual Educational Conversation Dataset…..……...27

Table 4.9: Ground Truth of Casual Sports Conversation Dataset…….…..27

Table 4.10: False Alarm of Casual Sports Conversation Dataset…….…..27

Table 4.11: Miss of Casual Sports Conversation Dataset……...28

Table 4.12: Speaker Error of Casual Sports Conversation Dataset……28

Table 4.13: Ground Truth of Casual Sports Conversation Dataset…….28

Table 4.14: False Alarm of Casual Sports Conversation Dataset……..28

Table 4.15: Miss of Casual Sports Conversation Dataset……..29

Table 4.16: Speaker Error of Casual Sports Conversation Dataset……29

Table 4.17: Ground Truth Casual Sports Innovation Dataset…….30

Table 4.18: False Alarm of Casual Sports Innovation Dataset……...30

Table 4.19: Miss of Casual Sports Innovation Dataset……...31

Table 4.20: Speaker Error of Casual Sports Innovation Dataset…….31

Table 4.21: Word Error Rate Calculated on Casual Political Conversation Dataset……...31

Table 4.22: Word Error Rate Calculated on Casual Educational Conversation Dataset……32

Table 4.23: Word Error Rate Calculated on Casual Sports Conversation Dataset…….32

Table 4.24: Word Error Rate Calculated on Casual Weather Innovation Dataset……..32

Table 4.25: Word Error Rate Calculated on Casual Innovation Conversation Dataset……..32

Table 4.26 : Findings on 5 Multilingual Dataset……...….32

Chapter 1
Introduction

In the evolving field of Natural Language Processing (NLP), speech-related tasks are gaining more attention, driven by the increasing use of voice-based systems like virtual assistants, automated transcription services, customer service applications, etc. **Speaker Diarization** and **Automatic Speech Recognition (ASR)** are two critical technologies in this space. Together, these techniques enable the transformation of raw audio into structured, usable data, advancing the capabilities of NLP in processing, understanding, and generating human-like interactions.

Automatic Speech Recognition (ASR) and Speaker Diarization have witnessed remarkable advancements in recent years, driven by breakthroughs in machine learning and deep learning techniques. These advancements have led to significant improvements in various speech-processing tasks such as transcriptions, speaker identification, and conversational analysis. One of the key benefits of these advancements is the ability to handle more complex acoustic patterns and speaker variations, enabling more accurate and contextually relevant speech-to-text results, even in multi-speaker environments. Additionally, ASR and diarization models have become more scalable and efficient, allowing for faster processing of large volumes of audio data, and facilitating real-time applications such as voice assistants and meeting transcriptions.

1.1 Automatic Speech Recognition (ASR)

Automatic Speech Recognition (ASR) is a technology that uses Machine Learning (ML) and Artificial Intelligence (AI) to turn spoken words into written text. This technology has grown a lot in the last ten years and is now used in many everyday apps. For example, YouTube and Instagram use ASR for real-time captions on videos, Spotify uses it to create text versions of podcasts, and Zoom, and Google Meet use it to transcribe meetings.

ASR works by listening to **what people say**, breaking down the sounds, and using smart algorithms to determine the words. It then converts those words into text that we can read. This technology is powerful because it can handle different accents and background noise, making it easier for everyone to understand and share spoken content.

Not only has the accuracy of ASR greatly improved, but it’s also become much more widely available, making it easier for everyone to use.

1.2 Audio Diarization

Diarization is the process of segmenting an audio recording into distinct sections, each associated with a different speaker. The primary goal of diarization is to determine "who spoke when" in a recording that features multiple speakers. This process is essential in various applications, such as transcribing meetings, analyzing conversations, and processing broadcast media.

Key Aspects of Diarization

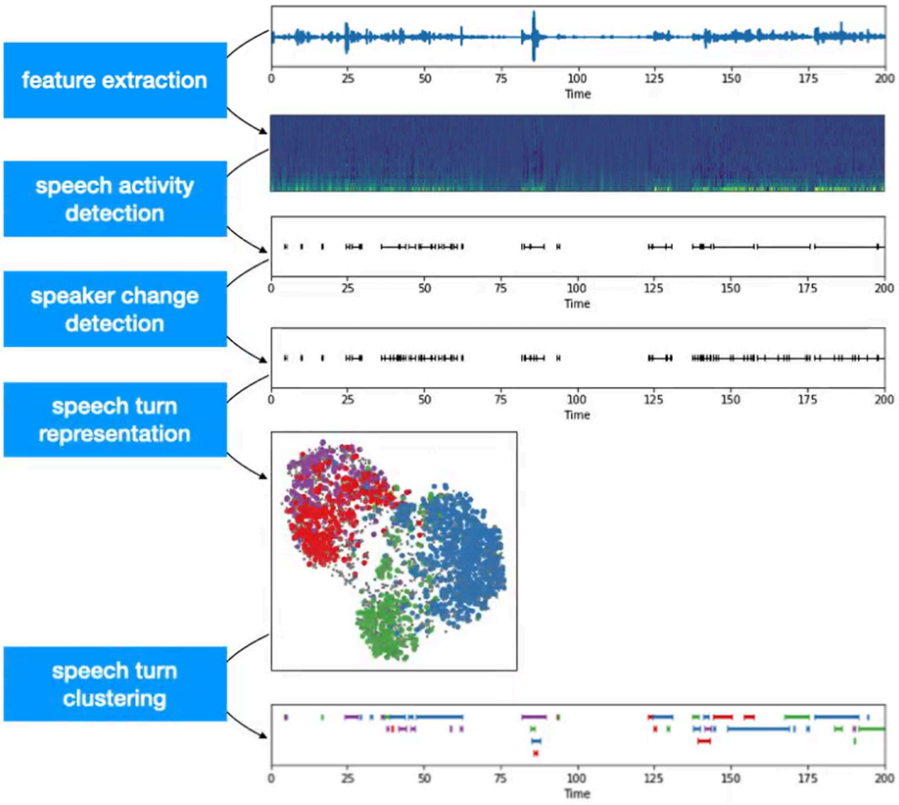


Figure 1.1 : Key aspects of the Diarization process through visualization

Speaker diarization is the process of separating and labeling different speakers in an audio recording. It starts by turning the raw audio into feature data that captures each speaker's unique voice characteristics, often using methods like Mel-Frequency Cepstral Coefficients (MFCCs). Then, Speech Activity Detection (SAD) identifies which parts of the audio contain speech and which parts are silent or have background noise. The next step is to find where one speaker stops and another begins, called Speaker Change Detection. Each section of speech is then turned into a detailed representation using tools like x-vectors, which hold information about the speaker's voice. Finally, the system groups similar speech segments together, so all parts spoken by the same person are put in the same group. This process is useful in many areas, like creating

accurate transcriptions, organizing meeting discussions, managing broadcast media, and analyzing call center conversations.

1.3 Sentence Transformation

In natural language processing (NLP), understanding the nuances of language, like semantic similarity between sentences, is a major challenge. With the rapid growth of textual data, advanced models have become essential. This research examines sentence transformation [4] using state-of-the-art NLP models to analyze semantic similarity, which measures how closely related two texts are based on meaning rather than just word matching. Traditional methods, such as word overlap, often miss deeper language patterns, but recent advancements like transformer-based architectures have greatly improved semantic understanding. We use the SentenceTransformer model, specifically all-MiniLM-L6-v2, [8] to create sentence embeddings that capture the essence of entire sentences. These embeddings allow for effective similarity comparisons using metrics like cosine similarity. Our study evaluates the semantic closeness of the sentence "It seems to be heard is heard" with sample texts, demonstrating how modern models accurately capture subtle linguistic nuances for robust semantic analysis.[17]

Chapter 2
Literature Survey

The development of speech recognition systems began with Alexander Graham Bell's work on converting sound waves into electrical impulses. One of the first systems, created by Davis et al., recognized spoken digits through an electronic circuit that analyzed formants and matched patterns using correlation coefficients, adapting to variations in speech. Over the last fifty years, speech recognition technology has advanced, supporting multiple languages and enabling efficient speech-to-text conversion. Despite the existence of over 7,300 languages, only a portion have well-developed systems, with languages like English, Russian, Chinese, and Hindi among those receiving significant research and development.

2.1 Approaches to ASR

2.1.1 Acoustic-Phonetic Approach

The acoustic-phonetic approach [5][10] to speech recognition, developed by Hemdal and Hughes, focuses on identifying and labeling distinct speech sounds using acoustic properties. It assumes a fixed set of phonetic units defined by features like nasality, voicing, formant locations, and frequency ratios. The method involves steps such as spectral analysis, feature detection, segmentation, labeling, and applying linguistic constraints for word recognition. Despite its structured design, this approach has not been successful for commercial use. Linguistic rules help the system access the lexicon, improving word interpretation and understanding.

2.1.2 Pattern Recognition Approach

The pattern recognition approach, introduced by Itakura in 1975 and expanded by Rabiner and Juang, has dominated speech recognition for the past six decades. It involves two main steps: pattern training and pattern comparison. This method stands out due to its use of a mathematical framework to create consistent speech pattern representations, enabling accurate comparisons. These representations can be in the form of speech templates or statistical models and are applied to recognize sounds, words, or phrases. During the comparison stage, the system matches new speech input with trained patterns, allowing precise classification. The approach's strength lies in its structured, mathematically rigorous analysis of speech data.

2.1.2.1 Template-Based Approach

The template-based approach stores a set of prototypical speech patterns as reference templates, creating a dictionary for recognizing spoken words. An unknown utterance is matched against

these templates to find the best match, thus determining the word category. Each word has a dedicated reference template, which minimizes errors from smaller acoustic variations but becomes inefficient as vocabulary size grows.

Watcher et al. addressed limitations of the Hidden Markov Model (HMM) [6], such as losing time dependencies and over-generalization, by applying a template-based method with Dynamic Time Warping (DTW) for continuous speech recognition. DTW does not require extensive modeling or training but faces challenges from a large search space, and its performance is often compared to HMM-based systems [6].

2.1.2.2 Stochastic Approach

The stochastic approach uses probabilistic models, like Hidden Markov Models (HMM) [6], to manage uncertainty and variability in speech, such as confusable sounds, speaker differences, and contextual influences. Compared to the template-based approach, HMM provides a more flexible and mathematically robust framework. By incorporating probabilities, this method effectively handles speech uncertainties, enhances the ability to differentiate similar sounds, and adapts to diverse speaking styles, making it a reliable solution for automatic speech recognition in varied conditions.

2.1.3 Knowledge-Based Approach

The knowledge-based approach to speech recognition mimics human analysis of speech, utilizing acoustic features to identify and characterize spoken language. It integrates aspects of both the acoustic-phonetic and pattern recognition methods, overcoming some of their limitations related to error analysis and system improvement. Traditional systems use heuristic rules based on linguistic knowledge and speech spectrogram observations to enhance performance in input representation, speech unit definition, and algorithm design.

Samouelian introduced a data-driven method for continuous speech recognition (CSR), using inductive inference to derive knowledge from a speech database, generating decision trees to handle speaker variability. However, limited recognition performance was observed, partly due to the small study sample size. Tripathy et al. improved this approach with a fuzzy inference algorithm for classifying English vowels, achieving better results than conventional Mel-Frequency Cepstral Coefficients (MFCC) analysis and highlighting the potential of knowledge-based techniques in advancing speech recognition accuracy.

2.1.4 Connectionist Approach

The connectionist approach to speech recognition focuses on integrating knowledge from multiple sources, using simple computing units interconnected in a network to simulate the functioning of the nervous system. Rather than encoding knowledge as individual rules or

probabilities, this method distributes it through the connections between units, representing uncertainty as patterns of activity. The simplicity and uniformity of these models make them well-suited for hardware implementation and efficient network simulation.

Connectionist models are seen as promising for cognitive tasks. For instance, Savage et al. improved recognition rates by integrating an Artificial Neural Network (ANN) with Vector Quantization (VQ), although the overall increase was modest. To reduce quantization errors in traditional VQ and Hidden Markov Models (HMM) [6], Debyeche et al. developed a hybrid Distributed Vector Quantization (DVQ) technique, with NN-DVQ showing significant error reduction. Hatulan et al. applied an ANN/HMM method to the Filipino language, training feed-forward ANNs whose outputs fed into an HMM, effectively predicting phoneme sequences for speech-to-text conversion.

2.1.5 Support Vector Machines

Support Vector Machines (SVMs) are state-of-the-art classifiers commonly used in pattern recognition, employing a discriminative approach that optimizes the margin between data samples and the classifier's decision boundary. SVMs are effective in both linear and nonlinear data classification but require variable-length data to be converted into fixed-length vectors before use.

SVMs are generalized linear classifiers that ensure better generalization through maximum-margin fitting and control of Vapnik-Chervonenkis (VC) dimensions. This allows them to operate effectively in high-dimensional spaces and handle complex nonlinear features while keeping model complexity under control.

Sendra et al. developed an SVM-based continuous speech recognizer, using SVMs for frame-level classification and a Token Passing algorithm—an adaptation of the Viterbi algorithm—to recognize word sequences. Their results showed that SVMs improved recognition accuracy on small databases but required considerable computational resources to maintain accuracy on larger databases.

2.2 Approaches to Diarization

Data from domains like Broadcast News (BN), Meetings, and Conversational Telephone Speech (CTS) vary in several ways, such as the quality of recordings, the presence of background noise, the number of speakers, how long each speaker talks, and the style of their speech. [19] These differences create unique challenges when it comes to indexing and processing speech data from these domains. For example, indexing speech from BN might involve handling clearer audio but more formal speech, while CTS may involve more casual speech with overlapping conversations

and poorer audio quality. Although some methods can work across multiple domains, each one still requires specialized approaches to deal with its particular challenges.

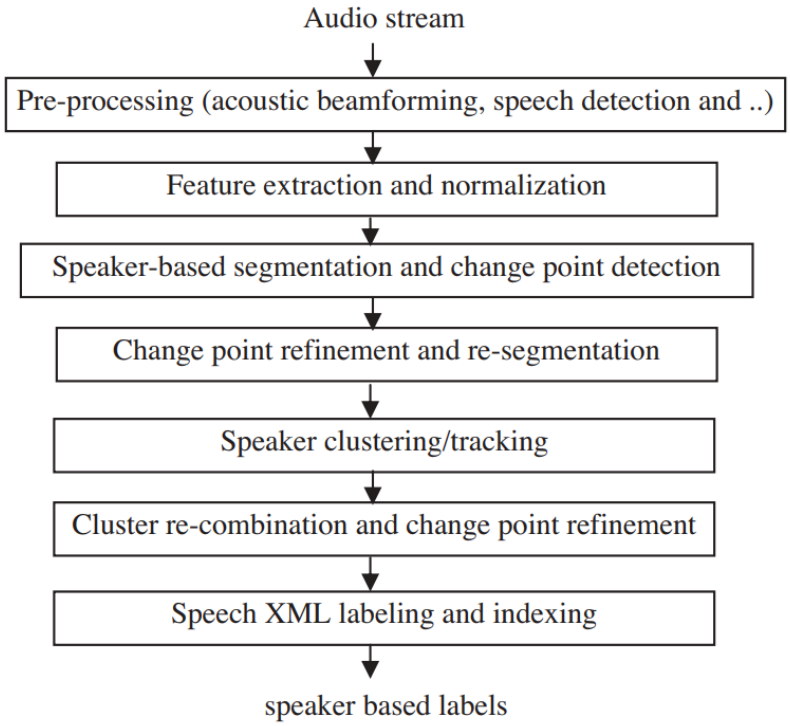


Figure 1.2. A general speaker diarization framework.

Speaker diarization systems typically follow a structured approach with key steps. The process often begins with detecting speech and non-speech segments or simply removing silence. In some cases, such as multi-microphone (MDM) meetings, this step may be preceded by acoustic beamforming, but that's not always necessary. These initial tasks are part of a broader preprocessing step.

Next comes feature extraction, a crucial stage in all speech and signal processing tasks. After that, the speech stream is segmented whenever there is a speaker change. Sometimes, this is followed by a refinement or re-segmentation step to improve the accuracy of those changes and reduce false alarms.

Following segmentation, the speech segments are grouped based on speaker characteristics through clustering. This clustering may be further refined to create more accurate speaker groups, which can help minimize false detections. The process typically ends with the formation of speaker clusters, though in some cases, it can continue with speaker labeling and indexing for more detailed results.

2.2.1 Preprocessing Approaches

The application of speaker diarization [20] in the meeting domain has led to the development of methods for managing multiple microphones, which are often deployed in different locations throughout a room. These microphones vary in type and quality, including wall-mounted microphones for speaker localization, lapel microphones, desktop microphones, and microphone arrays. The diversity in microphone setups necessitates new approaches for diarization systems that can effectively handle multi-channel input.

Approaches to Multi-Channel Diarization

Various algorithms have been proposed to extend mono-channel diarization to multi-channel environments. One approach involves performing speaker diarization independently on each channel, followed by a merging algorithm that integrates the results. Another method focuses on late-stage fusion, where diarization is performed on all channels, and only the one with the best signal-to-noise ratio (SNR) is used. Additionally, techniques like pre-processing the acoustic signals to merge them into a single channel before applying standard diarization algorithms have been explored.

Beamforming in Multi-Channel Systems

Since the NIST RT‘05 evaluation, acoustic beamforming has become the most common approach for multi-channel speaker diarization. Beamforming techniques are categorized into two types: data-independent (fixed) and data-dependent (adaptive). Fixed beamforming techniques, such as Delay-and-Sum (D&S), are simpler to implement but less effective at reducing noise compared to adaptive techniques. Adaptive beamforming, like the Generalized Sidelobe Canceller (GSC), offers better noise reduction but is more computationally intensive and sensitive to errors.

Post-Processing and Refinement

Post-processing techniques are often applied after beamforming to further enhance the signal. Approaches such as Wiener filtering have been used to improve signal quality by reducing noise. In real-world applications, post-processing is essential for optimizing the output of multi-channel diarization systems.

BeamformIt Toolkit

The BeamformIt toolkit is a widely used, open-source software that applies an enhanced delay-and-sum algorithm for acoustic beamforming. It allows for dynamic weighting of signal channels based on SNR and includes additional algorithms for selecting the optimal reference channel. Although more sophisticated beamforming methods exist, delay-and-sum remains the most reliable when microphone locations and configurations are unknown.

2.2.2 Feature Extraction and Normalization Approaches

Feature extraction plays a critical role in signal processing applications, significantly influencing system performance. In speaker diarization, features are extracted from acoustic signals to distinguish between different speakers in conversations. According to Kinnunen and Li (2010), ideal features for speaker modeling and discrimination should exhibit high between-speaker variability, low within-speaker variability, robustness against noise and distortion, frequent occurrence in speech, ease of measurement, and a relatively low number of features.

Commonly Used Features

Mel-frequency cepstral coefficients (MFCCs) are the most widely used features in speaker diarization systems. The order of MFCCs varies among different studies, with orders ranging from 12 to 24 commonly used. Other features include short-time energy (STE), zero-crossing rate (ZCR), pitch, spectrum magnitude, line spectrum pairs (LSPs), and perceptual linear prediction (PLP) coefficients. Additionally, phoneme duration, speech rate, silence detection, and prosodic features have also been explored for speaker diarization tasks.

Fusion of Features

A common approach in speaker diarization involves combining multiple features in parallel to leverage their collective discrimination power. For instance, fusion techniques using MFCCs, LSPs, and pitch features in a Bayesian Network [22] have demonstrated improved speaker segmentation performance. Similarly, Yamaguchi et al. (2005) introduced new features such as temporal stability of power spectra, spectral shape, and white noise similarities, which have been shown to enhance speaker segmentation systems. [23] [24]

Long-Term vs. Short-Term Features

While short-term features like MFCCs are effective in many systems, they may not fully capture the individual characteristics of speakers' voices or speaking behaviors. Research has demonstrated that combining long-term features, such as pitch, energy, formants, harmonics-to-noise ratio (HNR), and long-term average spectrum with short-term features, improves speaker diarization results. Long-term features provide additional insights into speaker behavior that short-term features may overlook.

Feature Normalization Techniques

Normalization techniques are essential for improving the performance of speaker diarization systems. Methods like RASTA-filtered cepstra mean and variance normalization, cepstral mean normalization (CMN), and feature warping are used to mitigate the impact of background noise and other non-speaker-related factors. Feature warping, in particular, has proven to be highly

effective in speaker verification tasks, [21] transforming the probability density function of features into a Gaussian shape prior to modeling.

TDOA and Acoustic Features

In multi-microphone (MDM) meeting diarization scenarios, time delay of arrival (TDOA) features are used for short-term speaker segmentation. TDOA features are often combined with acoustic features, such as MFCCs, to provide better speaker identity information. Studies have shown that integrating TDOA with spectral features improves the performance of diarization systems. Despite challenges posed by noise and reverberation in distant microphones, the combination of TDOA and acoustic features remains a standard approach in Rich Transcription Evaluation projects.

In summary, feature extraction and normalization are key elements of speaker diarization systems. Combining short-term and long-term features with techniques like TDOA significantly enhances the system's ability to separate and identify speakers in multi-channel environments accurately.

2.2.3 Speaker-Based Segmentation and Change Point Detection Approaches

Speech Activity Detection (SAD) or Voice Activity Detection (VAD) [7] is crucial in diarization frameworks as it identifies speech regions in an audio stream, which other subtasks heavily rely upon. Depending on the domain, non-speech regions may include silence, music, room noise, or cross-talk. Several approaches have been proposed to improve the accuracy of SAD systems, each with its advantages and limitations:

Common Approaches to SAD:

A. **Gaussian Mixture Models (GMMs):** These models classify speech/non-speech regions based on labeled training data. [15] The simplest form includes only two models for speech and non-speech, while more complex versions integrate various acoustic classes like noise and music (Nguyen et al., 2002; Sun et al., 2010).
B. **Feature-Based Models:** SAD can also involve the detection of silence or commercials by using energy constraints, phone recognizers, or word recognizers (Tranter et al., 2004). For example, **zero-crossing rate (ZCR)** and **short-time energy (STE)** have been applied in earlier approaches.
C. **Hybrid Approaches:** Some approaches combine model-based and unsupervised methods to ensure robustness under different acoustic conditions. These systems often start with energy-based detection and proceed to model-based detection, iteratively training speech/non-speech models (Anguera et al., 2006; Huijbregts et al., 2007).

D. **Energy-Based Detection**: Detecting speech by analyzing modulation energy or signal autocorrelation works well unless the audio has high-energy non-speech regions, such as loud background noise.

Challenges and Innovations:

A. **Robustness to Acoustic Changes**: Traditional model-based approaches rely on external data, making them less adaptive to environmental shifts. Hybrid models, which blend energy-based methods with HMM [6] or GMM techniques, aim to mitigate this limitation.
B. **Unsupervised Methods**: Newer approaches propose robust unsupervised techniques that require minimal training data. For example, derivative filters combined with finite state machines (FSM) have been employed to label speech and non-speech regions.
C. **Special Cases for Different Domains**: The SAD approaches vary depending on the domain, such as telephony, meetings, or broadcast news. For example, meeting audio contains varied noise sources like coughing or laughing, which necessitate careful segmentation to avoid misclassifying non-speech events as speech.

Example Systems:

A. **ICSI-SRI SAD System**: This system evolved from a model-based approach in 2005 to a two-step algorithm in 2006, focusing first on energy-based segmentation and then using HMM for realignment and training. [16] This method avoids the need for external training data and relies on the assumption that non-speech in meetings is typically silence (Anguera et al., 2006).

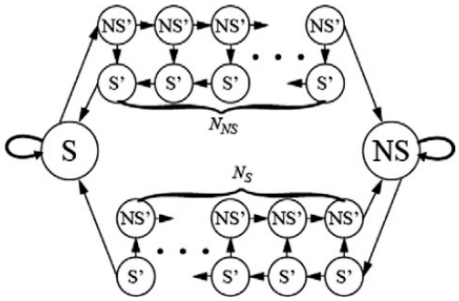


Figure 1.3. Structure of ICSI-SRI HMM-based speech/non-speech detector (Pfau and Ellis, 2001)

B. **LIA Meeting Diarization System**: LIA uses 12 linear frequency cepstral coefficients (LFCCs) augmented with their derivatives as features in an iterative Viterbi decoding process. This approach refines the speech/non-speech segmentation through state

duration rules and achieves robust performance with minimal external data (Evans et al., 2009).

Overall, SAD systems continue to evolve with hybrid methods and feature-rich approaches aimed at balancing robustness, adaptability, and accuracy across different acoustic environments.

2.2.4 Speaker Segmentation Approaches

Speaker segmentation and clustering are essential for tasks like speaker diarization. Segmentation methods include silence detection, energy-based, and model-based approaches, with hybrid techniques combining these for better accuracy. Clustering groups segments by speaker, evaluated using metrics like Average Cluster Purity (ACP). [14] Offline clustering methods, like bottom-up and top-down, optimize segment grouping, while online clustering handles real-time data. Hybrid and evolutionary HMM-based clustering combine multiple techniques to improve performance, balancing accuracy and efficiency based on the application's needs.

2.3 Survey of Commercially Available ASR Services and Speaker Diarization Capabilities

2.3.1. Google Speech Recognition (speech_recognition module)

- Languages Supported: Supports multiple languages including English, Bengali, Hindi, Spanish, French, German, Italian, and more, with varying degrees of accuracy depending on the language.
- Features: The speech_recognition module is a Python library that facilitates the integration of the Google Speech-to-Text [9] API and other speech recognition systems like Sphinx and Microsoft Bing Voice Recognition. It supports audio file formats like WAV, AIFF, FLAC, and MP3, and provides features like real-time transcription, background noise filtering, and automatic punctuation.
- Speaker Diarization: The module itself does not offer built-in speaker diarization, but can be used in combination with other libraries like pyannote-audio or services like Google Speech-to-Text API for speaker segmentation. It is best suited for simple transcription tasks.
- Open-Source: Yes, the speech_recognition module is open-source.

2.3.2. Amazon Transcribe

- **Languages Supported**: 37 languages.
- **Features**: Custom vocabulary filtering, auto-punctuation, supports WAV, MP3, FLAC, MP4 formats. Available in Java, Python, Node.js, Ruby, Go, PHP, .NET.
- **Speaker Diarization**: Supports real-time and batch transcriptions (up to 4 hours). In real-time transcription, speaker diarization is only available in "English-US"

and can detect 2-5 speakers, with accuracy decreasing as the number of speakers increases.
- **Open-Source**: No, Amazon Transcribe is a paid service.

2.3.3. Rev.ai

- **Languages Supported**: 31 languages.
- **Features**: Automatic Speaker Recognition, punctuation, capitalization, speaker diarization, timestamp generation. Supports various formats (WAV, MP3, OGG) and files up to 5 GB. SDKs available in Java, Python, Node.js.
- **Speaker Diarization**: Available as a default feature for recorded media files, and can label all speakers in the media.
- **Open-Source**: No, Rev.ai is a paid service.

2.3.4. IBM Watson Speech-to-Text

- **Languages Supported**: 13 language variants.
- **Features**: Pre-trained models, model training, fine-tuning, low-latency transcription, interim transcription, smart formatting, speaker diarization, word spotting, and filtering. SDKs available for Android, Go, Java, Node.js, Python, Ruby, .NET, Swift, Unity.
- **Speaker Diarization**: Available for telephony models with a limit of 6 speakers for accurate performance. Performance drops when the number of speakers exceeds six.
- **Open-Source**: No, IBM Watson is a paid service.

2.3.5. pyannote-audio

- **Languages Supported**: Primarily focused on English, but can be extended to other languages with additional training data.
- **Features**: pyannote-audio [12] is an open-source Python library for speaker diarization and speech activity detection. It provides pre-trained models for speaker segmentation, speaker embedding, and diarization, and supports fine-tuning models on custom datasets.
- **Speaker Diarization**: Offers robust diarization features like speaker segmentation and clustering. It can detect speaker boundaries and assign speakers to different segments. Ideal for offline tasks and research-based applications, and supports formats such as WAV and MP3.
- **Open-Source**: Yes, pyannote-audio is open-source. [13][18]

Chapter 3
Datasets

Five distinct datasets were chosen across diverse conversation domains for this research work. The selected datasets cover a broad range of topics, ensuring that the models are tested on various linguistic styles, contexts, and semantic structures. Each dataset also integrates elements like gender change and variations in person references, to ensure the models' robustness in handling pronoun shifts and diverse sentence structures. Together, these datasets provide a rigorous testing ground, evaluating the models' effectiveness in understanding and transforming sentences while maintaining meaning and adapting to gendered language and different personal perspectives. Below is a detailed description of each dataset:

3.1 Casual Political Conversation Dataset

The audio file named Audio1.wav is 157 seconds long and features a conversation between two speakers: one male and one female. The file is segmented into 14 parts for analysis. The English content comprises 20 words and 144 characters, while the Bengali section is more extensive, containing 317 words and 1968 characters. Additionally, the Hindi portion includes 33 words and 322 characters. For translation purposes, the Bengali text is translated into 352 English words, totaling 2003 characters, while the Hindi text is translated into 48 English words, with 333 characters.

A tabular description of the dataset.

Audio File Name	Audio1.wav
Length (Sec)	157
No of Speaker	2
Male Speaker Count	1
Female Speaker Count	1
Total No of Segmentation	14
Total English Text Words	20
Total English Text Characters	144
Total Bengali Text Words	317
Total Bengali Text Characters	1968
Total Hindi Text Words	33

Total Hindi Text Characters	322
Translated Bengali to English Words	352
Translated Bengali to English Characters	2003
Translated Hindi to English Words	48
Translated Hindi to English Characters	333

Table 3.1: Casual Political Conversation Dataset Summary

Speaker	Start	End	split_audio_file_path	English_Text	Bengali_Text	Translated_BE	Hindi_Text	Translated_HE
SPEAKER_01	0.0309688	1.04346875	/content/drive/MyDrive/split_audio/segment_0.wav	politics	পলিটিক্স	Politics	पॉलिटिक्स	Politics
SPEAKER_01	1.2290938	12.3159688	/content/drive/MyDrive/split_audio/segment_1.wav		পলিটিক্স যার বাংলা অর্থ এসে দাঁড়ায় রাজনীতি কিন্তু আজকের সমাজে রাজনীতি এবং পলিটিক্স এই দুটো ওয়ার্ডকে পাশাপাশি রাখলে তাদের ক্ষেত্রে কিন্তু কোন মিল খুঁজে পাওয়া যায় না	Politics, which means politics in Bengali, but in today's society, if these two words are put side by side, there is no similarity between them.	पॉलिटिक्स	Politics
SPEAKER_01	13.260969	26.7103438	/content/drive/MyDrive/split_audio/segment_2.wav	Helen Haven difference	পলিটিক্স এর বাংলা অর্থ অবশ্যই রাজনীতি কিন্তু পলিটিক্স যারা আজকের দিনে করেন তাদের সাথে রাজনীতি যারা করে গেছেন তাদের মধ্যে কিন্তু একটা বিস্তর ফারাক একটা হেলান হেভেন ডিফারেন্স	The Bengali meaning of politics is of course politics, but there is a big difference between those who do politics today and those who have done politics, a lean heaven difference.	पॉलिटिक्स सैड बांग्ला डिफरेंस	Politics Sad Bangla Difference

Figure 3.1: Snapshot of Casual Political Conversation Dataset

3.2 Casual Educational Conversation Dataset

The audio file named Audio2.wav has a duration of 88 seconds and features a conversation between two female speakers. The file is divided into 9 segments. It contains 10 English words and 59 characters, while the Bengali text consists of 172 words and 1013 characters. The Hindi segment includes 22 words and 161 characters. Upon translation, the Bengali content expands to 174 English words, amounting to 998 characters. Similarly, the Hindi text translates into 29 English words, totaling 173 characters.

A tabular description of the dataset.

Audio File Name	Audio2.wav
Length (Sec)	88
No of Speaker	2
Male Speaker Count	0
Female Speaker Count	2
Total No of Segmentation	9
Total English Text Words	10
Total English Text Characters	59
Total Bengali Text Words	172
Total Bengali Text Characters	1013
Total Hindi Text Words	22
Total Hindi Text Characters	161
Translated Bengali to English Words	174
Translated Bengali to English Characters	998
Translated Hindi to English Words	29
Translated Hindi to English Characters	173

Table 3.2: Casual Educational Conversation Dataset Summary

Speaker	Start	End	split_audio_file_path	English_Text	Bengali_Text	Translated_BE	Hindi_Text	Translated_HE
SPEAKER_00	0.520344	14.93159	/content/drive/MyDrive/split_audio/segment_0.wav		নমস্কার নমস্কার আপনি কেমন আছেন আমি ভালোই আছি আপনি কেমন আছেন আমিও ভালোই আছি আজকে আমরা শিক্ষার গুরুত্ব নিয়ে কথা বলব আপনি কি মনে করেন বর্তমান যুগে শিক্ষার ভূমিকা কতটা গুরুত্বপূর্ণ	hello hello how are you i am fine how are you i am also fine today we will talk about the importance of education do you think the role of education is important in this day and age	আজ কৈমরা	The camera today
SPEAKER_01	15.16784	18.34034	/content/drive/MyDrive/split_audio/segment_1.wav		আমি মনে করি শিক্ষা শুধু বই পড়া	I think education is just reading books	অমী মোন কোরী শিখর	Ami Mon Cori peak
SPEAKER_01	18.42472	28.81972	/content/drive/MyDrive/split_audio/segment_2.wav		এটা আমাদের মানসিকতার দক্ষতাকে গড়ে তোলে শিক্ষা আমাদের চিন্তাধারাও জ্ঞানকে প্রসারিত করে এবং বিভিন্ন বিষয়ে দক্ষতা সাহায্য করে	It develops our mental skills Education broadens our thinking and knowledge and helps us to master various subjects	মনুষ্যতা	humanity
SPEAKER_00	29.05597	43.16347	/content/drive/MyDrive/split_audio/segment_3.wav	online account on kitchen	ঠিক বলেছেন বিশেষ করে আজকের ডিজিটাল যুগে প্রযুক্তি শিক্ষার অংশ হয়ে উঠেছে অনলাইন এখন অনেক কিছু শেখা যায় স্কুল বা বিশ্ববিদ্যালয় না গিয়ে আমরা অনেক দক্ষতা অর্জন করতে পারি	Rightly said especially in today's digital age technology has become a part of education online now many things can be learned without going to school or university we can acquire many skills	আজ কে ডিজিটল প্রোজেক্ট শিখর অঁনলাইন অকাউঁট	Today's Digital Project Shikhar Online Account
						With the opportunity of online		

Figure 3.2: Snapshot of Casual Educational Conversation Dataset

3.3 Casual Sports Conversation Dataset

The audio file named Audio3.wav is 123 seconds long and features two speakers: one male and one female. It is divided into 17 segments. The English content consists of 54 words and 308 characters, while the Bengali text has 284 words and 1550 characters. The Hindi section includes 55 words and 464 characters. When translated, the Bengali text results in 315 English words, totaling 1556 characters, and the Hindi text translates into 83 English words, amounting to 499 characters.

A tabular description of the dataset.

Audio File Name	Audio3.wav
Length (Sec)	123
No of Speaker	2
Male Speaker Count	1
Female Speaker Count	1
Total No of Segmentation	17
Total English Text Words	54
Total English Text Characters	308
Total Bengali Text Words	284
Total Bengali Text Characters	1550
Total Hindi Text Words	55
Total Hindi Text Characters	464
Translated Bengali to English Words	315
Translated Bengali to English Characters	1556
Translated Hindi to English Words	83
Translated Hindi to English Characters	499

Table 3.3: Casual Political Sports Dataset Summary

Speaker	Start	End	split_audio_file_path	English_Text	Bengali_Text	Translated_BE	Hindi_Text	Translated_HE
SPEAKER_00	0.030969	0.537219	/content/drive/MyDrive/split_audio/segment_0.wav					
SPEAKER_00	1.127844	6.831594	/content/drive/MyDrive/split_audio/segment_1.wav		এখনতো ক্রিকেটে ওয়ানডে এবং টেস্ট এর সাথে সাথে t-20 একটা সংযোজন হয়েছে	Now t-20 is an addition to cricket along with ODIs and Tests	অখন্ড ক্রিকেট বনউ T20	Akhand Cricket ODI T20
SPEAKER_00	7.692219	10.00409	/content/drive/MyDrive/split_audio/segment_2.wav	to at home without	তো এখন বিরাট রহ	So now be great	তো অকাউট বিরাট রোহিত	So the account is Virat Rohit.
SPEAKER_00	10.67909	11.18534	/content/drive/MyDrive/split_audio/segment_3.wav		কি	what		
SPEAKER_00	11.75909	12.88972	/content/drive/MyDrive/split_audio/segment_4.wav	Jada	জাদু যা আশ	Magic is what it is	জউজা	Jadeja
SPEAKER_00	13.17659	14.34097	/content/drive/MyDrive/split_audio/segment_5.wav	11	এরা যেমন খেল	As they play	হীরা জামুন খেল	Diamond Jamun Game
SPEAKER_00	14.62784	17.10847	/content/drive/MyDrive/split_audio/segment_6.wav	play the piano teleport that has	এদের তো এখন একটা দাপট দেখায় যায়	Now they can see a power		
			/content/drive/MyDrive/split_audio/segment_7.		পুরনো দিনের সময় যখন টেস্ট এবং হচ্ছে ওয়ানডে	Who remembers the old days when there were		

Figure 3.3 Snapshot of Casual Political Sports Dataset

3.4 Casual Weather Conversation Dataset

The audio file named Audio4.wav has a duration of 116 seconds and features two speakers: one male and one female. It is segmented into 20 parts. The English text consists of 47 words and 274 characters, while the Bengali section includes 240 words and 1414 characters. The Hindi content comprises 29 words and 306 characters. The translation of the Bengali text yields 292 English words and 1435 characters, while the Hindi text translates into 40 English words, totaling 260 characters.

A tabular description of the dataset.

Audio File Name	Audio4.wav
Length (Sec)	116
No of Speaker	2
Male Speaker Count	1
Female Speaker Count	1
Total No of Segmentation	20
Total English Text Words	47
Total English Text Characters	274

Total Bengali Text Words	240
Total Bengali Text Characters	1414
Total Hindi Text Words	29
Total Hindi Text Characters	306
Translated Bengali to English Words	292
Translated Bengali to English Characters	1435
Translated Hindi to English Words	40
Translated Hindi to English Characters	260

Table 3.4: Casual Weather Coversation Dataset Summary

Speaker	Start	End	split_audio_file_path	English_Text	Bengali_Text	Translated_BE	Hindi_Text	Translated_HE
SPEAKER_00	0.030969	0.047844	/content/drive/MyDrive/split_audio/segment_0.wav	it was really interested	এবছরের ক্ষেত্রে ঠান্ডাটা এখনো যে পরেনি প্রত্যেক বারে মোটামুটি কালীপুজোর পর থেকে বা কালিপুজোর সময়	In this year's case, the	स्टैंडर्ड अकाउंट	Standard Account
SPEAKER_00	0.182844	6.409719	/content/drive/MyDrive/split_audio/segment_1.wav	1002	ঠান্ডাটা কিন্তু চলে	The cold is gone.	को	To
SPEAKER_00	6.831594	8.114094	/content/drive/MyDrive/split_audio/segment_2.wav	Kiki Evolution calculator	কি কি এবছর সেম ভাবে আসেনি এখনো ঠান্ডা বেশ গরম গরম	What has not come in	ठंडा ठंडा	cool cool
SPEAKER_01	8.839719	14.03722	/content/drive/MyDrive/split_audio/segment_3.wav	statistical	যথেষ্টই গরম	hot enough	स्टैटिसटिक्स	Statistics
SPEAKER_01	14.88097	16.28159	/content/drive/MyDrive/split_audio/segment_4.wav		কিন্তু এদিকে আবার বলছে 15 তারিখের পর থেকে	But meanwhile again	एयरपोर्ट	airport
					আমার ঠান্ডাটা ঝাঁকিয়ে বসবে মানে আগের বছরের ক্ষেত্রে ঠান্ডাটা কিন্তু যখন পড়েছিল তখন বেশ করেছিল আর			

Figure 3.4 : Snapshot of Casual Weather Dataset

3.5 Casual Innovation Dataset

The audio file named Audio5.wav is 78 seconds long and features two female speakers. It is segmented into 11 parts. The file does not contain any English words or characters. The Bengali content comprises 143 words and 833 characters, while the Hindi section includes 14 words and

112 characters. When translated, the Bengali text yields 140 English words and 793 characters, and the Hindi text translates into 18 English words, retaining 112 characters.

A tabular description of the dataset.

Audio File Name	Audio5.wav
Length (Sec)	78
No of Speaker	2
Male Speaker Count	0
Female Speaker Count	2
Total No of Segmentation	11
Total English Text Words	0
Total English Text Characters	0
Total Bengali Text Words	143
Total Bengali Text Characters	833
Total Hindi Text Words	14
Total Hindi Text Characters	112
Translated Bengali to English Words	140
Translated Bengali to English Characters	793
Translated Hindi to English Words	18
Translated Hindi to English Characters	112

Table 3.5: Casual Innovation Conversation Dataset Summary

Speaker	Start	End	split_audio_file_path	English_Text	Bengali_Text	Translated_BE	Hindi_Text	Translated_HE
SPEAKER_00	0.992844	1.667844	/content/drive/MyDrive/split_audio/segment_0.wav		নমস্কার	hello		
SPEAKER_00	3.203469	4.047219	/content/drive/MyDrive/split_audio/segment_1.wav		নমস্কার	hello		
SPEAKER_00	4.266594	5.498469	/content/drive/MyDrive/split_audio/segment_2.wav		আপনি কেমন আছেন	how are you	अपने कामों नाच	Dancing your chores
SPEAKER_00	5.937219	8.198469	/content/drive/MyDrive/split_audio/segment_3.wav		আমি ভালো আছি আপনি কেমন আছেন	i am fine how are you		
SPEAKER_01	8.502219	9.683469	/content/drive/MyDrive/split_audio/segment_4.wav		আমি ভালোই আছি	i am fine	आमी भालो ए	We spear a
			/content/drive/MyDriv		আজ আমরা উদ্ভাবন নিয়ে কথা বলবো আপনি কি মনে করেন উদ্ভাবন	Today we will talk about innovation how do you think		

Figure 3.5: Snapshot of Casual Innovation Dataset

Chapter 4
Experiments and Results

In this study, we conducted experiments on five distinct datasets: Casual Political Conversation dataset, Casual Educational Conversation Dataset, Casual Sports Conversation Dataset, Casual Educational Weather Dataset, and Casual innovation Conversation Dataset to extract various cause-and-effect events.

4.1 Diarization Error Rate

The Diarization Error Rate (DER) is a widely used evaluation metric in the field of speaker diarization, which is the process of segmenting and labeling audio based on the different speakers present. This metric plays a crucial role in assessing the performance of diarization systems, particularly in tasks like speech recognition, transcription, and audio analysis for various applications such as meeting recordings, broadcast news, and call center analytics.

Definition and Components

The Diarization Error Rate (DER) quantifies the errors made by a diarization system by comparing the system's output with a ground truth reference. DER is expressed as a percentage and encompasses three main types of errors:

1. **Missed Speech**: This refers to segments where speech was present in the ground truth but was not detected by the diarization system. Such errors occur when the system fails to recognize a speaker's presence, often due to low volume, background noise, or overlapping speech.
2. **False Alarm**: These are segments where the diarization system falsely identifies speech in parts of the audio where no speech exists. False alarms can result from background noises or environmental sounds being mistakenly labeled as speech.
3. **Speaker Confusion**: This occurs when the diarization system correctly detects speech but attributes it to the wrong speaker. Speaker confusion errors are critical as they directly impact the system’s ability to separate and label speakers accurately, which is essential for understanding conversational dynamics.

Mathematical Formula

The Diarization Error Rate is mathematically defined as:

$$\text{DER} = \frac{\text{Total Time of Errors}}{\text{Total Time of Speech}} \times 100$$

…… Formula 1

Where:

- **Total Time of Errors** is the sum of the durations of missed speech, false alarms, and speaker confusion errors.
- **Total Time of Speech** refers to the total duration of all speech present in the ground truth.

Expressing DER as a percentage helps standardize the metric, making it easier to compare performance across different systems and datasets.

Detailed Breakdown of DER Components

Missed speech in speaker diarization occurs when the system fails to detect parts of the conversation, often due to factors like poor audio quality, high levels of background noise, or overlapping speech. When one speaker's voice is masked by another's, the system may miss crucial segments of the conversation. This leads to incomplete transcriptions or analyses, impacting the overall recall rate of the system, as significant portions of speech are unaccounted for.

False alarms are another common issue, where non-speech sounds such as ambient noise, music, or environmental sounds are incorrectly recognized as speech. These errors degrade the system's precision, as they introduce incorrect information into the output. False alarms can clutter the transcription or speaker labels with irrelevant data, leading to confusion and a reduction in the overall accuracy of the diarization system.

Speaker confusion happens when the system incorrectly assigns speech to the wrong speaker, which is particularly challenging in settings with multiple speakers or voices that are acoustically similar. This issue is critical for applications like meeting transcriptions or conversational analysis, where accurately attributing statements to the correct speaker is essential. Misattributions can lead to significant misunderstandings, making it a key area for improvement in diarization systems.

Practical Implications of DER

A low Diarization Error Rate (DER) indicates a highly accurate speaker diarization system, with fewer errors in segmenting and labeling speakers. A high DER suggests difficulties in proper speaker separation, which can affect applications like meeting minutes generation, speech-to-text conversion, or voice-based interfaces. DER is a crucial metric that includes missed speech, false alarms, and speaker confusion, offering a comprehensive view of system performance. Despite technological advancements, achieving a consistently low DER remains challenging, especially in noisy environments, though ongoing research continues to improve system accuracy.

Ground Truth in Audio Diarization

In the context of audio diarization, *ground truth* refers to the accurate, human-annotated reference data that defines the true segmentation and speaker labels of an audio recording. It is essentially the benchmark against which the performance of a diarization system is evaluated. The ground truth specifies who is speaking and when with precise start and end times for each speaker's segment.

Components of Ground Truth for Audio Diarization

1. **Speaker Segments**: The audio is divided into segments, each attributed to a specific speaker. The start and end times for each segment are precisely marked, indicating when each speaker begins and stops talking.
2. **Speaker Labels**: Each segment is labeled with an identifier corresponding to the actual speaker. For example, in a meeting recording, labels might indicate Speaker A, Speaker B, etc., or use the speakers' names if known.

Ground truth data is essential for the development and refinement of diarization systems, as it ensures that the systems are tested and improved against a reliable and accurate reference.

4.1.1 DER on Casual Political Conversation Dataset

Ground Truth		
Speaker	Start	End
SPEAKER 00	0.283	89.034
SPEAKER 01	90.844	156.235

Table 4.1: Ground Truth of Casual Political Conversation Dataset

Comparing Table 4.2 with Figure 3.1, Calculating Flase Alarm, Miss and Speaker Error

False Alarm		
Ground truth	Pyannote value	False Alarm Duration
0.283	0.03096875	0.25203125
90.844	90.80159375	0.04240625
	Total False Alarm Duration	0.2944375

Table 4.2: False Alarm of Casual Political Conversation Dataset

Miss		
Ground truth	Pyannote value	False Alarm Duration
89.034	89.02971875	0.00428125
156.235	156.2259688	0.00903125
	Total Miss Duration	0.0133125

Table 4.3: Miss of Casual Political Conversation Dataset

Speaker Error		
	Total Speaker Error Duration	0

Table 4.4: Speaker Error of Casual Political Conversation Dataset

Using Formula 1 on the result derived from Table 4.2, Table 4.2, Table 4.4, **Diarization Error Rate (%)**

= 0.196019108 %

4.1.2 DER % on Casual Educational Conversation Dataset

Ground Truth		
Speaker	**Start**	**End**
SPEAKER 00	0.519	1.193
SPEAKER 01	1.157	2.269
SPEAKER 00	2.612	3.571
SPEAKER 01	3.648	5.892
SPEAKER 00	5.923	14.931
SPEAKER 01	15.189	28.698
SPEAKER 00	29.085	43.034
SPEAKER 01	43.349	55.923

SPEAKER 00	56.392	70.308
SPEAKER 01	70.742	79.286
SPEAKER 00	79.672	87.404

Table 4.5: Ground Truth of Casual Educational Conversation Dataset

Comparing Table 4.5 with Figure 3.2, Calculating Flase Alarm, Miss and Speaker Error

False Alarm		
Ground truth	**Pyannote value**	**False Alarm Duration**
14.93159375	14.931	0.00059375
15.189	15.16784375	0.02115625
28.81971875	28.698	0.12171875
43.16346875	43.034	0.12946875
55.93784375	55.923	0.01484375
79.34346875	79.286	0.05746875
79.672	79.63034375	0.04165625
87.46034375	87.404	0.05634375
	Total False Alarm Duration	0.44325

Table 4.6: False Alarm of Casual Educational Conversation Dataset

Miss		
Ground truth	**Pyannote value**	**False Alarm Duration**
0.52034375	0.519	0.00134375
29.085	29.05596875	0.02903125
43.34909375	43.349	9.375E-05
56.46096875	56.392	0.06896875
	Total Miss Duration	0.0994375

Table 4.7: Miss of Casual Educational Conversation Dataset

Speaker Error		
Ending Timeline	**Starting Timeline**	**Speaker Error Duration**
2.269	1.157	1.112
5.892	3.648	2.244
79.286	70.742	8.544
	Total Speaker Error Duration	11.9

Table 4.8: Speaker Error of Casual Educational Conversation Dataset

Using Formula 1 on the result derived from Table 4.6, Table 4.7, Table 4.8, **Diarization Error Rate (%)**

= 14.139418%

4.1.3 DER on Casual Sports Conversation Dataset

Ground Truth		
Speaker	**Start Time**	**End Time**
SPEAKER_00	0.097	22.443
SPEAKER_01	23.395	59.827
SPEAKER_00	61.128	104.116
SPEAKER_01	104.127	122.25

Table 4.9: Ground Truth of Casual Sports Conversation Dataset

Comparing Table 4.9 with Figure 3.3, Calculating Flase Alarm, Miss and Speaker Error

False Alarm		
Ground truth	**Pyannote value**	**False Alarm Duration**
0.097	0.03096875	0.06603125
23.395	23.33534375	0.05965625
61.128	61.10159375	0.02640625
	Total False Alarm	0.15209375

Table 4.10: False Alarm of Casual Sports Conversation Dataset

Miss		
Ground truth	**Pyannote value**	**False Alarm Duration**
22.443	22.44096875	0.00203125
59.827	59.80221875	0.02478125
	Total Miss Duration	0.0268125

Table 4.11: Miss of Casual Sports Conversation Dataset

Speaker Error		
Ending Timeline	**Starting Timeline**	**Speaker Error Duration**
104.127	101.0953438	3.0316562
	Total Speaker Error Duration	3.0316562

Table 4.12: Speaker Error of Casual Sports Conversation Dataset

Using Formula 1 on the result derived from Table 4.10, Table 4.11, Table 4.12, **Diarization Error Rate (%)**

= 2.610213374%

4.1.4 DER % on Casual Weather Conversation Dataset

Ground Truth		
Speaker	**Start**	**End**
SPEAKER_00	0.288	9.227
SPEAKER_01	9.257	16.282
SPEAKER_00	16.51	54.738
SPEAKER_01	55.319	73.922
SPEAKER_00	75.12	97.053
SPEAKER_01	97.596	114.36

Table 4.13: Ground Truth of Casual Sports Conversation Dataset

Comparing Table 4.13 with Figure 3.4, Calculating Flase Alarm, Miss and Speaker Error

False Alarm

Ground truth	Pyannote value	False Alarm Duration
0.288	0.03097	0.25703
9.257	8.83972	0.41728
16.51	16.4672	0.04278
54.73971875	54.738	0.00172
55.319	55.2797	0.03928
75.12	75.1078	0.01216
	Total False Alarm	0.77025

Table 4.14: False Alarm of Casual Sports Conversation Dataset

Miss		
Ground truth	**Pyannote value**	**False Alarm Duration**
9.227	8.11409	1.11291
16.282	16.2816	0.00041
73.922	73.8591	0.06291
97.053	97.0285	0.02453
97.61909375	97.596	0.02309
114.36	106.664	7.69591
	Total Miss Duration	8.91975

Table 4.15: Miss of Casual Sports Conversation Dataset

Speaker Error		
	Total Speaker Error Duration	0

Table 4.16: Speaker Error of Casual Sports Conversation Dataset

Using Formula 1 on the result derived from Table 4.14, Table 4.15, Table 4.16, **Diarization Error Rate (%)**

= 8.91975%

4.1.5 DER % on Casual Innovation Conversation Dataset

Ground Truth		
Speaker	**Start**	**End**
SPEAKER_00	0.984	1.598
SPEAKER_01	3.26	5.472
SPEAKER_00	6.008	8.246
SPEAKER_01	8.577	17.343
SPEAKER_00	17.818	33.488
SPEAKER_01	33.881	47.734
SPEAKER_00	48.212	60.565
SPEAKER_01	61.276	74.176
SPEAKER_00	74.69	78.144

Table 4.17: Ground Truth Casual Sports Innovation Dataset

Comparing Table 4.17 with Figure 3.5, Calculating Flase Alarm, Miss and Speaker Error

False Alarm		
Ground truth	Pyannote value	False Alarm Duration
8.577	8.50221875	0.074781
17.818	17.69909375	0.118906
33.56159375	33.488	0.073594
33.881	33.78096875	0.100031
61.276	61.23659375	0.039406
74.176	74.14596875	0.030031
74.69	74.66909375	0.020906
78.17909375	78.144	0.035094
	Total False Alarm	0.49275

Table 4.18: False Alarm of Casual Sports Innovation Dataset

Miss		
Ground truth	Pyannote value	False Alarm Duration
0.99284375	0.984	0.008844
8.246	8.19846875	0.047531
17.343	17.32784375	0.015156
60.565	60.56159375	0.003406
	Total Miss Duration	0.074937

Table 4.19: Miss of Casual Sports Innovation Dataset

Speaker Error		
Ending Timeline	**Starting Timeline**	**Speaker Error Duration**
5.472	3.26	2.212
60.565	48.212	12.353
78.144	74.69	3.454
	Total Speaker Error Duration	18.019

Table 4.20: Speaker Error of Casual Sports Innovation Dataset

Using Formula 1 on the result derived from Table 4.18, Table 4.19, Table 4.20, **Diarization Error Rate (%)**

= 23.82908654%

4.2 Word Error Rate (WER)

Total Words	Error Words	Word Error Rate %
317	8	2.52365930599

Table 4.21: Word Error Rate Calculated on Casual Political Conversation Dataset

Total Words	Error Words	Word Error Rate %
172	5	2.90697674419

Table 4.22: Word Error Rate Calculated on Casual Educational Conversation Dataset

Total Words	Error Words	Word Error Rate %
284	17	5.98591549296

Table 4.23: Word Error Rate Calculated on Casual Sports Conversation Dataset

Total Words	Error Words	Word Error Rate %
246	12	4.87804878049

Table 4.24: Word Error Rate Calculated on Casual Weather Innovation Dataset

Total Words	Error Words	Word Error Rate %
143	2	1.3986013986

Table 4.25: Word Error Rate Calculated on Casual Innovation Conversation Dataset

Gattering all findings from Table 3.1, Table 3.2, Table 3.3, Table 3.4, Table 3.4, Table 3.5, Table 4.21, Table 4.22, Table 4.23, Table 4.24, Table 4.25, constructing in a table.

Audio File Name	Audio1.wav	Audio2.wav	Audio3.wav	Audio4.wav	Audio5.wav
Length (Sec)	157	88	123	116	78
No of Speaker	2	2	2	2	2
Male Speaker Count	1	0	1	1	0
Female Speaker Count	1	2	1	1	2
Total No of Segmentation	14	9	17	20	11
Total English Text Words	20	10	54	47	0

Total English Text Characters	144	59	308	274	0
Total Bengali Text Words	317	172	284	240	143
Total Bengali Text Characters	1968	1013	1550	1414	833
Total Hindi Text Words	33	22	55	29	14
Total Hindi Text Characters	322	161	464	306	112
Translated Bengali to English Words	352	174	315	292	140
Translated Bengali to English Characters	2003	998	1556	1435	793
Translated Hindi to English Words	48	29	83	40	18
Translated Hindi to English Characters	333	173	499	260	112
False Alarm	0.2944375	0.44325	0.1520937	0.77025	0.49275
Miss Duration	0.0133125	0.0994375	0.0268125	8.91975	0.074937
Speaker Error	0	11.9	3.0316562	0	18.019
DER %	0.196019	14.139418	2.6102133	8.91975	23.829086
WER %	2.523659	2.9069767	5.9859154	4.87804878	1.3986013

Table 4.26 : Findings on 5 Multilingual Datase

4.3 Result Analysis:

In the conducted experiments, several critical observations emerged regarding factors that influence Diarization Error Rate (DER) and Word Error Rate (WER) in speech recognition systems. One of the most impactful factors was the occurrence of false alarms, where the system incorrectly identifies non-speech segments as speech. As false alarms increase, DER rises proportionally, as these erroneous segments skew the diarization output and reduce the accuracy of speaker tracking. Similarly, missed durations, where actual speech segments are overlooked, contribute to a higher DER by leaving parts of

the audio unaccounted for, resulting in an incomplete representation of the spoken content. Additionally, speaker errors misidentifications that result in speech segments being assigned to the wrong speaker add to DER, as they prevent the system from accurately mapping speaker changes throughout the audio stream.

The presence of same-gender speakers presented a notable challenge as well, as interactions between speakers with similar vocal characteristics increased DER. Differentiating between same-gender speakers proved complex for the system, leading to frequent misidentifications. Voice overlapping further exacerbated this issue, significantly elevating both DER and WER. Overlapping speech, where multiple people speak simultaneously, is inherently difficult for diarization systems to handle, as it requires segmenting and attributing speech amid layers of interwoven voices. Finally, noise within the channel, such as background interference or audio distortions, was found to increase both DER and WER by making it difficult for the system to isolate and process speech segments accurately.

These findings highlight the need for careful management of factors such as false alarms, missed durations, speaker errors, same-gender speaker interactions, voice overlaps, and channel noise to improve the precision of DER and WER in speech recognition systems. Addressing these elements can significantly enhance the overall performance, leading to more reliable speaker and word identification.

Chapter 5
Conclusions

In this thesis, we aim to explore the nuances of audio transcription across diverse variables, including multi-gender voices, varying recording durations, and channel conditions. Through a series of experiments, we analyze factors affecting the effectiveness and accuracy of audio diarization and automatic speech recognition (ASR). The findings provide insights into how audio characteristics, speaker variations, and technical configurations impact transcription quality.

5.1. Conclusion of the Present Work

5.1.1. Impact of Audio Length on Processing Time

As the duration of audio increases, the processing time for transcription and diarization also rises. This is primarily due to the increased computational demand required to analyze a longer audio segment, as each additional second demands additional processing for speech-to-text conversion, speaker identification, and segmentation. Efficient handling of extended audio lengths often requires optimized algorithms or processing resources to maintain performance and accuracy without significant time delays.

5.1.2. Gender Variability in Diarization Accuracy

Experiments indicate that diarization accuracy is notably higher when differentiating between male and female voices, compared to cases with two speakers of the same gender. This is because distinct vocal characteristics between genders make it easier for the diarization model to segregate speakers accurately. Voice pitch, tone, and timbre differences between male and female speakers assist in reducing the Diarization Error Rate (DER). However, when both speakers share similar vocal profiles (such as two males or two females), the model struggles more, resulting in a higher DER.

5.1.3. False Alarm Impact on Diarization Error Rate (DER)

False alarms in diarization occur when non-speech sounds or background noise are misclassified as speech from a speaker. An increase in false alarm rates directly elevates the DER. This happens because the algorithm inaccurately attributes speech to a speaker where there is none, leading to erroneous speaker boundaries and reducing overall accuracy. Techniques like enhanced noise filtering or silence detection can help in mitigating false alarms and improving DER.

5.1.4. Miss Duration and Its Effect on DER

Miss duration is defined as segments of audio containing speech that are unaccounted for by the diarization algorithm. An increase in miss duration leads to a rise in DER, as missed speech segments disrupt the continuity of speaker identification, causing the model to lose context. This misattribution or omission lowers the reliability of diarization, especially in environments with intermittent speech or varying vocal intensities.

5.1.5. Speaker Error and Diarization Error Rate (DER)

Speaker errors occur when the diarization model confuses one speaker for another, misassigning segments of speech. An increase in speaker errors directly impacts DER, as it implies that the model is failing to correctly distinguish between different speakers. This may be due to overlapping speech, similar vocal tones, or model limitations in detecting distinct speaker features.

5.1.6. ASR and Sentence Transformer Model Selection

The overall performance of the transcription application depends significantly on the quality of both the diarization and ASR models. The accuracy of ASR in particular is crucial, as it directly influences the word error rate (WER) and speaker-wise WER. Selecting a robust sentence transformer model that accurately captures the context and nuances of the spoken language is equally important, as it enhances the model's ability to differentiate subtle variations in dialogue.

5.1.7. Impact of Miss Duration on Word Error Rate (WER)

When miss duration increases, there is a direct effect on the speaker-wise WER, as speech segments are left unprocessed by the ASR model. Each missed word or phrase decreases the accuracy of the transcription output, leading to incomplete or inaccurate sentence structures. Reducing miss duration through optimized diarization helps minimize this impact and improve overall transcription quality.

5.1.8. False Alarm and Speaker Error Influence on Speaker-wise WER

Higher false alarm rates and speaker errors negatively impact speaker-wise WER, as these factors lead to inaccuracies in recognizing which words were spoken by which speaker. This is particularly detrimental in multi-speaker scenarios, as it introduces ambiguity in attributing words to the correct speaker, leading to potential misunderstandings in the transcription output.

5.1.9. Effect of Noise on Performance

Background noise introduces additional challenges for both diarization and ASR models. Noise, particularly in the form of environmental sounds or speech from unintended sources, can

interfere with speech clarity and accuracy, leading to elevated DER and WER. Implementing noise reduction techniques or using high-quality audio recording equipment can help mitigate these issues and improve transcription accuracy.

5.1.10. Audio Overlapping and Performance

Overlapping speech, where two speakers talk simultaneously, complicates both diarization and ASR processes. Overlapping speech often leads to erroneous speaker segmentation and lowers the reliability of the ASR output, as the model may fail to accurately transcribe words spoken by both individuals. Models capable of handling multi-speaker overlap are more effective in these cases, though such capabilities are still an area of active research and development.

5.1.11. Multi-Channel Audio and Channel Quality

In scenarios where multiple channels are used, such as multi-microphone setups, the quality of each audio channel significantly impacts transcription performance. Poor-quality recordings on one or more channels can distort the overall clarity of the audio, affecting both diarization and ASR results. Ensuring each channel provides clear and distinct audio can improve model performance by minimizing potential sources of error.

This study underscores the intricate relationship between audio properties, model selection, and transcription accuracy. Optimal performance in multi-speaker transcription relies on careful attention to diarization quality, ASR capabilities, and the handling of overlapping speech and noise. By adjusting model parameters and improving recording conditions, the DER and WER can be reduced, leading to more reliable and accurate transcriptions for multi-speaker audio. Future work may focus on enhancing model robustness to better handle complex audio environments, such as highly overlapping or noisy speech, and further refine ASR and diarization methodologies for improved performance in diverse use cases.

5.2. Scope of the Future Work

This thesis has highlighted numerous challenges and findings in multi-speaker audio transcription, particularly in diarization and ASR performance. Building upon these insights, there are several promising directions for future research and development that could further improve the accuracy, robustness, and efficiency of multi-speaker transcription systems. Below are key areas for potential future work:

5.2.1. Advanced Diarization Models with Improved Speaker Discrimination

Current diarization models show limitations, especially in differentiating speakers of similar vocal profiles and handling overlapping speech. Future work could explore advanced deep learning techniques, such as end-to-end neural diarization models, which combine speaker

segmentation and recognition in a unified model. Improved speaker embeddings and clustering methods could also enhance model accuracy, particularly in environments with multiple speakers of the same gender.

5.2.2. Real-Time Transcription and Diarization Solutions

Real-time transcription is increasingly important for applications like live captioning, virtual meetings, and customer service. Developing optimized, low-latency diarization and ASR models would enable real-time processing without compromising accuracy. Leveraging GPU acceleration, model pruning, and streaming ASR methods could improve processing speed, making transcription solutions feasible for real-time applications with multiple speakers.

5.2.3. Enhancement of Overlapping Speech Handling

Speech overlap [11] is one of the major challenges in multi-speaker transcription. Future research could focus on models specifically designed to handle overlapping speech by separating speakers on a more granular, frame-by-frame basis. Speaker separation techniques, which involve isolating individual voices within overlapping segments, could work in tandem with diarization to yield cleaner, more accurate transcriptions in multi-speaker scenarios.

5.2.4. Noise Robustness and Environmental Adaptation

Background noise and environmental factors significantly impact transcription quality. Future work could investigate noise-robust diarization and ASR models that leverage techniques such as noise suppression, acoustic scene analysis, [26] and adaptive noise models. Moreover, designing models that can automatically adjust to different environments would enhance the robustness of transcription in dynamic settings. [27]

5.2.5. Multi-Lingual and Code-Switching Capabilities

In many real-world scenarios, speakers may switch between languages or use phrases from different languages (code-switching). Future transcription models that can handle multi-lingual input and code-switching would expand the applicability of diarization and ASR systems. Research into language-agnostic speaker embeddings and training ASR models on multi-lingual datasets would be instrumental for these advancements.

5.2.6. Enhanced ASR Models with Contextual and Semantic Understanding

ASR models can benefit from better contextual and semantic understanding to reduce Word Error Rate (WER), particularly in multi-speaker environments. Incorporating sentence transformers and language models like BERT or GPT could improve the model's ability to understand the context, thus leading to more accurate transcriptions. Developing hybrid models

that combine ASR with natural language processing (NLP) capabilities may help in understanding nuanced language usage and correcting ASR errors based on context.

5.2.7. Exploring Multi-Channel and Spatial Audio Processing

For settings where multiple microphones or spatial audio recordings are used, integrating spatial information (such as sound source localization) could enhance diarization and speaker separation accuracy. Future work could explore spatial audio processing techniques to better segment speakers based on location, improving diarization performance [25] in settings like conference rooms, panel discussions, and surveillance.

5.2.8. Improvement of DER Metrics and Evaluation Methods

The current metrics, such as Diarization Error Rate (DER) and Word Error Rate (WER), do not always capture the full quality of multi-speaker transcription. Research into more comprehensive evaluation metrics that better reflect the quality of both diarization and ASR results would be valuable. Metrics that account for overlapping speech accuracy, speaker-wise segmentation, and error propagation between diarization and ASR could provide a more nuanced assessment of transcription quality.

5.2.9. Personalization and Adaptation for Speaker-Specific Characteristics

Personalization of transcription models based on specific speaker characteristics (such as dialect, speaking style, or vocal timbre) could further improve accuracy. Speaker-adaptive ASR and diarization models that learn speaker-specific nuances over time would be beneficial, especially for applications where speakers are known or frequently recurring (such as meetings or call centers).

5.2.10. Ethical Considerations and Privacy-Preserving Transcription

As multi-speaker transcription technologies become more powerful, privacy concerns and ethical considerations gain prominence. Future research could explore privacy-preserving diarization and ASR approaches that allow for speaker identification and transcription while safeguarding speaker anonymity and respecting privacy laws. Techniques such as federated learning and differential privacy may enable model improvements without compromising user data privacy.

5.2.11. User Feedback Integration for Continuous Model Improvement

Enabling models to learn and improve over time based on user feedback could refine transcription accuracy. For instance, users could correct misattributions or transcription errors, providing feedback that models could learn from in future sessions. Research into interactive or feedback-driven ASR and diarization systems could lead to models that continually improve and adapt based on specific use cases and user preferences.

Future work on multi-speaker transcription systems has the potential to address current limitations in diarization and ASR accuracy, processing speed, and adaptability to complex audio environments. By incorporating advanced modeling techniques, real-time capabilities, multi-lingual support, and ethical considerations, next-generation transcription systems could become more versatile, accurate, and widely applicable across various industries and use cases. These improvements would ultimately contribute to more reliable, context-aware, and user-centric multi-speaker transcription solutions.

Chapter 6
Bibliography

[1] J. W. Merrill, Automatic speech recognition, 1986

[2] Xugang Lu, Sheng Li, M. Fujimoto, Automatic Speech Recognition, 2019

[3] M. DíezL. BurgetFederico LandiniJ. Černocký, Analysis of Speaker Diarization Based on Bayesian HMM With Eigenvoice PriorsIEEE/ACM Transactions on Audio, Speech, and Language Processing 2020

[4] Jerry R. Hobbs & Ralph Grishman, The automatic transformational analysis of english sentences: An implementation, pp. 267-283, Dec 2010

[5] Yosuke Higuchi, M. Suzuki, Gakuto Kurata, Speaker Embeddings Incorporating Acoustic Conditions for Diarization 2020

[6] L. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, Feb 1989

[7] Yingjun Dong, Neil G. MacLaren, Yiding Cao, F. Yammarino, Shelley D. Dionne, M. Mumford, S. Connelly, Hiroki Sayama, G. Ruark, Speaker Diarization Using Stereo Audio Channels: Preliminary Study on Utterance Clustering, 2020

[8] Koichiro Yoshino, Shinsuke Mori, Tatsuya Kawahara, Language modeling for spoken dialogue system based on sentence transformation and filtering using predicate-argument structures, 2012

[9] Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, Zhong Meng, Ke Hu, A. Rosenberg, Rohit Prabhavalkar, Daniel S. Park, Parisa Haghani, Jason Riesa, Ginger Perng, H. Soltau, Trevor Strohman, B. Ramabhadran, Tara N. Sainath, P. Moreno, Chung-Cheng Chiu, J. Schalkwyk, Franccoise Beaufays, Yonghui Wu, Google USM: Scaling Automatic Speech Recognition Beyond 100 Languages, 2023

[10] Aonan ZhangQuan WangZhenyao ZhuJ. PaisleyChong Wang, Fully Supervised Speaker Diarization ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)2018

[11] Boakye, K., Trueba-Hornero, B., Vinyals, O., Friedland, G., 2008. Overlapped speech detection for improved speaker diarization in multiparty meetings, In: Proc. of ICASSP, pp. 4353–4356.

[12] Pyannote.Audio, Neural Building Blocks for Speaker Diarization, ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2020

[13] Delphine Charlet, Claude Barras and Jean-Sylvain Liénard, "Impact of overlapping speech detection on speaker diarization for broadcast news and debates", Proc. ICASSP 2013, May 2013.

[14] M. Siegler, U. Jain, B. Raj and R. Stern, “Automatic Segmentation, Classification, and Clustering of Broadcast News Audio,” Proc. of DARPA Speech Recognition Workshop, pp. 97-99, Chantilly, Virginia, Feb. 1997.

[15] Reynolds, D.A.: A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification. PhD thesis, Georgia Institute of Technology 1992

[16] S.N. Wrigley, G.J. Brown, V. Wan, and S. Renals, “Speech and crosstalk detection in multi-channel audio,” IEEE Trans. on Speech and Audio Processing, vol. 13, no. 1, pp. 84–91, 2005

[17] T.C.Rajapakse, Simple transformers, https://github.com/ThilinaRajapakse/ simpletransformers, 2019.

[18] A. Moi and N. Patry, Huggingface’s tokenizers, version 0.13.4, 2023.

[19] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Fried land, and O. Vinyals, “Speaker diarization: A review of recent research,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 2, pp. 356–370, 2012.

[20] K. Han and S. Narayanan, “A robust stopping criterion for ag glomerative hierarchical clustering in a speaker diarization sys tem,” in Interspeech 2010, September 26-30, Makuhari, Japan, 2010, pp. Interspeech–2010.

[21] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” IEEE Trans actions on Audio, Speech, and Language Processing, vol. 19, no. 4, pp. 788–798, 2011.

[22] P. Kenny, “Bayesian analysis of speaker diarization with eigen voice priors,” CRIM, Montreal, Technical Report, 2008.

[23] Y. Ephraim and D. Malah, “Speech enhancement using a minimum-mean square error short-time spectral amplitude esti mator,” IEEE Transactions on Acoustics, Speech, and Signal Pro cessing, vol. 32, no. 6, pp. 1109–1121, 1984.

[24] N. Ryant, M. Liberman, and J. Yuan, “Speech activity detection on youtube using deep neural networks.” in INTERSPEECH, 2013, pp. 728–731.

[25] S. Tranter, “Two-way cluster voting to improve speaker diarization performance,” in Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP’05). IEEE International Conference on, vol. 1. IEEE, 2005, pp. I–753.

[26] D. Moraru, S. Meignier, C. Fredouille, L. Besacier, and J. Bonas tre, “The ELISA consortium approaches in broadcast news speaker segmentation during the NIST 2003 rich transcription evaluation,” in Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP’04). IEEE International Conference on, vol. 1. IEEE, 2004, pp. I–373.

[27] A. Friedland, B. Vinyals, C. Huang, and D. Muller, “Fusing short term and long term features for improved speaker diarization,” in Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on. IEEE, 2009, pp. 4077–4080.