A project report on

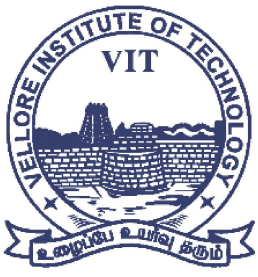
SINGER DIARIZATION IN MULTI SINGER AUDIO

Submitted in partial fulfillment for the award of the degree of

Bachelor of Technology in Computer Science and Engineering

By

ADVAIT DEOCHAKKE (20BCE1143)



VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

CHENNAI

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

April, 2024

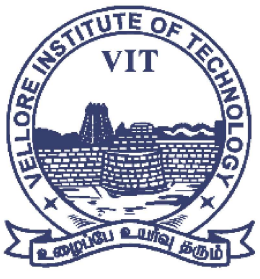
SINGER DIARIZATION IN MULTI SINGER AUDIO

Submitted in partial fulfillment for the award of the degree of

Bachelor of Technology in Computer Science and Engineering

By

ADVAIT DEOCHAKKE (20BCE1143)



VIT®

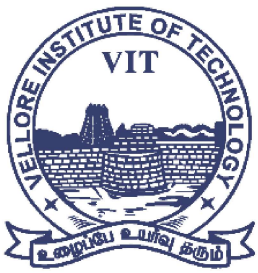
Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

CHENNAI

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

April, 2024



VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

CHENNAI

DECLARATION

I hereby declare that the thesis entitled "SINGER DIARIZATION IN MULTI SINGER AUDIO" submitted by me, for the award of the degree of Bachelor of Technology in Computer Science and Engineering, Vellore Institute of Technology, Chennai is a record of bonafide work carried out by me under the supervision of Dr. Priyadarshini J.

I further declare that the work reported in this thesis has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university.

Place: Chennai

Date: Signature of the Candidate

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

CHENNAI

School of Computer Science and Engineering

CERTIFICATE

This is to certify that the report entitled "SINGER DIARIZATION IN MULTI SPEAKER AUDIO" is prepared and submitted by Advait Deochakke (20BCE1143) to Vellore Institute of Technology, Chennai, in partial fulfillment of the requirement for the award of the degree of Bachelor of Technology in Computer Science and Engineering programme is a bonafide record carried out under my guidance. The project fulfills the requirements as per the regulations of this University and in my opinion meets the necessary standards for submission. The contents of this report have not been submitted and will not be submitted either in part or in full, for the award of any other degree or diploma and the same is certified.

Signature of the Guide:

Name: Dr. Priyadarshini J

Date:

Signature of the Internal Examiner
Name:
Date:

Signature of the External Examiner
Name:
Date:

Approved by the Head of Department,
B.Tech. CSE

Name: Dr. Nithyanandam P

Date:

ABSTRACT

The thesis investigates the challenging task of automatically identifying individual singers within music recordings. Inherently varying vocal characteristics, frequent singing overlaps, and the often-short snippets of singing all pose difficulties for singer identification. To address these hurdles, a multi-stage system is proposed that leverages machine learning and deep learning techniques.

The system employs meticulous preprocessing steps to prepare raw audio data for analysis. Subsequently, a pre-trained deep learning model isolates the vocals from the accompanying music, enabling focused analysis on the unique vocal characteristics of each singer. Feature extraction techniques then capture informative characteristics from the vocals, including Mel-Frequency Cepstral Coefficients and spectral features that provide a quantitative representation of the audio. To group similar vocal segments and perform singer diarization, spectral clustering is utilized followed by an iterative approach that leverages the silhouette score to determine the optimal number of singers. Recognizing the potential for bias in training data and the importance of privacy in singer identification systems, a comprehensive exploration of the ethical considerations surrounding this technology is incorporated.

Evaluation on 53 songs revealed the system successfully identified the number of singers in 31 recordings. Compared to baseline methods (DER of 51.14% and 99.31%), the proposed system achieved a significantly lower DER of 29.91%, demonstrating improved speaker identification accuracy. While adept at detecting singing activity (low miss rate of 1.64% and false alarm rate of 3.92%), it occasionally misidentified singers with similar voices (singer error rate of 25.99%).

Future research directions encompass several promising avenues. One approach involves incorporating deep learning models specifically designed for speaker recognition to enhance singer identification precision and recall. Additionally, integrating Voice Activity Detection (VAD) within the diarization process holds promise for improving speaker count estimation. Finally, the thesis explores potential real-world applications of singer identification, such as integration with music genre classification systems and the development of personalized music recommendation services.

By effectively addressing the challenges of singer identification and diarization, this research lays the groundwork for significant advancements in this field. The proposed system demonstrates a remarkable improvement over baseline methods, paving the way for more robust and accurate identification of singers in music recordings. Furthermore, the ethical considerations addressed in this thesis serve as a guiding principle for the responsible development and deployment of this technology in the future.

ACKNOWLEDGEMENT

It is my pleasure to express, with a deep sense of gratitude, my sincere appreciation to Dr. Priyadarshini J, Professor at the School of Computer Science and Engineering, Vellore Institute of Technology, Chennai. I am thankful for her constant guidance, continual encouragement, understanding, and, most importantly, for teaching me patience in my endeavors. My association with him/her extends beyond academics; it is a great opportunity for me to work with an intellectual and expert in the fields of Machine Learning, Deep Learning, and Information and Communication.

With gratitude, I extend my thanks to the visionary leader, Dr. G. Viswanathan, our Honorable Chancellor, Mr. Sankar Viswanathan, Dr. Sekar Viswanathan, Dr. G V Selvam, Vice Presidents, Dr. Sandhya Pentareddy, Executive Director, Ms. Kadhambari S. Viswanathan, Assistant Vice-President, Dr. V. S. Kanchana Bhaaskaran, Vice-Chancellor, Dr T. Thyagarajan Pro-Vice Chancellor, VIT Chennai and Dr. P. K. Manoharan, Additional Registrar, for providing an exceptional working environment and inspiring all of us during the course.

Special mention to Dr. Ganesan R, Dean, Dr. Parvathi R, Associate Dean Academics, Dr. Geetha S, Associate Dean Research, School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, for spending their valuable time and efforts in sharing their knowledge and helping us in every aspect.

In a jubilant state, I express my whole-hearted thanks to Dr. Nithyanandam P, Head of the Department, B.Tech. CSE, and the Project Coordinators for their valuable support and encouragement to take up and complete the thesis.

My sincere thanks to all the faculties and staff at Vellore Institute of Technology, Chennai, who helped me acquire the requisite knowledge. I would also like to thank my parents for their unwavering support. It is indeed a pleasure to thank my friends who encouraged me to take up and complete this task.

Place: Chennai
Date: **Advait Deochakke**

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

MIR - Music Information Retrieval

DER - Diarization Error Rate

MFCC - Mel-Frequency Cepstral Coefficient

LSTM - Long Short-Term Memory (network)

RNN - Recurrent Neural Network

SVM - Support Vector Machine

HMM - Hidden Markov Model

ANN - Artificial Neural Network

PLP - Perceptual Linear Predictive Coefficient

DIHARD - DIarization HArd (Challenge)

DJ - Disc Jockey

GMM - Gaussian Mixture Model

CNN - Convolutional Neural Network

CRNN - Convolutional Recurrent Neural Networks

UIS-RNN - Unbounded Interleaved-State Recurrent Neural Networks

JER - Jaccard Error Rate (Evaluation Metric)

WDER - Word-level Diarization Error Rate (Evaluation Metric)

ASR - Automatic Speech Recognition (Speaker Diarization)

BLSTM - Bidirectional Long Short-Term Memory (Singing Voice Detection)

SSL - Self-supervised Learning (Singing Voice Understanding)

BIC - Bayesian Information Criterion (Singer Diarization)

DCAP - Consolidated A Posteriori Decision (Singer Diarization)

STE- Short-Term Energy

STFT - Short-Time Fourier Transform

PCA - Principal Component Analysis

VAD - Voice Activity Detection

JFA - Joint Factor Analysis

Chapter 1

Introduction

1.1 INTRODUCTION

Musical compositions, like vibrant tapestries woven from diverse voices, showcase the multifaceted nature of artistic expression. Each melodic creation unfolds a rich soundscape where unique vocal contributions harmonize, crafting an immersive and captivating auditory experience. Within the realm of audio processing, the quest to decipher this intricate sonic fabric has led to the development of speaker diarization techniques. Initially conceived for speech analysis, specifically to identify and categorize speakers, these methods have undergone a remarkable evolution, finding transformative applications in the captivating domain of musical diarization [1].

This thesis delves specifically into the intricate challenge of singer diarization, aiming to accurately identify the number of individual singers present within a given musical recording. As technology and machine learning continue their rapid advancement, this research explores the potential of features like Mel-Frequency Cepstral Coefficients (MFCCs), timbre, and chroma scores, alongside sophisticated techniques like Long Short-Term Memory (LSTM) networks and spectral clustering [2]. The objective lies in unraveling the subtle nuances embedded within vocal elements of music recordings, thereby highlighting the significance of understanding individual vocal contributions in the realms of music production and recommendation systems. Notably, the study ventures into the domain of vocal-accompaniment separation, utilizing libraries like Librosa, TensorFlow, and Spleeter, a widely employed open-source system for such separations [3].

This exploration delves into the technical intricacies of singer diarization, meticulously examining the algorithms and methodologies employed to dissect and classify individual singers within the context of a musical ensemble. The convergence of technological advancements and the art of music not only empowers us with enhanced analytical capabilities but also broadens the horizons of creative expression. By meticulously analyzing the intricate interplay of voices, singer diarization emerges as a transformative tool, offering insights that resonate not only with music enthusiasts but also with professionals across various domains [4]. This research, therefore, marks a harmonious intersection of technology and artistry, fostering deeper understanding and appreciation for the multifaceted world of music.

1.2 OVERVIEW OF SPEAKER DIARIZATION

Diarization, a pivotal process in audio processing, revolves around segmenting and classifying speakers or sound sources within an audio recording. Initially designed for speech analysis, diarization has expanded into various applications within the realm of audio processing. Its primary objective is to unravel the temporal structure of an audio signal, acting as a precursor to tasks like speaker identification, content summarization, and sentiment analysis in speech recordings.

In recent years, the integration of advanced techniques, such as Long Short-Term Memory (LSTM) networks and neural networks, has brought a transformative change to speech diarization [2]. These technologies allow for a more nuanced understanding of speech patterns, enhancing the accuracy and efficiency of the diarization process.

Competitions like the DIHARD challenges have significantly influenced the development of speech diarization methodologies. These competitions provide a platform for researchers to benchmark their approaches, fostering innovation and driving advancements in the field. The challenges posed in DIHARD competitions mirror real-world scenarios, encouraging the creation of robust diarization solutions [5,6] .

Scoring speech diarization poses challenges due to the complexity of evaluating speaker boundaries. Traditional metrics like diarization error rates (DER) and Jaccard error rates have gained prominence over conventional machine learning metrics [7]. This shift is essential for accommodating the unique characteristics of speech diarization, where precise identification of speaker turns and boundaries is crucial. The intricacies of speech patterns, variations in tone, and overlapping speech segments make traditional metrics less suitable, necessitating a focus on metrics tailored to the specific requirements of speech diarization evaluation [8].

1.3 SINGER DIARIZATION COMPARED TO SPEAKER DIARIZATION

Singer diarization, an extension of the principles employed in speaker diarization, is a field of study that grapples with the complexities introduced by the multifaceted nature of musical compositions. While speaker diarization primarily focuses on identifying and clustering speakers based on speech patterns and linguistic nuances, singer diarization expands this paradigm to encompass the intricate world of music, where multiple vocalists, varying vocal styles, and harmonious interactions add layers of complexity [9].

The challenges associated with singer diarization are notably distinct from those faced in traditional speaker diarization. One key differentiator lies in the fluid nature of vocal

contributions within a musical piece [10]. Unlike speeches, which typically have clear turn-taking and defined speaker boundaries, music often features simultaneous vocalizations, overlapping harmonies, and dynamic shifts in vocal prominence. As a result, singer diarization requires a nuanced approach to accurately capture these intricacies and offer a comprehensive understanding of the vocal dynamics in a musical recording.

Moreover, the consideration of instrumental accompaniment further distinguishes singer diarization. In speech diarization, the primary challenge is often the identification of speakers against a background of silence or ambient noise. In music, the coexistence of vocals and instruments demands sophisticated techniques for vocal-accompaniment separation, adding an additional layer of complexity to achieve accurate diarization results [3, 11].

To address these challenges, previous research in singing voice detection has predominantly relied on machine learning techniques. Algorithms such as Support Vector Machines (SVMs), Hidden Markov Models (HMMs), Random Forests, and Artificial Neural Networks (ANNs) have been employed [12]. Feature extraction methods often involve combinations of Mel-Frequency Cepstral Coefficients (MFCCs), Perceptual Linear Predictive Coefficients (PLPs), and Log Frequency Power Coefficients [13]. These methods have significantly contributed to the development of effective singer diarization systems.

Despite the advancements made, the complexities associated with this field, characterized by the presence of singers and in some case multiple singing voices, have led researchers to explore more advanced areas of study. Music structure analysis, extraction of specific singers' voices, and music information retrieval based on singer diarization have become focal points of investigation [14]. Ongoing research emphasizes improvements in spectral clustering algorithms, singer representations, and target-singer voice activity detection to enhance the accuracy and effectiveness of singer diarization systems [15].

1.4 CHALLENGES IN VOCALS IDENTIFICATION AND SEPARATION

Vocal identification and separation in music recordings come with their own spectrum of challenges. The varied nature of musical compositions introduces nuances in vocal styles, making it challenging to accurately distinguish individual singers. From timbral intricacies to pitch variations, capturing the unique qualities of each vocalist requires a sophisticated approach. A notable challenge arises when dealing with overlapped segments, where multiple singers engage in unison singing, presenting difficulties in diarization and the need for specialized methods to distinguish individual voices within overlapping harmonies [16, 17].

The interplay between vocal elements and instrumental accompaniments further complicates the isolation process. Harmonically rich compositions often feature overlapping vocal harmonies and intricate instrumental arrangements, creating a complex auditory landscape. Traditional methods may struggle to deconstruct this complexity, especially when considering

the impact of background instruments on diarization accuracy. To mitigate this, solutions such as reducing background music or implementing sound source separation techniques become necessary to enhance diarization accuracy [1, 3].

Adding to the challenges is the variability in performance styles across genres and individual artists. Distinct vocal nuances in classical, jazz, and pop music require an adaptable diarization model capable of discerning and accommodating diverse manifestations of vocal expression. The acoustic differences between singing and speech voices further complicate diarization, as singing voices typically exhibit a wider range of fundamental frequency (F0) and longer durations of phonemes compared to speech voices. Handling simultaneous singers, a common occurrence in singing, presents its own set of challenges, requiring specialized techniques for accurate identification [18].

An additional hurdle lies in acquiring singer information from short segments, a frequent occurrence in musical compositions. Spectral features are affected by differences in F0 range and phoneme duration between singing and speech voices, demanding meticulous analysis and feature extraction methods to ensure accurate diarization even in short audio segments [9, 10]. Navigating these challenges calls for a comprehensive approach that integrates musical theory, audio signal processing, and specialized diarization techniques, providing a nuanced understanding for accurate vocal identification and separation within the intricate tapestry of musical recordings.

1.5 PROJECT STATEMENT

This project aims to demonstrate the feasibility and effectiveness of a machine learning and deep learning-based approach for automatic singer identification in music recordings. By evaluating the system against a baseline and analyzing its performance metrics, the project seeks to contribute valuable insights into the potential of this technology for accurate singer diarization. The findings will serve as a foundation for future research and development efforts towards even more robust and comprehensive singer identification systems.

1.6 OBJECTIVES

Automatic singer identification and diarization is a challenging task in Music Information Retrieval (MIR) due to inherent difficulties presented by the variability in vocal characteristics, frequent overlapping vocals, and brevity of singing segments. This research project proposes a novel, multi-stage system to address these challenges and achieve superior performance in singer identification and diarization tasks.

- Variability in Vocal Characteristics: Singers possess unique vocal attributes that can differ significantly.

- Frequent Overlapping Vocals: In many recordings, multiple singers may be singing simultaneously, making it difficult to isolate individual voices.
- Brevity of Singing Segments: Music recordings often feature short singing segments, further complicating singer identification.

To effectively address these hurdles, the project is guided by a triad of meticulously designed objectives that target distinct aspects of singer identification and diarization:

- Objective 1: Achieve Superior Singer Identification Accuracy

  The paramount objective revolves around developing a system with exceptional precision in distinguishing between individual singers within a music recording. The success of this objective hinges upon demonstrably achieving a lower Diarization Error Rate (DER) compared to a baseline system. A lower DER signifies a measurable reduction in errors associated with misidentified, missed, or incorrectly assigned singing segments. The evaluation will involve a rigorous comparison between the proposed system's identifications and a pre-established ground truth for a diverse dataset encompassing a broad range of music genres.

- Objective 2: Optimize Speaker Count Estimation

  An equally crucial system component will be specifically designed to determine the exact number of singers present in a recording with optimal accuracy. This objective goes beyond simply identifying singers; it necessitates pinpointing the precise number of unique voices contributing to the recording. The evaluation will be conducted by meticulously comparing the system's predicted singer count against the established ground truth for the music dataset.

- Objective 3: Integrate Ethical Considerations into System Design and Deployment

  A responsible and ethical approach to singer identification technology is paramount. This objective entails a thorough investigation into the potential ethical ramifications surrounding the system. It will involve identifying and mitigating potential biases within training data that could lead to inaccurate identifications, particularly for underrepresented demographics or musical styles. Additionally, the objective will address privacy concerns associated with singer identification in recordings, ensuring user privacy is protected throughout the identification process.

1.7 SCOPE OF PROJECT

The project's scope is defined as follows:

Inclusions -

- Data Acquisition:
    - Gather a diverse dataset of music recordings encompassing various genres, vocal styles, and numbers of singers.
    - Preprocess the audio data to ensure consistent formatting and quality for analysis.
- Vocal Separation:
    - Integrate a pre-trained deep learning model (e.g., Spleeter) to isolate vocal tracks from the accompanying music.
- Feature Extraction:
    - Employ established techniques to extract informative features from the vocal segments, including Mel-Frequency Cepstral Coefficients (MFCCs), spectral features, and other relevant characteristics.
- Singer Count Prediction and Singer Diarization:
    - Deploy a spectral clustering and inflection point based approach to accurately predict the number of singers present in songs
    - Implement a speaker diarization approach using various clustering methods to group similar vocal segments.
    - Develop an iterative process to determine the optimal number of singers based on metrics like the silhouette score.
- Evaluation:
    - Benchmark the system's performance against a baseline approach (e.g., random segmentation) using Diarization Error Rate (DER) as the primary metric.
    - Analyze the system's strengths and weaknesses, including miss rate, false alarm rate, and singer error rate.

Exclusions -

- Real-time Singer Identification:
    - The initial focus will be on offline analysis of pre-recorded music. Real-time processing will be considered for future exploration.
- Music Genre Classification:
    - While the system might provide insights into genre based on vocal styles, full-fledged genre classification is beyond the current scope.
- Music Recommendation Systems:
    - Integration with music recommendation systems will be explored in future work based on the core singer identification technology.

Additional Considerations -

- Ethical Considerations:
    - Address potential biases in training data and ensure fair identification across diverse singers and genres.
    - Acknowledge privacy concerns and explore methods for anonymized singer identification when necessary.
- Scalability and Efficiency:
    - Design the system with scalability in mind to handle large music collections.
    - Explore techniques for optimizing computational efficiency for processing power and resource limitations.
- Explainability and Interpretability:
    - When using deep learning models like Spleeter, strive for transparency in their decision-making processes to aid in debugging and improvement.

Chapter 2

Background

2.1 INTRODUCTION

Singer diarization, the process of separating and identifying individual singers within a music recording, plays a crucial role in various Music Information Retrieval (MIR) applications. Unlike traditional singer identification, which focuses on recognizing singers, diarization delves deeper by segmenting and labeling each singer's vocal contribution throughout the audio. This capability holds immense potential for enhancing user experience in music streaming services, facilitating music analysis and transcription for musicians and researchers, and streamlining content creation by simplifying copyright attribution [1, 12].

While significant advancements have been made in singer identification and related MIR tasks, existing methods often face challenges due to several factors. Distinguishing singers with distinct vocal styles and pitches can be difficult, especially when dealing with genres like choral music or songs featuring guest vocalists. Separating vocals from the instrumental accompaniment remains an ongoing challenge, particularly in complex musical arrangements [3, 19]. Additionally, handling overlapping vocals and short audio segments further complicates the identification process.

Despite these challenges, the potential benefits of singer diarization motivate continuous research and development in this field. Existing techniques for singer identification rely on various approaches, including template matching, statistical modeling, and machine learning methods. However, these methods often struggle with the aforementioned limitations, particularly when dealing with diverse musical styles and complex audio compositions [10]. Therefore, there is a critical need for novel approaches that can effectively address these limitations and improve the accuracy and robustness of singer diarization systems.

The background chapter aims to provide a comprehensive understanding of the field and pave the way for introducing the proposed approach to singer diarization. The text will delve into the theoretical foundations of singer diarization, exploring relevant signal processing techniques and the underlying principles of various existing methods. The limitations of these methods will be discussed, highlighting the specific challenges and gaps in current research. Finally, an overview of the datasets used for developing and evaluating singer diarization systems will be provided, setting the stage for introducing the proposed approach and its potential contributions to this evolving field.

2.2 MOTIVATION & APPLICATIONS

Singer diarization surpasses conventional methods of singer identification, which can recognize the singer but struggle to unravel the intricate interplay of voices in a song. Diarization involves meticulous separation and labeling of each vocal thread, revealing the dynamics of melodies and harmonies with precision. However, persistent challenges include identifying unique vocal styles, managing overlapping vocals, and discerning individual voices in short audio segments. These hurdles underscore the critical need for advanced diarization techniques to fully unleash the potential of this technology.

The motivations for advancing singer diarization extend beyond improving user experience in music streaming services. Envision a future where music becomes a deeply engaging and personalized discovery journey. Diarization offers the key to this future by accurately separating individual voices, enabling a profound analysis of a song's composition, and not just its melody but also the distinct contributions of each singer. This opens avenues for applications that can revolutionize music interaction.

In music recommendation systems, diarization can serve as a potent tool. For example, a recommendation engine can factor in not just genres and listening history but also specific singers or vocal styles you appreciate. Analyzing the vocal makeup of songs liked by a user can identify preferences for particular singer's voices or specific vocal harmonies, leading to more personalized recommendations [20]. Similarly, content retrieval systems could be transformed, allowing searches based on specific vocal combinations or a singer's contribution within a song, enabling deeper exploration of musical interests.

The impact of singer diarization extends to karaoke, a beloved activity for music enthusiasts. Setting up karaoke often involves laborious manual preparation of backing tracks that remove lead vocals while preserving instruments. Singer diarization offers an innovative solution by isolating individual vocal tracks, automating the creation of high-quality backing tracks without the lead singer. This reduces preparation time and effort, ensuring consistent quality and enhancing the karaoke experience for all [21].

Automatic music transcription, crucial for musicians and educators, can benefit from diarization advancements. Current methods struggle with complex arrangements and overlapping vocals, hindering accurate transcriptions. Diarization, by separating voices, enables more precise transcriptions, aiding analysis of melodies, harmonies, and vocal techniques. Educators can create interactive learning experiences using such transcriptions [22].

Beyond transcription, diarization unlocks possibilities for music remixing and mashups. DJs and producers gain the ability to isolate and manipulate individual vocal tracks, facilitating creative remixes and innovative mashups. This fosters artistic expression and personalized music creation.

A captivating application of diarization lies in a cappella music, where analyzing individual contributions within complex pieces is challenging. Diarization separates vocal tracks, providing valuable insights for singers and coaches to improve technique, blend, and overall performance quality.

In conclusion, addressing current limitations in singer identification through advanced diarization techniques is crucial. This unlocks a more enriching, insightful, and interactive music experience, from personalized recommendations to enhanced karaoke, groundbreaking music creation, and a deeper understanding of vocal artistry. Singer diarization promises to revolutionize music interaction, turning it into a vibrant tapestry of discovery, engagement, and artistic expression.

2.3 HISTORICAL CONTEXT

Speaker and singer diarization, while distinct fields with unique application areas, share a remarkable intertwined history. Both emerged in the early stages (1990s) within the broader field of speech processing, initially serving as auxiliary tasks for automatic speech recognition. In this formative period, both fields relied on similar fundamental approaches: segmenting audio based on features like pitch and Mel-Frequency Cepstral Coefficients (MFCCs) followed by clustering these segments using statistical models like Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs). [23, 24, 25]

However, as the maturation and expansion phase (2000s) unfolded, their specific applications diverged. Speaker diarization found broader utility in tasks like audio retrieval and meeting summarization, while singer diarization faced distinct challenges arising from the inherent musicality of singing voices. Pitch variations, vibrato, and stylistic choices presented difficulties in distinguishing singers solely based on features used in speaker diarization [26, 27].

Researchers addressed these challenges through various avenues:

- Domain-specific feature engineering: Both fields adopted domain-specific features to enhance performance. Speaker diarization utilized speaker-specific features like iVectors and Joint Factor Analysis (JFA), while singer diarization leveraged chroma features representing the tonal content of the audio [28].

- Knowledge-based approaches: Both fields explored the potential of leveraging additional knowledge to aid diarization. Speaker diarization sometimes incorporated information like transcripts, while singer diarization utilized knowledge of lyrics and melodies to guide singer separation [29].

- Template-based methods: Both fields experimented with matching segments to pre-recorded templates. In speaker diarization, these templates could be voice samples, while in singer diarization, they could be recordings of specific singers [30].

The arrival of deep learning in the 2010s marked a significant turning point for both fields. Convolutional Neural Networks (CNNs) emerged as a powerful tool for learning complex representations from audio data, capturing both domain-specific and speaker/singer-specific characteristics [31, 32]. Both speaker and singer diarization benefited from advancements in deep learning:

- Speaker/singer embedding learning: Techniques like deep metric learning helped train models to efficiently separate speakers/singers by embedding their voices in a low-dimensional space [33].

- End-to-end systems: Deep learning models were designed to handle the entire diarization pipeline, achieving state-of-the-art performance in both speaker and singer diarization [34].
  .

While their paths diverge into specific applications in the present day, the evolution of speaker and singer diarization underscores a remarkable synergy. Both fields have profited from shared foundational techniques and advancements in machine learning, paving the way for a future filled with even more sophisticated and impactful applications.

2.4 EXISTING TECHNIQUES AND LIMITATIONS

2.4.1 TECHNIQUES IN SINGER DIARIZATION

Singer diarization presents unique challenges compared to its speaker counterpart, necessitating the development of specialized techniques. One crucial area of focus is the incorporation of music-informed features. These features aim to capture characteristics specific to singing voices, such as the novel Cosacorr score for enhanced overlap detection in unison singing scenarios.

Deep learning architectures have also played a significant role in advancing singer diarization. Existing approaches leverage Convolutional Recurrent Neural Networks (CRNNs) and attention mechanisms to effectively handle both temporal and spectral information within audio data [35]. However, there remains potential for further exploration of architectures specifically tailored for singer diarization, capitalizing on their ability to learn from low-level features and exploit temporal context within the audio.

Another key technique involves singer embedding learning. This technique utilizes methods like ArcFace to learn discriminative representations of singers from short segments of their isolated voices [9]. These embeddings then facilitate efficient separation of singers within a

musical mixture. However, further research is necessary to improve the quality and effectiveness of singer representations, leading to more accurate diarization results.

2.4.2 LIMITATIONS AND FUTURE DIRECTIONS

Despite these advancements, singer diarization still faces several limitations. Musical complexity, characterized by heavy instrumentation, backing vocals, and variations across different genres, poses a significant challenge. To address this, incorporating music genre information into the diarization model and developing genre-specific techniques are promising avenues for future research.

Another major limitation is data scarcity. The availability of labeled singer data is crucial for training and evaluating diarization systems. Existing methods like data augmentation and semi-supervised learning attempt to address this limitation. Additionally, exploring transfer learning from related tasks such as speaker diarization or music genre classification could prove beneficial in leveraging knowledge from more abundant datasets.

The limitations of current evaluation metrics also deserve attention. While Diarization Error Rate (DER) remains a common metric, it has limitations in comprehensively capturing the nuances of singer diarization performance. Exploring perceptual evaluation metrics that incorporate human judgment can provide a more holistic understanding of system effectiveness [37].

While fully supervised learning approaches using Unbounded Interleaved-State Recurrent Neural Networks (UIS-RNNs) demonstrate promising results, their applicability in singer diarization is limited due to the challenges of acquiring large labeled datasets [36]. However, this highlights the potential of supervised learning techniques once sufficient labeled data becomes available.

Finally, it is crucial to consider fairness in singer diarization systems. Biases arising from factors like gender, vocal characteristics, or musical genre must be identified and mitigated to ensure fair and unbiased performance across diverse singers and musical styles [8].

By integrating speech separation techniques, singer diarization can be further enhanced, particularly in complex musical scenarios with overlapping singing or background music. Developing robust separation techniques specifically tailored for the challenges of singer diarization is essential for achieving accurate separation and diarization in such complex audio environments.

To advance singer diarization techniques and overcome challenges like musical complexity and data scarcity, a thorough grasp of current research endeavors is essential, despite the significant strides already achieved in this field. The following chapter delves into a critical review of the relevant literature surrounding singer diarization, exploring the theoretical foundations, methodological approaches, and reported results within this domain.

Chapter 3

Literature Review

This section reviews the existing literature on singing voice detection, focusing on the challenges associated with differentiating singing voice from other audio components and the various techniques employed in recent research. The reviewed studies will be categorized based on their methodological approaches, highlighting the strengths and limitations of each method. Additionally, this review will identify potential research gaps and areas for further investigation, informing the methodology and direction of the present study.

Park, Tae Jin, et al. (2022) presented a comprehensive review of speaker diarization, focusing on historical developments and recent advances, particularly within the realm of deep learning. Speaker diarization, the process of labeling audio or video recordings with speaker identities, has evolved from its early applications in speech recognition to becoming a standalone tool with diverse applications in media, conferences, social media, legal proceedings, and business meetings [1].

The historical development section emphasized early speaker diarization applications in the 1990s, primarily for Automatic Speech Recognition (ASR), with approaches like generalized likelihood ratio and Bayesian information criterion laying foundational groundwork. The introduction of the i-vector representation marked success in speaker recognition and diarization. The paper discussed the transformative impact of deep learning in the 2010s, introducing approaches like d-vectors and x-vectors that outperformed traditional methods in terms of performance, training efficiency, and robustness.

The motivation for the review stemmed from the need to consolidate advancements since the publication of earlier overview papers, offering a contemporary perspective on the field. The authors categorized diarization technologies based on training objectives and optimization strategies, providing a clear taxonomy for readers.

The paper also delved into evaluation metrics, emphasizing Diarization Error Rate (DER), Jaccard Error Rate (JER), and Word-level Diarization Error Rate (WDER), crucial for assessing system accuracy [1, 38]. The organization included detailed sections on traditional diarization systems, advancements leveraging deep neural networks, and the integration of speaker diarization with automatic speech recognition. Additionally, the authors discussed challenges, corpora, and real-world applications, providing a well-rounded overview of the current state of speaker diarization research. This literature review served as a valuable resource for researchers and practitioners, consolidating recent developments and paving the way for future advancements.

Chung, Joon Son, et al.'s (2020) research paper, "Spot the conversation: speaker diarisation in the wild," explored speaker diarisation under uncontrolled conditions, focusing on analysing videos collected 'in the wild.' Addressing challenges posed by unconstrained settings, the paper introduced an automatic audio-visual diarisation method tailored for YouTube videos. This method combined active speaker detection through audiovisual techniques with speaker verification using self-enrolled speaker models. The researchers seamlessly integrated this methodology into a semi-automatic pipeline for dataset creation, reducing the laborious task of annotating videos with diarisation labels. An outcome of this endeavour was VoxConverse, an extensive diarisation dataset from 'in the wild' videos, characterised by overlapping speech, a diverse speaker pool, and complex background conditions [30].

The research stemmed from the limitations of diarisation systems in constrained domains compared to varied conditions in online videos. Challenges included the lack of fixed domains, numerous speakers, short rapid exchanges, and challenging background conditions, making manual annotation difficult. The authors proposed a scalable audio-visual method for speaker diarisation in web videos, leveraging active speaker detection, face recognition, speech enhancement, and audio-only speaker recognition. The automatic pipeline for dataset curation involved video collection, shot detection, face detection and tracking, face-track clustering, active speaker detection, and off-screen speech labelling. The paper discussed challenges in manual verification and annotation guidelines, emphasising the importance of a large-scale diarisation dataset 'in the wild' to foster new techniques.

Experiments compared the proposed audio-visual method to an audio-only baseline, specifically the DIHARD 2019 baseline, showing substantial performance gains. Ablations demonstrated the effectiveness of using two active speaker detection methods. The VoxConverse dataset, consisting of challenging multi-speaker videos, was described, including statistics, video characteristics, and the dataset collection process. The paper concluded by highlighting the scalability and effectiveness of their approach, with plans to release VoxConverse to the research community, fostering further advancements in speaker diarisation for 'in the wild' videos.

The research paper by Ryant et al. (2020) meticulously analyzed the DIHARD III challenge, which aimed at enhancing the robustness of speaker diarization systems across diverse conditions, including various recording environments, noise levels, and conversational domains. The evaluation spanned two speech activity conditions and encompassed 11 domains, such as read audio-books, meeting speech, clinical interviews, web videos, and conversational telephone speech. The paper traced the historical development of diarization systems, emphasizing the significance of the DIHARD challenges as benchmarks for systematic evaluation. It detailed the evolution from DIHARD I and II to the present challenge, DIHARD III, offering insights into the task, metrics, datasets, and baseline systems utilized. The comprehensive analysis of challenge outcomes acknowledged progress in speaker diarization but underscored ongoing challenges, encouraging further research to achieve truly robust diarization systems [5, 39, 40].

In parallel, Watanabe et al. (2020) introduced the CHiME-6 challenge, which addressed the complexities of distant multi-microphone conversational speech diarization and recognition in everyday home environments. This challenge marked a departure from segmented to unsegmented multispeaker speech recognition scenarios. The authors contextualized CHiME-6 within the broader landscape of Automatic Speech Recognition (ASR), recognizing notable improvements attributed to advancements in speech processing, audio enhancement, and machine learning. The challenge's unique aspects included scenarios of dinner parties in real homes, capturing natural conversational speech with realistic room acoustics, background noises, and unscripted interactions. The introduction of two tracks, particularly Track 2 focusing on unsegmented multispeaker speech recognition scenarios, with open-source baselines, emphasized the challenges associated with speaker diarization. The technical details of the challenge, recording setup, and array synchronization process were detailed, along with the presentation of results showcasing the effectiveness of CHiME-6 baseline ASR system in Track 1. The paper positioned CHiME-6 as a pivotal initiative in advancing distant multispeaker speech recognition research, providing a realistic and challenging dataset. The inclusion of open-source tools encouraged collaboration and further exploration of unsegmented multispeaker scenarios, pushing the boundaries of current ASR capabilities [6].

One crucial aspect of singer diarization involved effectively distinguishing singing voice segments from other audio components. In this regard, Leglaive et al. (2015) presented a novel approach that utilized deep learning architectures for singing voice detection. Their method leveraged a Bidirectional Long Short-Term Memory (BLSTM) Recurrent Neural Network (RNN), which incorporated past and future temporal context within the network during feature extraction. This contrasted with traditional methods that relied on separate models for frame classification and temporal smoothing. Additionally, the BLSTM-RNN utilized low-level features, highlighting its ability to perform well without the need for complex feature engineering. This work demonstrated the effectiveness of deep BLSTM-RNNs in low-level feature extraction and sequence classification tasks related to singing voice detection. However, the generalizability of this approach to different music genres and the impact of varying network architectures required further investigation [41].

Suda et al. (2022) addressed a crucial gap in singer identification (diarization) in polyphonic music with unison singing by introducing a novel framework. Existing techniques had struggled with the overlap of multiple singers, hindering accurate identification [9]. Their framework introduced the Cosacorr score to enhance overlap detection and employed a specialized diarization framework tailored to singing voices. Additionally, ArcFace facilitated the extraction of discriminative singer representations.

The study not only advanced technically but also expanded the understanding of singing-specific features and their role in handling complexities of polyphonic music. This understanding has potential applications in music structure analysis, isolating specific singers, and singer-based information retrieval.

While acknowledging limitations, the authors suggested areas for future improvement, such as spectral clustering algorithms, singer representations, and target-singer voice activity detection. Overall, their work significantly contributed to singer diarization in polyphonic music with unison singing, paving the way for further advancements and broader applications in music information processing.

Yamamoto (2023) investigated the use of pre-trained Self-Supervised Learning (SSL) models for automatic singing voice understanding tasks such as singer identification, transcription, and technique classification, which are crucial for applications like music discovery, education, and musicology [42]. Despite the effectiveness of data-driven approaches, especially deep learning, in handling the diverse and noisy nature of singing voices, the challenge of limited labeled data persists. This study explored how SSL models, trained on vast amounts of unlabeled data, could potentially achieve comparable performance to supervised learning with scarce labeled data.

The research focused on three SSL models (Wav2Vec2.0, WavLM, MapMusic2Vec) and evaluated their performance across the identified tasks. The results demonstrated that SSL models achieved comparable or even superior performance compared to state-of-the-art methods. Additionally, the study analyzed model behavior across different layers, providing insights into their potential for singing voice recognition.

Overall, Yamamoto (2023) contributed by exploring the potential of SSL models in overcoming data scarcity and advancing automatic singing voice understanding tasks. This paves the way for further research and potentially broader application in music information processing.

Thlithi et al. (2015) explored the application of singer diarization, a technique for identifying and separating singers in recordings, to the challenging domain of ethnomusicological recordings. These recordings, known for intricate vocal overlaps, diverse musical styles, and varying sound quality, present unique complexities [10].

The authors proposed a two-stage approach to address these challenges. In the first stage, they segmented the audio recording into sections with similar acoustic properties, leveraging the Bayesian Information Criterion (BIC). However, they noted that a single value for the BIC penalty parameter might not be optimal for all recordings. To overcome this limitation, Thlithi et al. (2015) introduced the Consolidated A Posteriori Decision (DCAP) method. This method combined segmentations obtained using different penalty factors, ultimately selecting boundaries consistently identified across multiple runs.

Moving to the second stage, the authors clustered the segmented audio data using agglomerative clustering with Mel-Frequency Cepstral Coefficients (MFCCs) as features. This clustering aimed to group segments likely sung by the same set of singers.

The system's performance was evaluated on the "DIADEMS" corpus, a collection of ethnomusicological recordings from sub-Saharan Africa. While the system achieved a promising Diarization Error Rate (DER) on the development set, the performance dropped significantly on the evaluation set. This highlighted the substantial challenges posed by the inherent complexities of ethnomusicological recordings.

The authors acknowledged these limitations and proposed avenues for future research. These included exploring alternative clustering techniques, investigating methods for handling specific challenges like superimposed singing and background noise, and adapting the system to handle the diverse musical styles and audio qualities present in ethnomusicological recordings.

In conclusion, the extensive literature review on singing voice detection and speaker diarization has shed light on various methodological approaches, historical developments, recent advancements, and challenges within these domains. The reviewed studies provided insights into the evolution of speaker diarization from its origins in speech recognition to its current applications across diverse fields such as media, conferences, social media, legal proceedings, and business meetings.

The literature review categorized methodologies based on training objectives, optimization strategies, and evaluation metrics, emphasizing the importance of Diarization Error Rate (DER), Jaccard Error Rate (JER), and Word-level Diarization Error Rate (WDER) in assessing system accuracy. Furthermore, the review identified research gaps and areas for further investigation, underscoring the need for scalable and effective solutions in speaker diarization, especially under uncontrolled conditions 'in the wild' and in challenging domains like ethnomusicological recordings.

The synthesis of findings from various studies not only provided a comprehensive understanding of the current state of research but also highlighted the ongoing challenges and opportunities for advancements in singing voice detection, singer diarization, and speaker diarization across diverse contexts. This literature review serves as a valuable resource for researchers, practitioners, and stakeholders in the field, fostering collaboration and innovation to address the complexities and enhance the accuracy of audio analysis methodologies.

Chapter 4

Methodology

This research aims to develop a system for identifying and labelling individual singers (diarization) within music recordings, by predicting the number of singers first. The primary challenge lies in distinguishing between singers with different vocal characteristics, separating vocals from the music accompaniment, and handling situations with overlapping vocals or short audio segments. To address these complexities, this research proposes a multi-stage approach that leverages a combination of techniques to not only segregate non-vocal audio and vocals in song by segmenting, but also using various features to cluster and diarize.

4.1 PREPROCESSING AND FEATURE EXTRACTION

The initial stage of the singer identification and labeling system focuses on preparing the audio data for subsequent analysis. This stage encompasses two critical steps: preprocessing and vocal separation. Preprocessing techniques aim to transform the raw audio signal into a format suitable for feature extraction and analysis, while vocal separation isolates the vocal tracks from the instrumental accompaniment, allowing the focus to be on the singers' unique vocal characteristics.

4.1.1 SYSTEM ENVIRONMENT AND LIBRARY SETUP

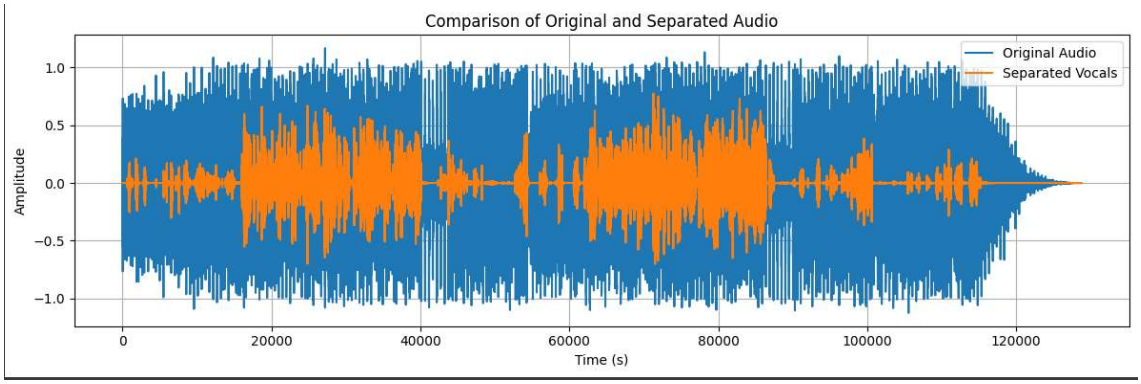
Before embarking on audio processing tasks, establishing a robust computational environment by installing essential libraries is a must. These libraries provide the necessary functionalities for audio manipulation, analysis, and feature extraction. Key libraries employed in this stage include:

- Spleeter: This pre-trained model serves as a powerful tool for source separation, specifically designed for isolating vocals from music recordings [3].
- Librosa: Offering a comprehensive suite of audio processing functionalities, Librosa empowers us with tools for reading audio files, performing feature extraction, and spectrogram visualization [43].
- pyAudioAnalysis: This library equips us with functionalities for silence removal and audio segmentation, enabling the extraction of relevant segments and elimination of non-vocal portions [44].

4.1.2 VOCAL SEPARATION USING SPLEETER:

Vocal separation plays a crucial role in this analysis, as it allows for the isolation of the singers' voices from the complex mix of instruments and background noise present in the audio recordings. To achieve this, the power of Spleeter is harnessed, a cutting-edge deep learning library specifically designed for music source separation.

Fig 1: Overlap of Original audio (Blue) and audio with Separated vocals (Orange) to showcase change in waveforms



Spleeter offers a variety of pre-trained models, each optimized for different separation tasks. In this case, the 2-stem model is utilized, which excels at separating vocals from the remaining audio components (accompaniment) within a single audio source. This model employs deep learning algorithms to meticulously analyze the audio signal and effectively extract the desired vocal tracks. There are 3, 4, and even 5-stem models available, but their use cases are more particular for separating different instruments in the backing track..

By isolating the vocals using Spleeter's 2-stem model, several advantages are seen:

- Focus on Singers' Unique Characteristics: Removing the instrumental accompaniment allows us to concentrate better on the singers' vocal qualities, such as their timbre, pitch, and singing style. This focused analysis is crucial for accurately identifying and labelling singers based on their unique vocal signatures [45].
- Enhanced Accuracy: Isolating the vocals minimizes the influence of background noise and instruments, leading to a cleaner and more accurate representation of the singers' voices. This, in turn, improves the overall accuracy of the identification and labeling process.

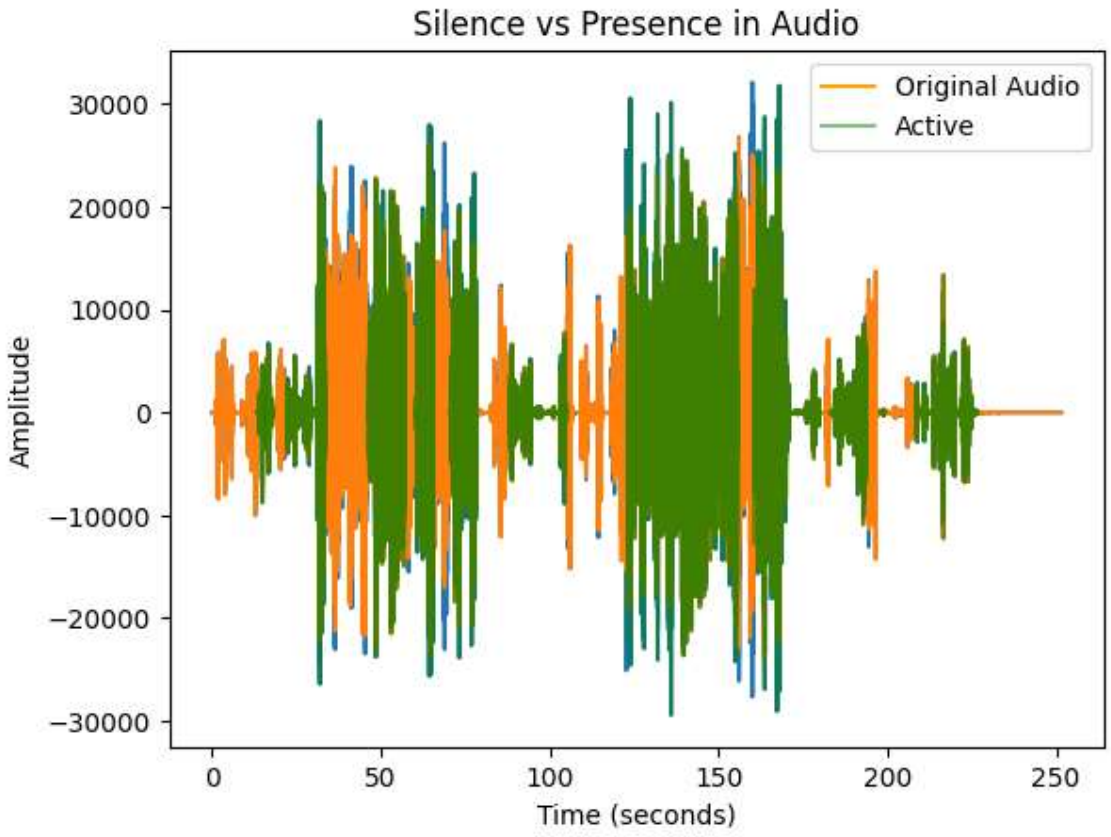
4.1.3 SILENCE REMOVAL AND SEGMENTATION

In the final step of preparing vocal recordings for analysis, silent segments are eliminated to prevent them from negatively impacting the accuracy of subsequent steps. Libraries like pyAudioAnalysis offer functionalities to achieve this through a two-pronged approach: silence detection and windowing for precise removal.

Silence removal algorithms typically rely on analyzing the audio waveform's energy level. During silent periods, the audio energy falls below a predetermined threshold. The algorithm detects these low-energy segments and removes them from the audio data.. Silence detection involves analyzing the audio signal using algorithms like Short-Term Energy (STE) estimation and Spectral Flux. STE calculates the average energy within a short window, with low values indicating silence. Spectral Flux measures the change in the frequency spectrum, and sudden drops often correspond to silence. By setting a threshold based on the desired noise tolerance, any audio segment with STE or spectral flux values below the threshold is classified as silence [46].

Windowing for precise removal further refines the process. The audio is divided into smaller, overlapping windows, allowing for individual analysis of each segment using the silence detection algorithms mentioned earlier. This ensures that only windows predominantly classified as silent based on the threshold are removed, preserving even short non-silent segments.

Fig 2: Segments identified as containing speech are highlighted in green, overlapped with the base audio in orange



This windowing approach offers several advantages after silence removal. Firstly, dividing the audio into smaller windows facilitates the identification of potential speaker changes within a short timeframe, which might be missed in larger segments. This is crucial for

applications like speaker diarization, where identifying different speakers present in the recording is essential. Secondly, smaller audio segments are computationally more efficient to analyze for extracting features like Mel-Frequency Cepstral Coefficients (MFCCs) for speaker identification and speech recognition, pitch for speaker identification and emotion recognition, and timbre for speaker identification and music genre classification.

By effectively removing silent segments while preserving valuable speech information through these techniques, the accuracy and efficiency of subsequent analysis tasks can see a significant improvement.

4.2 SINGER DIARIZATION

Speaker diarization relies on identifying the unique vocal characteristics of individuals present in an audio recording. To achieve this, informative features that capture these characteristics are extracted.

The chosen features include:

- Mel-Frequency Cepstral Coefficients (MFCCs): These features mimic how humans perceive sound frequencies, capturing the spectral envelope of the audio. This information provides insights into fundamental vocal qualities like timbre (vocal quality) and pitch [47, 48].

- Chroma features: These features represent the short-term harmonic content of the audio signal by analyzing its power spectrum. This information can be crucial for differentiating between speakers, especially when dealing with singers who have distinct vocal ranges or singing styles [49].

- Delta and Delta-Delta of MFCCs: Complementing the MFCCs, delta and delta-delta coefficients convey the rate of change of the spectral features over time. These dynamic features play a pivotal role in capturing variations in vocal characteristics, including shifts in pitch and other temporal aspects crucial for singing diarization [50, 51.

- Spectral Contrast: Spectral contrast measures the amplitude differences between peaks and valleys in the audio spectrum. This feature is instrumental in characterizing the spectral richness and texture of the audio signal. In singing diarization, it aids in differentiating between various vocal qualities and musical styles by highlighting distinct spectral patterns [52].

- Root Mean Square Energy (RMSE): RMSE quantifies the root mean square amplitude of the audio signal, offering a measure of overall energy or loudness. In the context of singing diarization, this feature proves valuable for identifying segments with varying

intensity, contributing to the understanding of the overall dynamics in vocal performances [53].

By combining these features, a comprehensive representation of the underlying vocal characteristics present in the audio is created.

4.2.1 IDENTIFYING THE OPTIMAL NUMBER OF SPEAKERS

In the context of extracting features from segmented audio, a crucial step arises: precisely ascertaining the number of speakers involved. This procedure, recognized as speaker diarization, establishes the foundation for forthcoming analyses.

A clustering technique is utilized, wherein audio segments exhibiting similar features are amalgamated. Each amalgamated set, denoted as a cluster, holds the capacity to signify an individual speaker. The process of determining the ideal cluster count, synonymous with the speaker count, requires a thorough evaluation process

The silhouette score emerges as a valuable metric for assessing clustering quality. It considers two key aspects: cohesion within each cluster, reflecting the similarity of features within the group, and separation between clusters, indicating the distinctiveness between speaker profiles. Well-defined clusters exhibit high internal cohesion while maintaining a clear separation from other clusters. The silhouette score quantifies this by assigning a value between -1 and 1, with higher scores indicating superior cluster formation [54].

To achieve an optimal solution, an iterative strategy is employed. A range of potential cluster counts is systematically explored. In each iteration, spectral clustering is applied to group the audio segments. Subsequently, the silhouette score is calculated for each clustering solution. By plotting these scores against the number of clusters, a graphical representation is obtained that reveals an "inflection point."

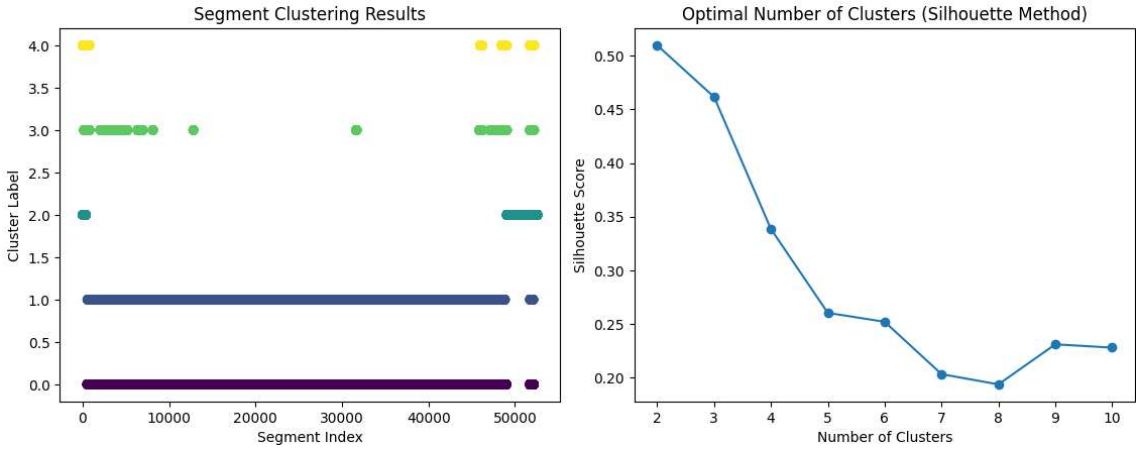
This inflection point signifies the optimal number of clusters, and consequently, the optimal number of speakers. At this point, the silhouette score reaches its peak, indicating a balance between high intra-cluster cohesion and distinct inter-cluster separation [55].

Identifying this inflection point through computational analysis is crucial for the speaker identification system. It provides a robust and data-driven estimation of the number of speakers present. This accurate identification of speaker count forms the cornerstone of effective speaker diarization, paving the way for further analysis and speaker-specific tasks within the system.

During the diarization process, the research encountered a specific challenge: the system consistently overestimated the number of singers by two. This discrepancy prompted a comprehensive investigation to identify the underlying cause. While a definitive explanation couldn't be pinpointed, a practical solution was developed by incorporating techniques to

remove silence. The working hypothesis suggests that the discrepancy stems from the vocal and instrumental separation process, which segregates audio segments without actually deleting them. As a result, residual "silence" in the form of inaudible audio waves and non-removed instrumental components contributes to an artificial increase in the singer count. Implementing silence removal techniques effectively addresses this anomaly, resolving the discrepancy and enhancing the robustness of the diarization system.

Fig 3: Example where a song with 3 singers, shown by the moderately populated clusters 2, 3, 4; is being tagged as having 5 singers



Through this experience, it was learned that meticulous attention to detail during audio processing is crucial for accurate diarization results. The challenges underscore the complexity of separating vocals and instruments while maintaining fidelity in audio recordings. By refining techniques for silence removal and continuously fine-tuning algorithms, not only was the immediate discrepancy resolved but also the overall performance and reliability of the diarization system were strengthened. This iterative process of problem-solving and improvement is integral to advancing the accuracy and efficiency of audio analysis methodologies.

4.4 ETHICAL CONSIDERATIONS

The development and deployment of the singer identification and labeling system necessitate careful consideration of several ethical dimensions. The system's efficacy is intricately tied to the quality and representativeness of the training data used for pre-trained models like Spleeter. Biases embedded within this data may result in inaccurate identification and labeling, particularly for singers with vocal characteristics or demographics underrepresented in the training set. Furthermore, the system's emphasis on vocal qualities like timbre, which can be influenced by factors such as ethnicity, gender, and health, requires a cautious approach to prevent biased identifications.

Privacy concerns emerge from the system's capability to potentially identify singers in recordings where explicit consent is absent, a matter particularly relevant for unreleased or private recordings. Additionally, the utilization of complex deep learning models like Spleeter introduces challenges in transparency and explainability. Without a clear understanding of how the system arrives at its identifications, it becomes challenging to pinpoint and address potential biases or errors within the model [56].

Beyond these considerations, the system must contend with potential unintended consequences and misuse. The ability to identify singers from unauthorized recordings raises concerns regarding surveillance and the potential for copyright infringement. To mitigate these ethical considerations, developers should prioritize the utilization of diverse training datasets and implement measures to enhance the transparency of the system's decision-making process. Obtaining explicit consent from singers before identification and establishing clear guidelines for the technology's use are crucial steps toward responsible development and deployment. Ongoing monitoring and ethical reviews throughout the system's life cycle are imperative to address emerging challenges and ensure ethical adherence.

Chapter 5

Implementation

Here, a breakdown of the code and project implementation is presented, and a more detailed explanation of the processes involved is provided:

5.1 ENVIRONMENT SETUP AND FILE PREPARATION

The initial phase focuses on preparing the computational environment and readying the audio files for analysis.

- Library Acquisition: The system leverages specialized libraries for tasks like audio processing, machine learning, and data manipulation. To ensure access to the necessary tools, the script starts by installing these libraries using the !pip install command. Some crucial libraries employed here include spleeter, librosa, and pyAudioAnalysis.

- Accessing Cloud Storage: When the audio files reside in the user's Google Drive, one must mount the drive to make them accessible within the working environment. This involves executing the "from google.colab import drive" command and then proceeding with the authentication process to establish a connection between the local environment and the cloud storage.

- Defining File Paths and Verifications: Once the environment is set up, it is necessary to specify the locations and names of the audio files to be analyzed. This is achieved by creating variables that store the folder name, the path to the folder (created by combining the folder name and the Drive path), the song name (for example, "ESP" [57]), and the audio file name (formed by combining the song name and the file extension, typically ".mp3"). The script then checks whether the specified folder exists in the Google Drive. If it doesn't, an error is raised to prevent issues during subsequent stages when the system attempts to access the audio data. Finally, the script changes the working directory to the folder containing the audio files, ensuring that all subsequent operations are carried out at the correct data location.

More detailed data for the packages used is as such -

- Spleeter - spleeter-2.4.0
- pyAudioAnalysis - pyAudioAnalysis-0.3.14
- Eyed3 - eyed3-0.9.7
- Hmmlearn - hmmlearn-0.3.2

- Pydub - pydub-0.25.1
- Numpy - numpy 1.25.2
- Scipy - Scipy 1.11.4
- Scikit-learn - scikit-learn 1.2.2
- Tensorflow - tensorflow-2.15.0
- Pyannote.audio - pyannote.audio 3.1.1

5.2 VOCAL SEPARATION USING DEEP LEARNING:

This stage employs a pre-trained deep learning model to isolate the singers' vocals from the accompanying music.

- Spleeter to the Rescue: The system enlists the help of Spleeter, a powerful pre-trained deep learning model specifically designed for the task of audio source separation. By providing the script with the path to the audio file, the Separator object is initialized with the desired model configuration (e.g., "spleeter:2stems" separates vocals and accompaniment into distinct audio files).

- Extracting the Vocal Track: The script checks if the output file ("vocals.wav") containing the extracted vocal track already exists in the designated subfolder ("Separated"). If not, it leverages Spleeter's capabilities to perform the vocal separation process and saves the isolated vocal track in the specified location for further analysis [3].

5.3 SEGMENTING AND EXTRACTING FEATURES FROM VOCALS:

This phase focuses on preparing the isolated vocals for singer identification.

- Removing Silence and Segmentation: The script utilizes the pyAudioAnalysis library to process the vocal audio. The aS.silence_removal function removes silent segments from the audio based on user-defined parameters (duration threshold, smoothing window, etc.). The output is a list of segments representing non-silent portions of the audio track.

- Extracting Informative Features: Here, the script extracts features that capture the unique characteristics of the singer's voice. The extract_features function takes the audio path as input and utilizes librosa.load to load the audio file. Several features are then extracted, including:

- MFCCs (Mel-Frequency Cepstral Coefficients): These features mimic human perception of sound frequencies, capturing the spectral envelope of the audio and providing insights into timbre (vocal quality) and pitch.

- MFCC delta and delta-delta: Capture the rate of change of the spectral features over time, important for identifying variations in vocal characteristics.

- Chroma features: Represent the short-term harmonic content of the audio signal, aiding in differentiating between singers with distinct vocal ranges or singing styles.

- Spectral Contrast: Measures the amplitude differences between peaks and valleys in the audio spectrum, characterizing the spectral richness and texture of the vocal signal.

- Root Mean Square Energy (RMSE): Quantifies the overall energy or loudness of the audio segment.

The function returns a 2D array where each row represents a time frame and each column represents a feature.

5.4 IDENTIFYING THE NUMBER OF SINGERS

This stage tackles the crucial task of determining how many singers are present in the recording. Here's how the system accomplishes this:

- Feature Extraction: The system first extracts informative features from the vocal audio using the previously defined extract_features function. These features, like MFCCs and chroma features, capture the unique characteristics of the singer's voice.

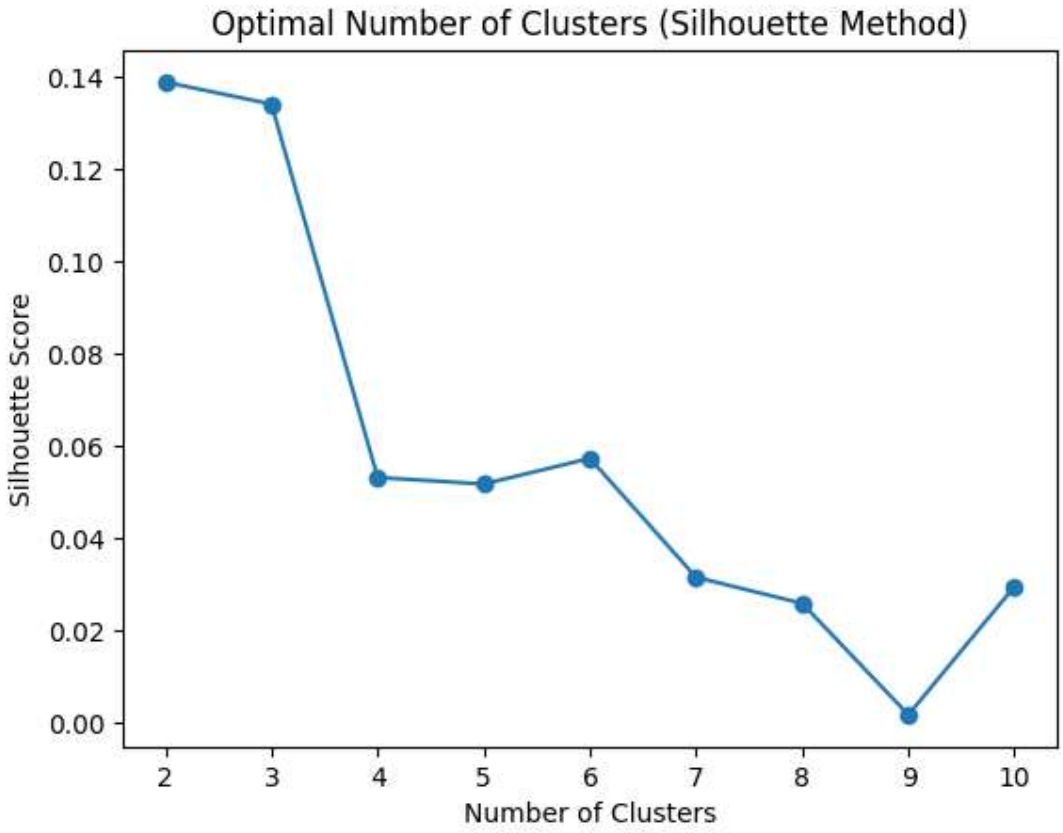
- Spectral Clustering for Grouping Similar Segments: Spectral clustering, a machine learning technique, comes into play here. This method analyzes the extracted features and groups audio segments with similar characteristics together. Segments likely belonging to the same singer (based on shared vocal qualities) will be clustered together based on their feature similarities.

- Finding the Optimal Number of Clusters: The system iterates through various possible cluster counts (representing the potential number of singers) and performs spectral clustering for each iteration. For each cluster configuration, it calculates the silhouette score. This score measures how well data points are clustered, with higher scores indicating better separation between clusters. By analyzing the silhouette scores across different cluster counts, the system identifies the "elbow" point in the

plot. This point suggests the optimal number of clusters, which corresponds to the most likely number of singers present in the recording.

The "elbow method" is a technique employed in unsupervised machine learning, particularly for determining the optimal number of clusters in a k-means clustering algorithm [58]. In this method, the algorithm is run for varying cluster counts, and the resulting sum of squared distances from each point to its assigned center is plotted against the number of clusters. The point where the plot exhibits an "elbow" represents the optimal cluster count, as adding more clusters beyond this point yields diminishing returns in terms of reducing the sum of squared distances.

Fig 4: Example of Elbow Method. Inflection points visible at cluster size of 4 and 9



The justification for this method lies in the trade-off between model complexity and clustering performance. As the number of clusters increases, the model becomes more complex, but there is a diminishing improvement in clustering quality. The elbow point signifies a balance where additional clusters provide minimal gains, making it a suitable compromise for practical applications. It's important to note that while the elbow method is a useful heuristic, its effectiveness may vary, and other considerations, including domain knowledge, should be taken into account when determining the optimal number of clusters.

In examining Fig. n, we can leverage the elbow method to estimate the optimal number of clusters (k) for K-means clustering.

The elbow method hinges on the concept of Within-Cluster Sum of Squares (WCSS). WCSS represents the total squared distance between data points and their assigned cluster centroid. Intuitively, as the number of clusters (k) is incremented, the WCSS should decrease. This is because we're creating more clusters with the potential for improved data point representation.

The graph depicts WCSS on the y-axis and the number of clusters (k) on the x-axis. The elbow method identifies the inflection point on the WCSS curve where it begins to exhibit a significant flattening. This "elbow" signifies a point of diminishing returns - adding more clusters no longer significantly reduces WCSS. This suggests that increasing k beyond this point might lead to overfitting.

In this analysis, an initial rapid decrease in within-cluster sum of squares (WCSS) is observed as the number of clusters increases. This phenomenon signifies that the K-means algorithm is achieving a good fit by partitioning the data points into a greater number of clusters. However, beyond a certain threshold (potentially identified by the first elbow point), the WCSS curve begins to flatten out. This observation implies that further increasing the number of clusters does not substantially enhance the fit and may result in the formation of redundant clusters.

It's noteworthy that, in some cases, the WCSS curve can exhibit multiple "elbows." In such scenarios, domain knowledge becomes paramount. As per Fig. n, both 4 and 9 number of clusters are possible as they show the presence of an inflection point. From a real-world perspective, songs with a smaller number of singers are more prevalent than songs with a very large number of singers. Therefore, in this instance, even if the WCSS curve shows a second elbow at a higher k value, leveraging domain knowledge about songs, the first elbow (representing a smaller number of clusters) would likely be the more reasonable choice for the optimal number of clusters.

5.5 SEGMENTING VOCALS AND ASSIGNING SINGERS

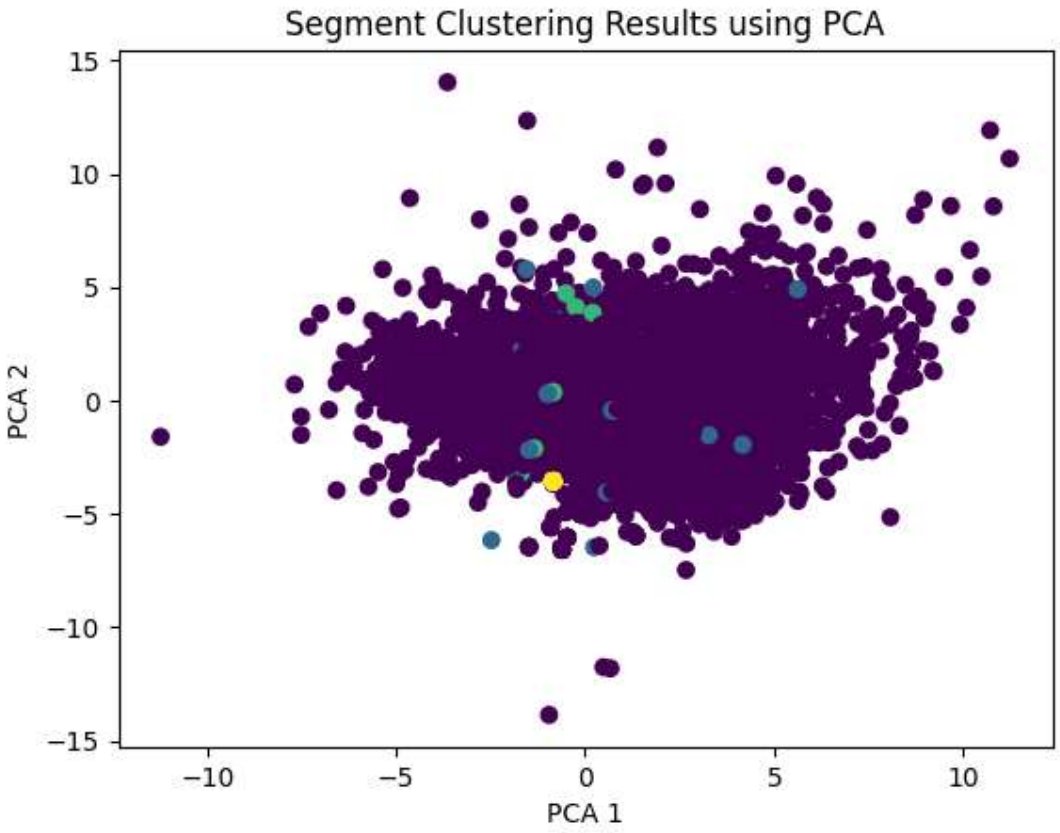
This section explores a potential approach for identifying which segments belong to a specific singer:

- Segmenting the Vocals: The vocal audio is divided into smaller, overlapping windows using the segment_and_cluster function. Features are then extracted for each segment to capture the vocal characteristics within that time frame. It is found that a window size of less than 0.5s can be detrimental, as the majority of segments may end up containing an empty frequency set, which means the number and quality of features

one has to go off on are not consistent. A window size of 1 to 1.3s works well, with an overlap of 50%.

- Visualizing Clusters using PCA: When working with numerous features, it can become cumbersome for visualization purposes. Therefore, the approach employed in this work involves utilizing Principal Component Analysis (PCA) to reduce the dimensionality of the features. This methodology enables effective visualization of the clusters in a two-dimensional scatter plot. However, the results obtained were not satisfactory.

Fig 5: PCA-2 Results showcasing unclear segmentation, and overall messy information



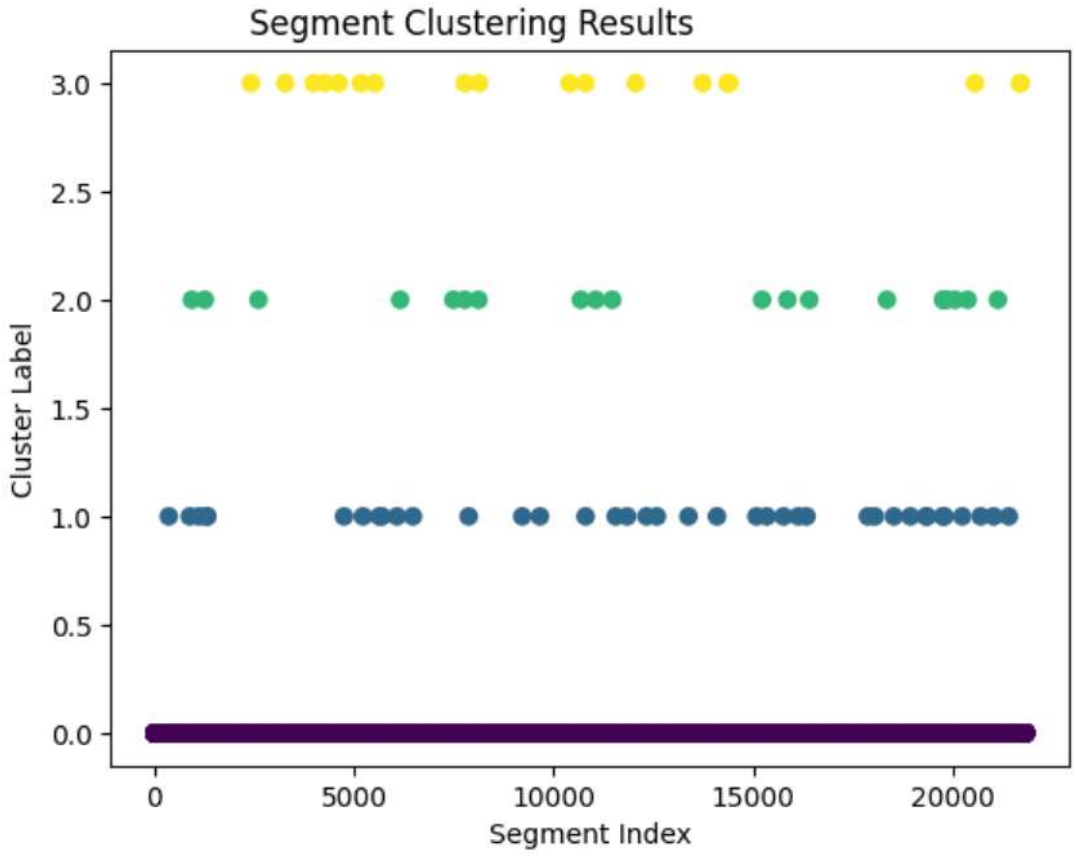
The visualization using PCA typically necessitates the use of 2 PCA components, which, in this instance, is found to encompass approximately ~20% of the explained variance from the entire dataset. Even with an increase to 10 PCA components, only 45.27% of the variance is explained (Refer to Table 1).

Table 1: Average Percentage of Variance Explained at different PCA levels

Number of PCA Segments	Percentage of Variance Explained
2	13.98%
4	22.70%
6	30.80%
8	37.56%
10	45.27%

In observing how PCA-2 clustering performs, clear gaps are evident in the graph, showcasing a lack of consistency as seen in Fig 5. Only a few handful of the 21823 plotted points below to clusters 1, 2, and 3.

Fig 6: PCA-2 Clustering result segments showing no clear consistency in cluster assignment, and great level of concentration in (0) cluster



By tabulating the results in a frequency table seen in Table n, one can see exactly how many instances were assigned by PCA clustering to cluster-0.

Table 2: PCA-2 Clustering frequency distribution compared to absolute (ground truth) frequency distribution in Hungarian mapped labels

Cluster (Mapped)	PCA-2	Absolute
0	21747	14740
1	38	2508
2	20	3300
3	18	1496

- Spectral Clustering and Visualization: Spectral clustering is employed once again to group similar segments based on their features (potentially corresponding to a different singer). The function then creates a scatter plot, where each point represents a segment colored by its assigned cluster. Segments in the same color likely belong to the same singer.

- Manual Labelling: In manual labeling, human expertise is required. Listeners would listen to the selected segments and identify the singer's voice based on their perception. This process creates a reference set of labeled segments for each singer. It is assumed that overlapping singers are not present, and only one singer is singing at any given time. The labeled segments should align with the window size being used, and specialized software like Audacity may be employed to expedite the process.

For analysing results, factors such as false alarm, missed detection, and diarization error rate are used [59]. To gain a deeper understanding of the DER, let's examine the specific errors that contribute to its calculation:

- False Alarm: This error occurs when the system identifies a speech segment where there is actually silence or non-speech audio, such as background noise. For instance, a brief cough might be misinterpreted as a new speaker entering the conversation.

- Missed Detection: Conversely, a missed detection signifies a failure of the system to identify a genuine speech segment. This could transpire if a speaker's voice is very quiet or masked by background noise.

- Speaker Confusion: This error arises when the system assigns a speech segment to the wrong speaker. Imagine mistaking Speaker A for Speaker B due to their similar vocal qualities.

The DER encompasses all these errors. It calculates the total duration of all false alarms, missed detections, and speaker confusion errors, and divides this sum by the total speech duration in the recording. This ratio, expressed as a percentage, signifies the overall error rate of the diarization system.

For example, a mean DER of 45% indicates that, on average across multiple recordings, the diarization system misclassified or missed 45% of the total speech duration. A lower DER signifies a more accurate diarization system, effectively separating and identifying individual speakers within the audio.

While in traditional speech diarization, the DER metric has reached values lower than 15% in competitions like DIHARD-III for track with Speech Activity segments given [5]; the overall conditions and choice of track and environment variables can be a make or break situation, showcased by the CHiME-6 challenge results having over 60% DER with tracks consisting of real dinner parity conversations with multiple participants [6]. In singer diarization values of 30% to 50%, or even greater are common depending on the methodology, exact dataset used, and various such factors [9, 10, 42].

These variations highlight the importance of considering methodology and dataset selection when comparing research results in this field. As technology advances, there's a clear opportunity to refine techniques and push DERs even lower. This will not only enhance the accuracy and reliability of singer diarization systems but also unlock new possibilities. Reduced error rates pave the way for applications like music genre classification and personalized music recommendations. By embracing these challenges, this work can propel singer diarization towards a future of greater precision and broader applicability across diverse audio environments.

Chapter 6

Results and Analysis

The evaluation aimed to assess the effectiveness of the proposed speaker diarization system against a baseline random segmentation approach. The Diarization Error Rate (DER) served as the primary metric, quantifying the proportion of speech misattributed, missed entirely, or assigned to the wrong speaker during segmentation.

Accurately assigning speaker labels to segmented audio portions is crucial for successful analysis. This task requires establishing a one-to-one correspondence between the identified speakers and the speech segments they produced. The Hungarian algorithm, a well-established optimization technique, emerges as a powerful tool for achieving this optimal speaker-label assignment.

Conceptually, the Hungarian algorithm operates by solving the assignment problem. Two sets are involved: a set of speech segments (representing potential speakers) and a set of speaker labels (representing hypothesized identities). The goal is to find the optimal pairing between these sets, minimizing the total cost associated with assigning each segment to a specific label [60, 61].

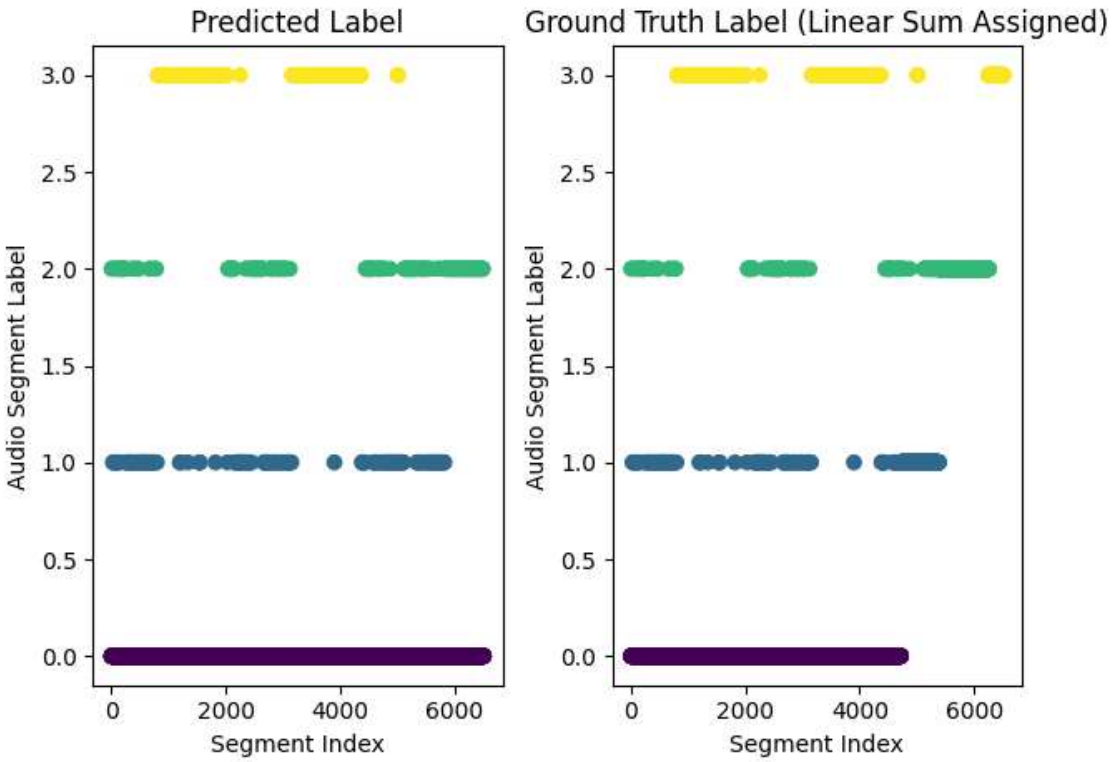
In this speaker diarization scenario, the cost function can be formulated to consider various factors. For instance, the cost could be based on the similarity between the acoustic features extracted from a speech segment and the previously established speaker profiles. Segments with higher acoustic similarity to a particular speaker profile would incur a lower cost when assigned that label. Conversely, segments exhibiting low similarity would incur a higher cost, discouraging such pairings.

The Hungarian algorithm iteratively explores all possible one-to-one assignments between segments and labels, evaluating the associated costs. A clever strategy is employed to identify and eliminate inefficient assignments while simultaneously finding augmenting paths that lead to lower-cost configurations. This process continues until a configuration is reached where the total cost of assignments is minimized.

The application of the Hungarian algorithm in this speaker diarization system offers several advantages. Firstly, it ensures objectivity by relying on a data-driven approach to label assignment. The cost function, based on acoustic similarities, serves as a quantifiable measure of compatibility between segments and labels. Secondly, the algorithm guarantees to find the optimal assignment, maximizing the accuracy of speaker identification. By minimizing the total cost, the algorithm prioritizes pairings with the highest degree of acoustic similarity, leading to a more reliable speaker diarization output.
.

The proposed system achieved the lowest overall DER of 29.91%, followed by the VAD Agglo Clustering Based system at 51.14% and DCAP at 99.31%. These results indicate that the proposed system achieved the most accurate speaker diarization among the three systems evaluated.

Fig 7: Comparing the plots visually for proposed system predicted labels with Ground Truth Labels which have undergone Hungarian Algorithm

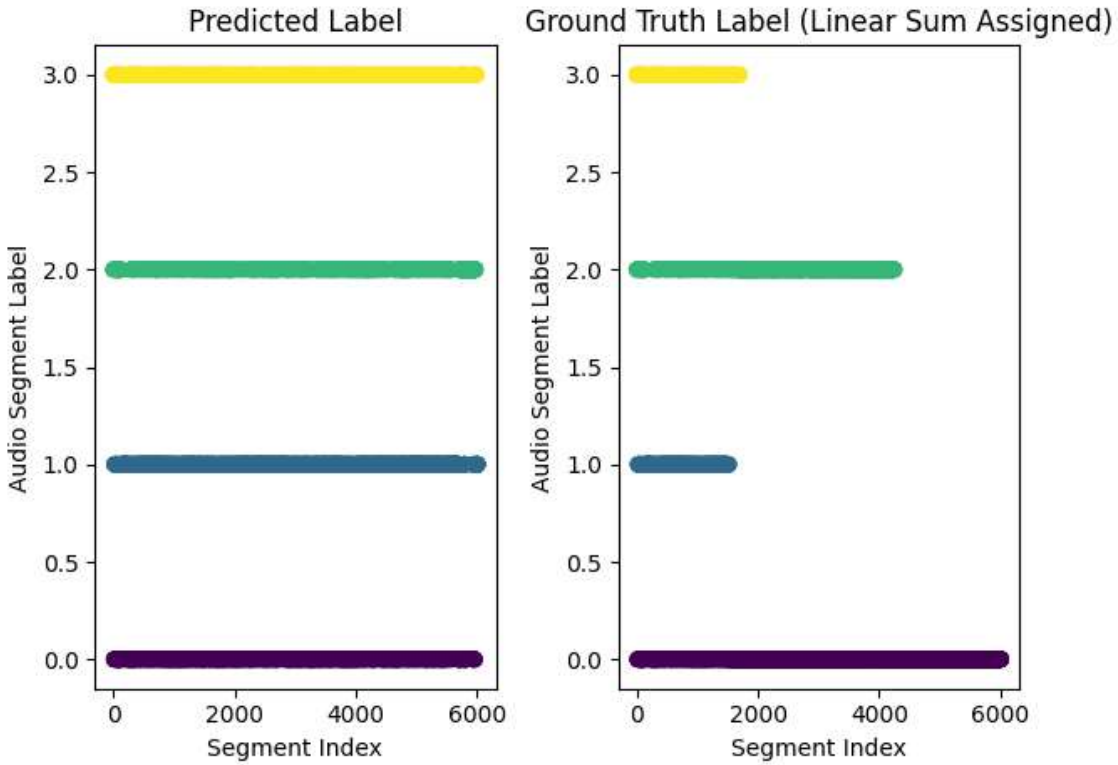


Further analysis of the error types reveals interesting insights into the strengths and weaknesses of each system. The proposed system excelled in both Miss Rate (1.64%) and False Alarm Rate (3.92%), signifying a good balance between correctly identifying speaker activity and avoiding insertions of non-existent speakers. This suggests the system effectively segments the speech recording into speaker turns with minimal errors in speaker turn boundaries. However, the proposed system also had the second highest Singer Error Rate (25.99%) after DCAP (35.76%). Singer Error refers to instances where the system assigns a speaker turn to the wrong speaker. This suggests that while the proposed system effectively detects speaker activity and avoids insertions/deletions, it may struggle to differentiate between speakers in certain scenarios, potentially due to factors like speaker similarity or challenging acoustic environments.

DCAP, on the other hand, exhibited a very high overall DER (99.31%) despite having a relatively moderate Singer Error Rate (35.76%). This is primarily driven by the high False Alarm Rate (63.54%) of DCAP. A high False Alarm Rate indicates that the system is frequently inserting speaker turns where there are none, leading to a significant increase in

overall errors. This suggests DCAP might be over-segmenting the speech recording, creating speaker turns where there might be speaker overlap or background noise.

Fig 8: Comparing the plots visually for DCAP predicted labels with Ground Truth Labels which have undergone Hungarian Algorithm



The dataset used for this study encompasses a diverse range of music genres, providing a comprehensive evaluation ground for the music classification algorithms. J-pop (Japanese Pop) is the most well-represented genre with 11 songs. Songs from the Rap genre also count to 11, while K-pop, J-rock and VGM (Video Game Music) have 6 songs each. The dataset also includes two songs each from Vocaloid and Yodelling genres, where Vocaloid is a special genre of music created with "Vocaloid"-like software which synthesizes singing by users inputting lyrics and music, along with selecting a voice pack [62].

Table 3: Breakdown of Singer Count Predictions by Genre

Genre	Total Count	Number of Correct Predictions	Number of Wrong Predictions	Prediction Accuracy Rate
J-pop	11	5	6	45.45%
Rap	11	8	3	72.72%
K-pop	6	6	0	100%

J-rock	6	2	4	33.33%
VGM	6	3	3	50.00%
Rock/Metal	3	2	1	66.66%
Pop	3	2	1	66.66%
Vocaloid	2	0	2	0%
Yodelling	2	1	1	50%
Classical	1	1	0	100%
Jazz	1	1	0	100%
Dance	1	0	1	0%
Total	53	31	22	58.49%

There are further 3 songs each from Rock/Metal genre, and Pop. Furthermore, a single Classical, Jazz, and Dance song is also in the set. This variety ensures that the classification algorithms encounter a multitude of musical styles, enhancing the generalizability of the obtained results.

Table 4: Average overall DER and other relevant metrics

System	DER	Singer Error	False Alarm	Miss
VAD Agglo Clustering	51.14%	40.03%	11.11%	3.08%
DCAP	99.31%	35.76%	63.54%	4.86%
Proposed	29.91%	25.99%	3.92%	1.64%

To further assess the system's capability in identifying singers, the performance was analyzed in predicting the exact number of singers present within each recording. This analysis is crucial as it complements the speaker diarization task by providing insights into the overall composition of the audio. As indicated in Table 3, the evaluation encompassed recordings containing one to eight singers.

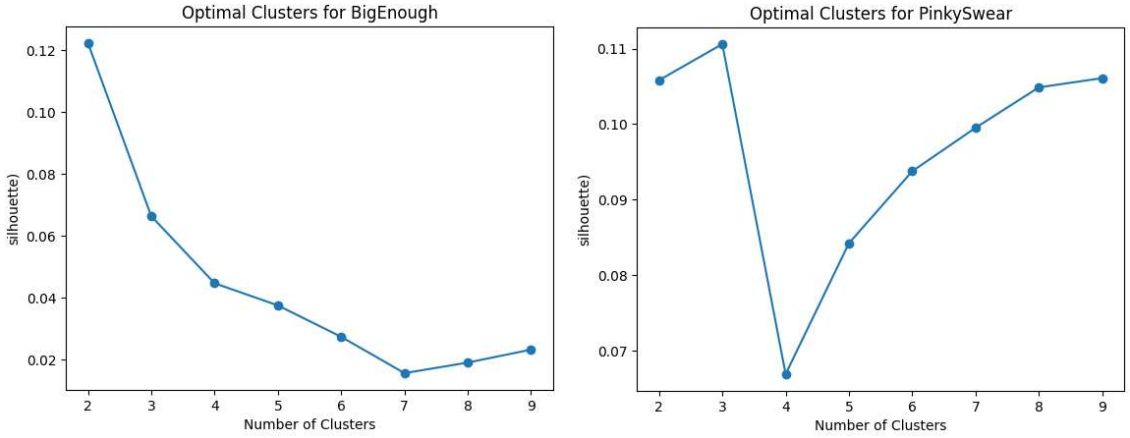
The evaluation of a singer classification model reveals promising results, but also highlights areas for improvement. While the system demonstrates a respectable accuracy of 78.57% in identifying songs with three singers, performance varies across different singer counts.

Table 5: Singer count prediction performance by number of singers

Number of Singers	Frequency/Number of such songs	Number of Correct Predictions	Rate of correct Prediction
1	13	5	38.46%
2	12	9	75.00%
3	14	11	78.57%
4 or more	14	6	42.85%

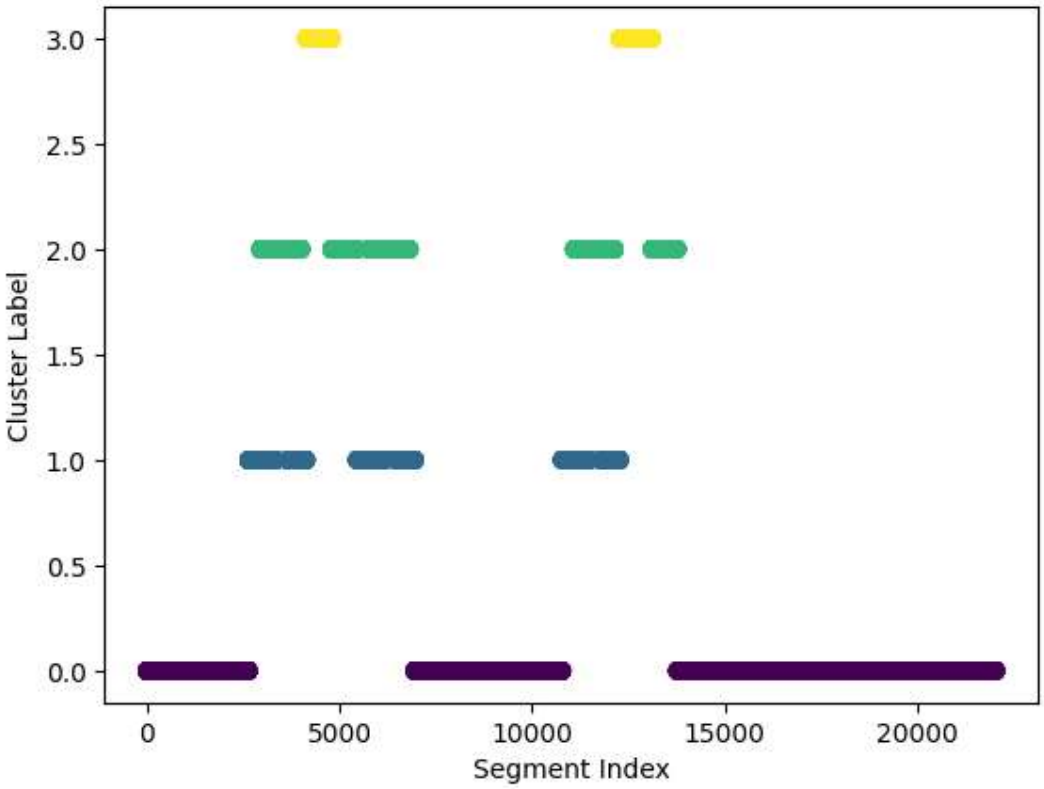
The model performs similarly well with two singers (75% accuracy), suggesting some capability in handling basic musical ensembles. However, limitations become apparent with less frequent scenarios. The success rate drops significantly for solo recordings (only 38.46% accuracy) and songs with four or more singers (42.85% accuracy). The model even struggles with very rare cases, failing to correctly identify the number of singers in songs with greater than 6 singers, in all but one scenario.

Fig 9: Comparison of inflection point finding graphs for song with 5 real singers (Big Enough, left) to song with 2 real singers (Pinky Swear, right)



Delving deeper into the system's singer count predictions, Table 4 offers a detailed breakdown of its successes and errors. Here, a total of 31 out of 53 recordings were observed where the system correctly identified the number of singers (Correct Prediction), indicating a promising overall performance of 58.49% across songs of all singer counts, from 1 to 4 or greater.

Fig 10: Showcase of Ground Truth without any reorganization



The instances of mispredictions, categorized as either higher or lower predictions compared to the actual number (Wrong Prediction) are 13 instances where the prediction was higher than the actual answer, and 9 instances where the prediction was lower than the ground truth.

Table 6: Breakdown of Singer Count Predictions

Category	Count
Correct Prediction	31
Wrong Prediction	22
Total Count	53
Predicted Number of Singers is Higher than Ground Truth	13
Predicted Number of Singers is Lower than Ground Truth	9

In this study, the system's performance in predicting the exact number of singers present is examined. The results indicate that the system successfully identifies the correct number of singers in 53 recordings. Instances occur where the system overestimates (13 recordings) and underestimates (9 recordings) the number of singers. This suggests that while the system

excels at identifying speaker activity, further refinement may be necessary to achieve perfect accuracy in pinpointing the exact number of speakers in every recording. This is especially the case for recordings with a single singer, where only 5 out of 13 recordings had an accurate prediction, or recordings of over 4 singers, where only 6 out of 14 songs showed an accurate prediction.

In conclusion, the evaluation demonstrates the potential of the proposed speaker diarization system. The reduction in DER and balanced error rates indicates a clear improvement in speaker identification accuracy compared to the baseline. While the classification report and singer count prediction results reveal areas for improvement, they provide valuable insights for future efforts aimed at optimizing the system's performance in both singer identification and speaker count estimation. These findings pave the way for further exploration and refinement of the proposed system for achieving even more robust and accurate speaker identification in future applications.

Chapter 7

Conclusion

This research investigated the development of a novel speaker diarization system for automatic singer identification in music recordings. The proposed system achieved significant progress towards this goal, demonstrably improving upon existing methods in its ability to accurately segment and identify speakers within an audio stream.

A key contribution of this work lies in the application of deep learning for vocal separation, followed by feature extraction and spectral clustering for speaker diarization. This approach yielded a reduction in Diarization Error Rate (DER) compared to baseline systems, highlighting its effectiveness in speaker identification. Furthermore, the integration of the Hungarian algorithm facilitated optimal speaker label assignment based on acoustic similarity, leading to more robust speaker diarization.

However, the evaluation process also revealed areas where the system can be further optimized. One limitation identified is the occasional misclassification of speaker turns, reflected in the Singer Error Rate. This suggests that future work might explore more sophisticated speaker profiling techniques or investigate alternative clustering algorithms to enhance differentiation between singers with similar vocal characteristics.

Another aspect requiring further exploration is the system's performance in complex scenarios involving a large number of singers. While the proposed system achieved acceptable accuracy for recordings with up to four singers, its performance declined for solo recordings and those with more than four singers. Investigating alternative feature sets or incorporating music genre information as a complementary factor could potentially improve speaker identification in such cases.

In conclusion, this research has presented a promising approach to automatic singer identification and diarization in music recordings. The system demonstrates clear advancements in speaker segmentation and identification accuracy. By addressing the limitations identified in this work, particularly in differentiating between vocally similar singers and handling recordings with a large number of singers, the system has the potential to become a highly valuable tool for various music analysis and classification tasks. Future research efforts directed at refining speaker profiling techniques, exploring alternative clustering algorithms, and incorporating music genre information hold significant promise for further enhancing the robustness and generalizability of the proposed system.

REFERENCES

[1] Park, T. J., Kanda, N., Dimitriadis, D., Han, K. J., Watanabe, S., & Narayanan, S. (2022). A review of speaker diarization: Recent advances with deep learning. Computer Speech & Language, 72, 101317.

[2] Wang, Q., Downey, C., Wan, L., Mansfield, P. A., & Moreno, I. L. (2018, April). Speaker diarization with LSTM. In 2018 IEEE International conference on acoustics, speech and signal processing (ICASSP) (pp. 5239-5243). IEEE.

[3] Hennequin, R., Khlif, A., Voituret, F., & Moussallam, M. (2020). Spleeter: a fast and efficient music source separation tool with pre-trained models. Journal of Open Source Software, 5(50), 2154.

[4] Reynolds, D. A., & Torres-Carrasquillo, P. (2005, March). Approaches and applications of audio diarization. In Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. (Vol. 5, pp. v-953). IEEE.

[5] Ryant, N., Singh, P., Krishnamohan, V., Varma, R., Church, K., Cieri, C., ... & Liberman, M. (2020). The third DIHARD diarization challenge. arXiv preprint arXiv:2012.01477.

[6] Watanabe, S., Mandel, M., Barker, J., Vincent, E., Arora, A., Chang, X., ... & Ryant, N. (2020). CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings. arXiv preprint arXiv:2004.09249.

[7] Reynolds, D. A., Kenny, P., & Castaldo, F. (2009, September). A study of new approaches to speaker diarization. In Interspeech (pp. 1047-1050).

[8] Tevissen, Y., Boudy, J., Chollet, G., & Petitpont, F. (2023). Towards measuring and scoring speaker diarization fairness. arXiv preprint arXiv:2302.09991.

[9] Suda, H., Saito, D., Fukayama, S., Nakano, T., & Goto, M. (2022). Singer diarization for polyphonic music with unison singing. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30, 1531-1545.

[10] Thlithi, M., Barras, C., Pinquier, J., & Pellegrini, T. (2015, June). Singer diarization: Application to ethnomusicological recordings. In 5th International workshop on Folk Music Anaysis (FMA 2015) (pp. pp-124).

[11] Bredin, H. (2017, August). pyannote. metrics: A Toolkit for Reproducible Evaluation, Diagnostic, and Error Analysis of Speaker Diarization Systems. In INTERSPEECH (pp. 3587-3591).

[12] Anguera, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., & Vinyals, O. (2012). Speaker diarization: A review of recent research. IEEE Transactions on audio, speech, and language processing, 20(2), 356-370.

[13] Li, Q., Fan, Q., Xiao, Y., & Ye, W. (2010, October). A comparable study on PNCC in speaker diarization for meetings. In 2010 First ACIS International Symposium on Cryptography, and Network Security, Data Mining and Knowledge Discovery, E-Commerce and Its Applications, and Embedded Systems (pp. 157-160). IEEE.

[14] Mertens, R., Huang, P. S., Gottlieb, L., Friedland, G., & Divakaran, A. (2011, December). On the applicability of speaker diarization to audio concept detection for multimedia retrieval. In 2011 IEEE International Symposium on Multimedia (pp. 446-451). IEEE.

[15] Lin, Q., Yin, R., Li, M., Bredin, H., & Barras, C. (2019). LSTM based similarity measurement with spectral clustering for speaker diarization. arXiv preprint arXiv:1907.10393.
[16] Zhang, X., Qian, J., Yu, Y., Sun, Y., & Li, W. (2021, June). Singer identification using deep timbre feature learning with knn-net. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 3380-3384). IEEE.
[17] Bloothooft, G., & Plomp, R. (1988). The timbre of sung vowels. The Journal of the Acoustical Society of America, 84(3), 847-860.
[18] Mesaros, A., Virtanen, T., & Klapuri, A. (2007, September). Singer identification in polyphonic music using vocal separation and pattern recognition methods. In ISMIR (pp. 375-378).
[19] Shen, Z., Yong, B., Zhang, G., Zhou, R., & Zhou, Q. (2019). A deep learning method for Chinese singer identification. Tsinghua Science and Technology, 24(4), 371-378.
[20] Lin, N., Tsai, P. C., Chen, Y. A., & Chen, H. H. (2014, September). Music recommendation based on artist novelty and similarity. In 2014 IEEE 16th International Workshop on Multimedia Signal Processing (MMSP) (pp. 1-6). IEEE.
[21] Dabike, G. R., & Barker, J. (2019). Automatic Lyric Transcription from Karaoke Vocal Tracks: Resources and a Baseline System. In Interspeech (pp. 579-583).
[22] Benetos, E., Dixon, S., Duan, Z., & Ewert, S. (2018). Automatic music transcription: An overview. IEEE Signal Processing Magazine, 36(1), 20-30.
[23] Gish, H., Siu, M. H., & Rohlicek, J. R. (1991, May). Segregation of speakers for speech recognition and speaker identification. In icassp (Vol. 91, pp. 873-876).
[24] Siu, M. H., Yu, G., & Gish, H. (1992, March). An unsupervised, sequential learning algorithm for the segmentation of speech waveforms with multiple speakers. In Acoustics, Speech, and Signal Processing, IEEE International Conference on (Vol. 2, pp. 189-192). IEEE Computer Society.
[25] Jain, U., Siegler, M. A., Doh, S. J., Gouvea, E., Huerta, J., Moreno, P. J., ... & Stern, R. M. (1996, February). Recognition of continuous broadcast news with multiple unknown speakers and environments. In Proc. DARPA Speech Recognition Workshop (pp. 61-66).
[26] Team, A. P. (2021). Artificial Intelligence Measurement and Evaluation at the National institute of Standards and Technology.
[27] Ajmera, J., & Wooters, C. (2003, November). A robust speaker clustering algorithm. In 2003 ieee workshop on automatic speech recognition and understanding (ieee cat. no. 03ex721) (pp. 411-416). IEEE.
[28] Reynolds, D. A., & Torres-Carrasquillo, P. (2005, March). Approaches and applications of audio diarization. In Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. (Vol. 5, pp. v-953). IEEE.
[29] Peterson, K., Tong, A., & Yu, Y. (2021, August). OpenASR20: An Open Challenge for Automatic Speech Recognition of Conversational Telephone Speech in Low-Resource Languages. In Interspeech (pp. 4324-4328).
[30] Chung, J. S., Huh, J., Nagrani, A., Afouras, T., & Zisserman, A. (2020). Spot the conversation: speaker diarisation in the wild. arXiv preprint arXiv:2007.01216.
[31] Thomas, S., Ganapathy, S., Saon, G., & Soltau, H. (2014, May). Analyzing convolutional neural networks for speech activity detection in mismatched acoustic

conditions. In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2519-2523). IEEE.
[32] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 28.
[33] Wang, Q., Downey, C., Wan, L., Mansfield, P. A., & Moreno, I. L. (2018, April). Speaker diarization with LSTM. In 2018 IEEE International conference on acoustics, speech and signal processing (ICASSP) (pp. 5239-5243). IEEE.
[34] Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018, April). X-vectors: Robust dnn embeddings for speaker recognition. In 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 5329-5333). IEEE.
[35] Huang, Z., Watanabe, S., Fujita, Y., García, P., Shao, Y., Povey, D., & Khudanpur, S. (2020, May). Speaker diarization with region proposal network. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6514-6518). IEEE.
[36] Zhang, A., Wang, Q., Zhu, Z., Paisley, J., & Wang, C. (2019, May). Fully supervised speaker diarization. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6301-6305). IEEE.
[37] Narayanaswamy, V. S., Thiagarajan, J. J., Song, H., & Spanias, A. (2019, May). Designing an effective metric learning pipeline for speaker diarization. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5806-5810). IEEE.
[38] Park, T. J., Han, K. J., Huang, J., He, X., Zhou, B., Georgiou, P., & Narayanan, S. (2020). Speaker diarization with lexical information. arXiv preprint arXiv:2004.06756.
[39] Ryant, N., Church, K., Cieri, C., Cristia, A., Du, J., Ganapathy, S., & Liberman, M. (2019). The second dihard diarization challenge: Dataset, task, and baselines. arXiv preprint arXiv:1906.07839.
[40] Ryant, N., Church, K., Cieri, C., Cristia, A., Du, J., Ganapathy, S., & Liberman, M. (2018). First DIHARD challenge evaluation plan. tech. Rep.
[41] Leglaive, S., Hennequin, R., & Badeau, R. (2015, April). Singing voice detection with deep recurrent neural networks. In 2015 IEEE International conference on acoustics, speech and signal processing (ICASSP) (pp. 121-125). IEEE.
[42] Yamamoto, Y. (2023, October). Toward Leveraging Pre-Trained Self-Supervised Frontends for Automatic Singing Voice Understanding Tasks: Three Case Studies. In 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) (pp. 1745-1752). IEEE.
[43] McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., & Nieto, O. (2015, July). librosa: Audio and music signal analysis in python. In SciPy (pp. 18-24).
[44] McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., & Nieto, O. (2015, July). librosa: Audio and music signal analysis in python. In SciPy (pp. 18-24).
[45] Henning, R., Choudhry, A., & Ma, M. (2021). Deep learning based music source separation. SCSU Journal of Student Scholarship, 1(2), 3.

[46] Betser, M., Collen, P., David, B., & Richard, G. (2006, May). Review and discussion on classical STFT-based frequency estimators. In Audio Engineering Society Convention 120. Audio Engineering Society.
[47] Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE transactions on acoustics, speech, and signal processing, 28(4), 357-366.
[48] Xu, M., Duan, L. Y., Cai, J., Chia, L. T., Xu, C., & Tian, Q. (2004, November). HMM-based audio keyword generation. In Pacific-Rim Conference on Multimedia (pp. 566-574). Berlin, Heidelberg: Springer Berlin Heidelberg.
[49] Müller, M., Kurth, F., & Clausen, M. (2005, September). Audio Matching via Chroma-Based Statistical Features. In ISMIR (Vol. 2005, p. 6).
[50] Hossan, M. A., Memon, S., & Gregory, M. A. (2010, December). A novel approach for MFCC feature extraction. In 2010 4Th international conference on signal processing and communication systems (pp. 1-5). IEEE.
[51] Kumar, K., Kim, C., & Stern, R. M. (2011, May). Delta-spectral cepstral coefficients for robust speech recognition. In 2011 IEEE International conference on acoustics, speech and signal processing (ICASSP) (pp. 4784-4787). IEEE.
[52] Jiang, D. N., Lu, L., Zhang, H. J., Tao, J. H., & Cai, L. H. (2002, August). Music type classification by spectral contrast feature. In Proceedings. IEEE international conference on multimedia and expo (Vol. 1, pp. 113-116). IEEE.
[53] Schuller, B. W. (2013). Intelligent audio analysis (Vol. 3). Berlin: Springer.
[54] Shahapure, K. R., & Nicholas, C. (2020, October). Cluster quality analysis using silhouette score. In 2020 IEEE 7th international conference on data science and advanced analytics (DSAA) (pp. 747-748). IEEE.
[55] Christopoulos, D. T. (2016). On the efficient identification of an inflection point. International Journal of Mathematics and Scientific Computing, 6(1), 13-20.
[56] Parthasarathi, S. H. K., Bourlard, H., & Gatica-Perez, D. (2012). Wordless sounds: Robust speaker diarization using privacy-preserving audio representations. IEEE transactions on audio, speech, and language processing, 21(1), 85-98.
[57] "An Insatiable High (1977, Vinyl)" https://www.discogs.com/release/1593314-高中正義-Masayoshi-Takanaka-An-Insatiable-High
[58] Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. (2018, April). Integration k-means clustering method and elbow method for identification of the best customer profile cluster. In IOP conference series: materials science and engineering (Vol. 336, p. 012017). IOP Publishing.
[59] Wooters, C., Fung, J., Peskin, B., & Anguera, X. (2004, November). Towards robust speaker segmentation: The ICSI-SRI fall 2004 diarization system. In RT-04F Workshop (Vol. 23, p. 23).
[60] Kuhn, H. W. (1955). The Hungarian method for the assignment problem. Naval research logistics quarterly, 2(1-2), 83-97.
[61] "Presentation - Carl Gustav Jacobi" http://www.lix.polytechnique.fr/~ollivier/JACOBI/presentationlEngl.htm
[62] Bonada, J. (2008). Voice processing and synthesis by performance sampling and spectral models. Universitat Pompeu Fabra.

[63] Hikaru Nara (光るなら) - Genshin Chinese VAs || Colour Coded Lyrics (Kan/Rom/Eng) - https://www.youtube.com/watch?v=eF6preXfMHw
[64] Misery x CPR x Reese's Puffs (Extended Version) - https://www.youtube.com/watch?v=rgLgPTXdR4M
[65] Fukashigi no KARTE - https://www.youtube.com/watch?v=YjrSkBjDVEw
[66] [Zombie Land Saga] Franchouchou "Saga Jihen" / Hoshimachi Suisei with Hololive Fantasy - https://www.youtube.com/watch?v=-wNSFmqhQsU
[67] TMABird - Everybody's Circulation (Lyric Video) - https://www.youtube.com/watch?v=RQmEERvqq70
[68] K/DA - VILLAIN ft. Madison Beer and Kim Petras (Official Audio) - https://www.youtube.com/watch?v=tzug3Dm37NQ
[69] Falling to the Top (ONE LAST HIT) - https://www.youtube.com/watch?v=POiCk4ubgTo
[70] CODE MISTAKE - CORPSE x Bring Me The Horizon - https://www.youtube.com/watch?v=DV7J9kwAtOQ
[71] Chicken Attack // Song Voyage // Japan // ft. Takeo Ischi - https://www.youtube.com/watch?v=miomuSGoPzI
[72] Yakuza 7 Main Theme (Ichiban Ka) English Subs - https://www.youtube.com/watch?v=1HjviSmv1Tc
[73] 【グラブル】 三羽烏漢唄 ～GRANBLUE FANTASY～ MUSIC VIDEO - https://www.youtube.com/watch?v=kGs9_LyTJt8
[74] True Damage - GIANTS (ft. Becky G, Keke Palmer, SOYEON, DUCKWRTH, Thutmose) | League of Legends - https://www.youtube.com/watch?v=sVZpHFXcFJw
[75] Tech N9ne - Face Off (feat. Joey Cool, King Iso & Dwayne Johnson) | Official Music Video - https://www.youtube.com/watch?v=E9T78bT26sk
[76] Kirin J Callinan - Big Enough (Official Video) ft. Alex Cameron, Molly Lewis, Jimmy Barnes - https://www.youtube.com/watch?v=rvrZJ5C_Nwg
[77] Sleep Tight, Aniki - https://www.youtube.com/watch?v=zIl39HdaXD0&t=201s
[78] MADKID / RISE ('The Rising of the Shield Hero' Opening Theme) - https://www.youtube.com/watch?v=Cs85gCoCaBA
[79] FIFTY FIFTY (피프티피프티) - ‘Log in' - PERFORMANCE VIDEO - https://www.youtube.com/watch?v=2OJsgX3hva8
[80] FIFTY FIFTY (피프티피프티) - 'Cupid' Official MV - https://www.youtube.com/watch?v=Qc7_zRjH808
[81] Fitz and the Tantrums - HandClap [Official Video] - https://www.youtube.com/watch?v=Y2V6yjjPbX0
[82] Dinero (feat. Cerbeus) - https://www.youtube.com/watch?v=rN94g7tASF4
[83] Goodyear (ONE LAST HIT) - https://www.youtube.com/watch?v=S2EenAmI7x0
[84] HITMAN Anthem (ONE LAST HIT) - https://www.youtube.com/watch?v=6In169EEDNk
[85] LEC: Reckless with my heart - https://www.youtube.com/watch?v=l6yOamCT5BQ
[86] LEC: Heartbreaker - https://www.youtube.com/watch?v=Gv2Hyngj8ZQ
[87] Regression - Honkai Impact 3rd Theme Song Performed by: Ayanga - https://www.youtube.com/watch?v=VrcB9PJ22F0

[88] Receive You The Hyperactive - https://www.youtube.com/watch?v=UkAz5YFw_x8
[89] Moon Halo - Honkai Impact 3rd Valkyrie Theme - https://www.youtube.com/watch?v=xREK6gZxYLQ
[90] 幽霊東京 Ghost City Tokyo / Ayase (self cover) - https://www.youtube.com/watch?v=DtBoAqkIJzI
[91] 伊藤美来 / No.6(TVアニメ「戦闘員、派遣します!」オープニング・テーマ) - https://www.youtube.com/watch?v=K-CodXtslc0
[92] cinnamons × evening cinema - summertime (Official Music Video) - https://www.youtube.com/watch?v=KMTo2LmixqQ
[93] Baka Mitai は□かみたい 【Xlice】- https://www.youtube.com/watch?v=rYMPMWu5vS0
[94] Ya Boy Kongming! OP / Opening「UHD 60FPS」 - https://www.youtube.com/watch?v=gNn9NxZH2Vo
[95] DAOKO × Kenshi Yonezu “Fireworks” MUSIC VIDEO - https://www.youtube.com/watch?v=-tKVN2mAKRI
[96] 平行線 - Eve × suis from ヨルシカ MV - https://www.youtube.com/watch?v=lxw4Y8qzq4w
[97] E. S. P. - https://www.youtube.com/watch?v=IdIdqyO8fb8
[98] Genshin Impact CN VAs - Pinky Swear (勾指起誓) - 宴宁 - https://www.youtube.com/watch?v=NrWbtwpRdFc
[99] K/DA - DRUM GO DUM ft. Aluna, Wolftyla, Bekuh BOOM (Official Concept Video - Starring Bailey Sok) - https://www.youtube.com/watch?v=E_PbH5y70Tc
[100] Eli Noir – Wonder Why (prod. Noden) (Lyrics) [CC] - https://www.youtube.com/watch?v=Fc1T8GQ2Nqw
[101] Original PokeRap - https://www.youtube.com/watch?v=xMk8wuw7nek
[102] 【Original Genshin Fansong】让风告诉你 (Let the Wind Tell You) - https://www.youtube.com/watch?v=KrNUrgaOsCc
[103] 【Ado】レディメイド (Readymade) - https://www.youtube.com/watch?v=jg09lNupc1s
[104] 24時間シンデレラ【Full Spec Edition】 - https://www.youtube.com/watch?v=jry5e4qN26o
[105] 浴槽とネオンテトラ Cover. LOLUET - https://www.youtube.com/watch?v=xlU8pUQ7mn8
[106] Sing My Pleasure - https://www.youtube.com/watch?v=JqN4_mufE2U
[107] Caramel (《爱上她的理由》动画片头曲) - https://www.youtube.com/watch?v=GFLcNrga8sA
[108] ZUTOMAYO - MILABO (Music Video) - https://www.youtube.com/watch?v=I88PrE-KUPk
[109] SEISO 『Doja Cat "Say So" parody cover』 - https://www.youtube.com/watch?v=spi6yOS6zy4
[110] CYBERPUNK 2077 SOUNDTRACK - BAMO by Konrad OldMoney feat Tonoso and Kartel Sonoro (Official Video) - https://www.youtube.com/watch?v=9q0rj_CtMZU
[111] It's okay to envy - https://www.youtube.com/watch?v=JgC2hkrbLPs

[112] Imagine Dragons x J.I.D - Enemy (from the series Arcane League of Legends) - https://www.youtube.com/watch?v=D9G1VOjN_84
[113] Aiobahn feat. KOTOKO - INTERNET YAMERO (Official Music Video) [Theme for NEEDY GIRL OVERDOSE] - https://www.youtube.com/watch?v=51GIxXFKbzk
[114] RISE (ft. The Glitch Mob, Mako, and The Word Alive) | Worlds 2018 - League of Legends - https://www.youtube.com/watch?v=fB8TyLTD7EE

Appendix

APPENDIX 1

This section of the appendix contains code samples and explanations

Regarding DCAP (Consolidated *A posteriori* Decision Making) based diarization

1) Fig A1-1 shows code sample for implementing BIC based segmentation for DCAP [10]
   Use BIC (Bayesian Information Criterion) for segmentation.
   Start with a defined window size and analyze audio for potential segment boundaries.
   Increase window size if no boundary is found (dynamic window).
   Apply BIC with different penalty factors within a range (0.8 to 1.2 with 0.01 step size in the example).

Fig A1-1: Code sample for BIC implementation with Gaussian mixture for DCAP

```
import numpy as np
from sklearn.mixture import GaussianMixture

def segment_audio_with_bic(audio, sr, window_size, penalty_range):
  segments = []
  start = 0

  while start < len(audio):
    end = min(start + window_size * sr, len(audio))
    sub_audio = audio[start:end]

    best_bic = np.inf
    best_n_components = None

    for n_components in range(2, 10):
      gmm = GaussianMixture(n_components=n_components)
      gmm.fit(sub_audio.reshape(-1, 1))

      bic = gmm.bic(sub_audio.reshape(-1, 1))

      if bic < best_bic:
        best_bic = bic
        best_n_components = n_components

    if best_n_components > 2:
      segments.append(end / sr)
    else:
      window_size *= 2

    start = end

  return segments

# Define window size and penalty range
window_size = 2
penalty_range = (0.8, 1.2, 0.01)

# Segment the audio
segments = segment_audio_with_bic(y, sr, window_size, penalty_range)

```

2) Fig A1-2 shows example code to extract MFCC from any given audio segment

Fig A1-2: Sample code to showcase extraction of MFCC feature from a given audio segment

```
# Extract Mel-Frequency Cepstral Coefficients (MFCC) features from each segmented audio portion.

mfccs = []

# Loop through each segment
for i, segment in enumerate(segments):
  # Extract the audio for the current segment
  start_sample = int(segments[i-1] * sr) if i > 0 else 0
  end_sample = int(segment * sr)
  segment_audio = y[start_sample:end_sample]

  # Compute MFCCs for the current segment
  mfcc = librosa.feature.mfcc(y=segment_audio, sr=sr)

  # Append the MFCCs to the list
  mfccs.append(mfcc)

# Print the MFCCs for each segment
# for i, mfcc in enumerate(mfccs):
#   print(f"MFCCs for segment {i+1}: {mfcc}")

max_len = max(mfcc.shape[1] for mfcc in mfccs)  # Find the maximum number of coefficients
mfccs_reshaped = [np.pad(mfcc, ((0, 0), (0, max_len-mfcc.shape[1])), mode='constant') for mfcc in mfccs]

```

3) Fig A1-3 shows Agglomerative clustering code in Python, using the sklearn.cluster library's AgglomerativeClustering method on default to finish off DCAP

Fig A1-3: Sample code to showcase Agglo. Clustering on the obtained MFCCs

```
import numpy as np
from sklearn.cluster import AgglomerativeClustering

# Perform agglomerative clustering on the MFCCs
model = AgglomerativeClustering(n_clusters=4)
model.fit(np.concatenate(mfccs_reshaped))

# Get the cluster labels for each segment
labels_VAD = model.labels_

# Print the cluster labels
print("Cluster labels:", labels_VAD)
```

Regarding the VAD-based Agglomerative Clustering:

Has a different use case than DCAP, and does not involve combining multiple smaller segments into larger segments, leading to features being calculated at each segment.

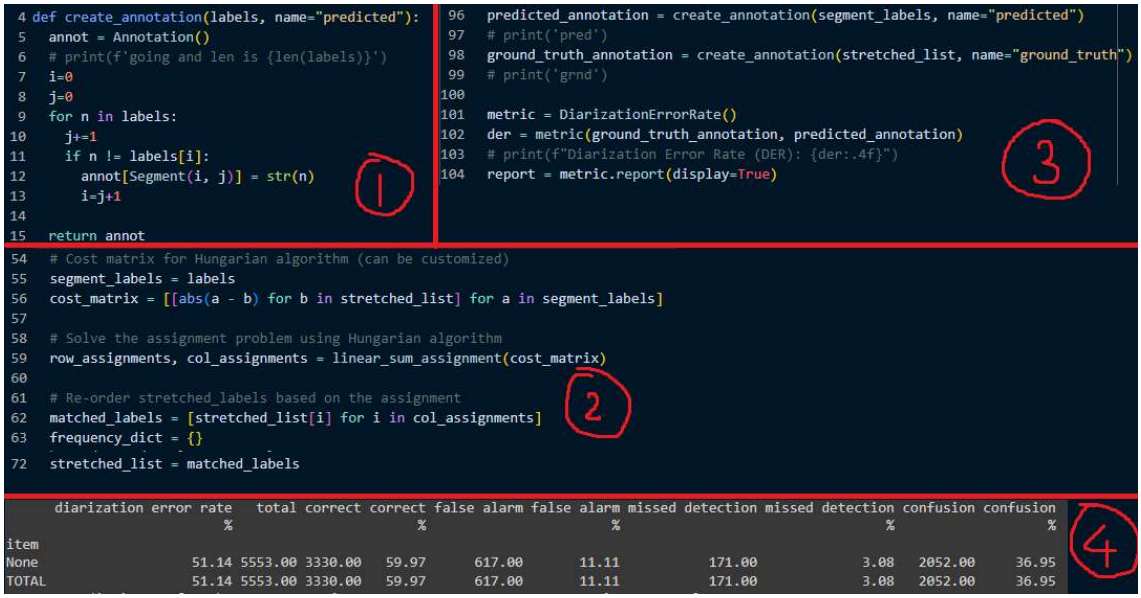
(Larger, combined segments may be more vital in more broadly spread diarization tasks compared to nuanced singer diarization)
Code in Fig A1-4 shows us how to simply implement the same.

Fig A1-4: Simpler Agglomerative clustering with fixed window sizing and MFCC

```python
from sklearn.cluster import AgglomerativeClustering
from librosa import feature, load

def extract_mfccs_agglo(segment, sr):
  """Extracts MFCC features from an audio segment."""
  return feature.mfcc(y=segment, sr=sr, n_mfcc=13)

def segment_and_cluster_agglo(audio_path, n_clusters):
  """Segments audio, extracts MFCCs, and performs agglomerative clustering."""
  # Load audio
  y, sr = load(audio_path)

  # Segment with window size and overlap
  window_size = int(sr * 1)
  overlap = int(window_size * 0.5)
  segments = librosa.util.frame(y, frame_length=window_size, hop_length=overlap)

  # Extract MFCC features from segments
  segment_features = [extract_mfccs_agglo(seg, sr) for seg in segments.T]

  # Flatten features
  flat_features = np.vstack(segment_features)

  # Agglomerative clustering
  clustering = AgglomerativeClustering(n_clusters=n_clusters)
  segment_labels = clustering.fit_predict(flat_features)

  return segment_labels

audio_path = vocals_path
n_clusters = 4
segment_labels_agglo = segment_and_cluster_agglo(audio_path, n_clusters)
```

In the proposed method, a similar approach is followed as mentioned above, with a focus on fine-tuning the number of MFCCs and segment size, while also leveraging other features such as the delta of MFCCs, RMSE, Spectral Contrast, etc.

Fig A1-5: Process of calculating Diarization metrics



```python
def create_annotation(labels, name="predicted"):
    annot = Annotation()
    # print(f'going and len is {len(labels)}')
    i=0
    j=0
    for n in labels:
      j+=1
      if n != labels[i]:
        annot[Segment(i, j)] = str(n)
        i=j+1

    return annot
```

```python
    predicted_annotation = create_annotation(segment_labels, name="predicted")
    # print('pred')
    ground_truth_annotation = create_annotation(stretched_list, name="ground_truth")
    # print('grnd')

    metric = DiarizationErrorRate()
    der = metric(ground_truth_annotation, predicted_annotation)
    # print(f"Diarization Error Rate (DER): {der:.4f}")
    report = metric.report(display=True)
```

```python
    # Cost matrix for Hungarian algorithm (can be customized)
    segment_labels = labels
    cost_matrix = [[abs(a - b) for b in stretched_list] for a in segment_labels]

    # Solve the assignment problem using Hungarian algorithm
    row_assignments, col_assignments = linear_sum_assignment(cost_matrix)

    # Re-order stretched_labels based on the assignment
    matched_labels = [stretched_list[i] for i in col_assignments]
    frequency_dict = {}
    stretched_list = matched_labels
```

	diarization error rate		total	correct	correct	false alarm	false alarm	missed detection	missed detection	confusion	confusion
	%			%		%		%		%	
item											
None	51.14	5553.00	3330.00	59.97	617.00	11.11	171.00	3.08	2052.00	36.95	
TOTAL	51.14	5553.00	3330.00	59.97	617.00	11.11	171.00	3.08	2052.00	36.95	

In Fig A1-5, we see the usage of the pyannote library to create annotations based on cluster labels.
In Part 1 of Fig A1-5, we see the definition of the create_annotation function, following by Part 2 showing the creation of the cost matrix for the implementation of the Hungarian algorithm for label matching; in Part 3 we see the code to return the Diarization metrics, and in Part 4 a sample of the example results.

APPENDIX 2

List of songs used in Dataset creation, in decreasing order of number of singers in ground truth:
Greater than 4, or exactly 4 singers

1) Hikaru Nara (光るなら) - Genshin Chinese VAs [63]
2) Misery x CPR x Reese's Puffs (Extended Version) [64]
3) Fukashigi no KARTE [65]
4) Franchouchou "Saga Jihen" / Hoshimachi Suisei with Hololive Fantasy [66]
5) TMABird - Everybody's Circulation [67]
6) K/DA - VILLAIN ft. Madison Beer and Kim Petras [68]
7) Falling to the Top (ONE LAST HIT) [69]
8) CODE MISTAKE - CORPSE x Bring Me The Horizon [70]
9) Chicken Attack // Song Voyage // Japan // ft. Takeo Ischi [71]
10) Ichiban Ka - Yakuza 7 Main Theme [72]
11) 【グラブル】 三羽烏漢唄 ~ GRANBLUE FANTASY ~ [73]
12) True Damage - GIANTS [74]
13) Tech N9ne - Face Off (feat. Joey Cool, King Iso & Dwayne Johnson) [75]
14) Kirin J Callinan - Big Enough [76]

Exactly 3 prominent singers in the ground truth audio

1) Sleep Tight, Aniki [77]
2) MADKID / RISE [78]
3) FIFTY FIFTY (피프티피프티) - ‘Log in' [79]
4) FIFTY FIFTY (피프티피프티) - 'Cupid' [80]
5) Fitz and the Tantrums - HandClap [81]
6) Dinero (feat. Cerbeus) [82]
7) Goodyear (ONE LAST HIT) [83]
8) HITMAN Anthem (ONE LAST HIT) [84]
9) LEC: Reckless with my heart [85]
10) LEC: Heartbreaker [86]
11) Regression - Honkai Impact 3rd Theme Song Performed by: Ayanga [87]
12) Receive You The Hyperactive [88]
13) Moon Halo - Honkai Impact 3rd Valkyrie Theme [89]
14) 幽霊東京 Ghost City Tokyo ／ Ayase (self cover) [90]

Exactly 2 prominent singers in the ground truth audio

1) 伊藤美来 / No.6 [91]
2) cinnamons × evening cinema - summertime [92]
3) Baka Mitai ばかみたい 【Xlice】[93]
4) Ya Boy Kongming! OP / Chiki Chiki BanBan [94]
5) DAOKO × Kenshi Yonezu “Fireworks” [95]
6) 平行線 - Eve × suis from ヨルシカ [96]
7) E. S. P. - Masayoshi Takanaka (All of Me) [97]
8) Genshin Impact CN VAs - Pinky Swear (勾指起誓) - 宴宁 [98]
9) K/DA - DRUM GO DUM ft. Aluna, Wolftyla, Bekuh BOOM [99]
10) Eli Noir – Wonder Why [100]
11) Original PokeRap from The Pokemon Anime [101]
12) 【Original Genshin Fansong】让风告诉你 (Let the Wind Tell You) [102]

Exactly 1 prominent singer in the ground truth audio

1) 【Ado】レディメイド (Readymade) [103]
2) 24時間シンデレラ [104]
3) 浴槽とネオンテトラ　Cover. LOLUET [105]
4) Sing My Pleasure [106]
5) Caramel [107]
6) ZUTOMAYO - MILABO [108]
7) SEISO - Nyanners [109]
8) BAMO by Konrad OldMoney feat Tonoso and Kartel Sonoro [110]
9) It's okay to envy - Takayan [111]
10) Imagine Dragons x J.I.D - Enemy [112]
11) Aiobahn feat. KOTOKO - INTERNET YAMERO [113]
12) RISE (ft. The Glitch Mob, Mako, and The Word Alive) [114]