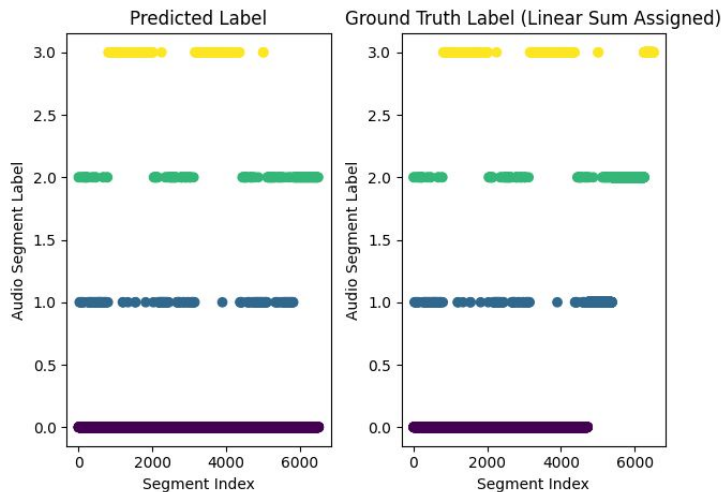

Singer Diarization in Multi Singer Audio

Advait Deochakke - 20BCE1143

Under guidance of Dr. Priyadarshini J

What is Diarization?

In short, “Who Spoke When?”. Does not concern itself with identifying speaker identity outside the scope of the audio. Labelled as Speaker A, Speaker B, etc. instead of eg. John, Harry, and so on.



What is Singer Diarization?

Like speaker diarization but for singing voices, identifies who is singing and when in an audio recording. Has to deal with variation of pitch, backing tracks, high levels of audio distortion, etc.



1. Intro

→ **Identifies count of singers in music**

Reduce complexity in Diarization

→ **Explore and Tackle Challenges**

Distinguishing vocal styles, separating vocals from instruments, etc

→ **Techniques**

MFCCs, Spectral Clustering, Timbre and Chroma scores, Silhouette

→ **Applications**

Music analysis, recommendation Sys, understanding vocal contributions

Problem Statement

Automatic Singer Identification is Hard

- Demonstrate machine learning and deep learning efficacy in automatic singer diarization within music recordings.
- Compare system performance against a baseline to assess accuracy and effectiveness.
- Gather valuable data on technology's potential for precise singer diarization via singer count prediction.
- Lay groundwork for future research, enhancing robustness of singer identification systems.

Research Objectives

- High Accuracy Singer Identification
 - Lower Diarization Error Rate (DER) than baseline system.
 - Identify singers precisely in various music genres (evaluate on dataset).
- Accurate Singer Count Estimation
 - Pinpoint exact number of singers in a recording.
 - Compare predicted singer count to established ground truth.
- Ethical Considerations
 - Mitigate bias in training data to ensure fair identification.
 - Protect user privacy during the identification process.

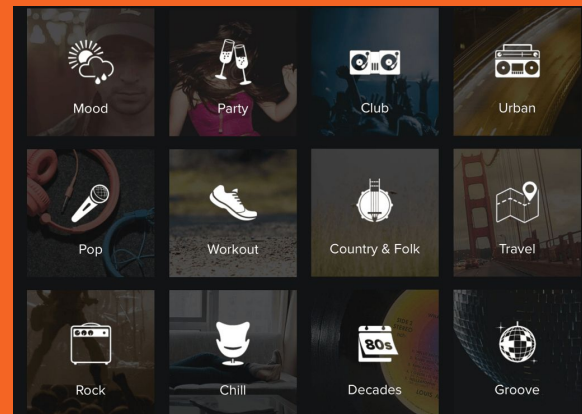
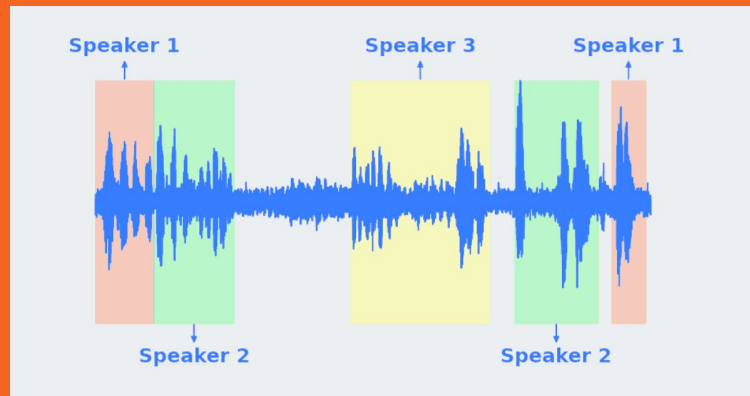
Project Scope

Inclusions

- Dataset Curation
- Vocal Separation
- Feature Extraction
- Singer Counting & Diarization
- Evaluation

Exclusions

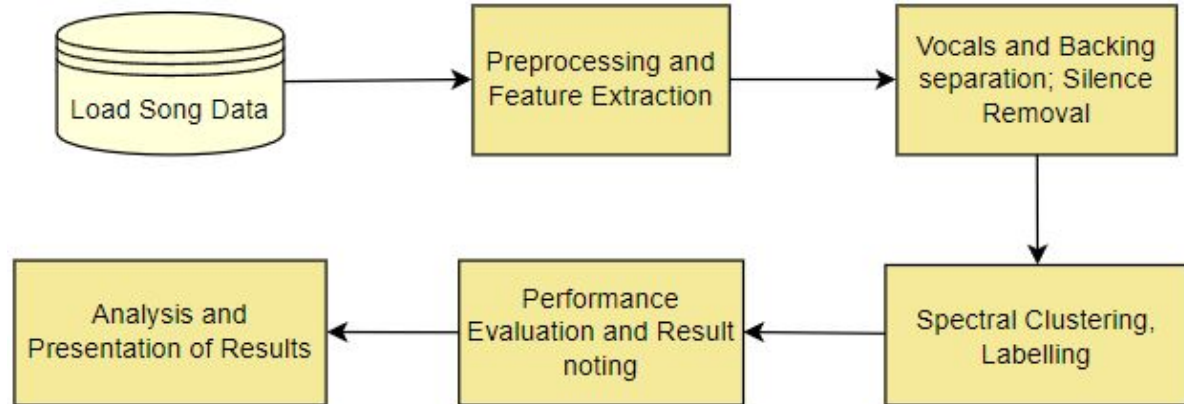
- Real-time Singer Identification
- Music Genre Classification
- Music Recommendation Systems



Proposed System

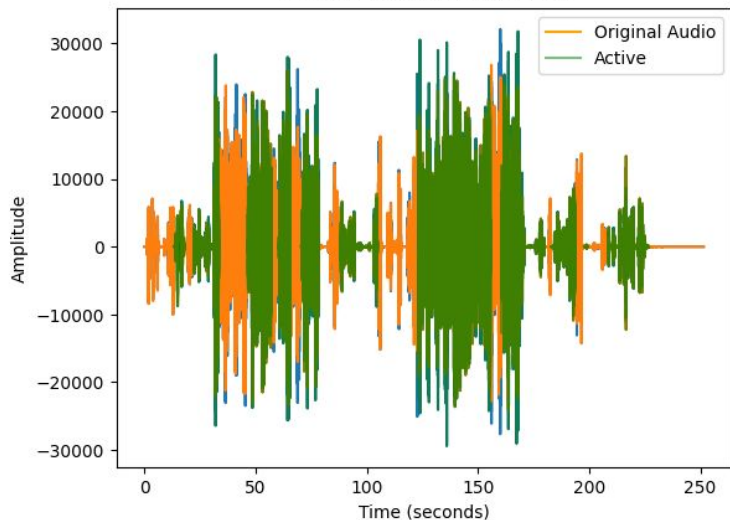
- **Feature extraction captures vocal characteristics:**
MFCCs (timbre, pitch), chroma (vocal range),
delta/delta-delta MFCCs (dynamics).
 - **Spectral clustering for singer count estimation:**
Groups similar vocal segments based on features.
 - **Silhouette score optimization:**
Identifies optimal number of clusters (singers) based on
Silhouette and elbow method.
-

System Diagram



Feature Extraction

Silence vs Presence in Audio

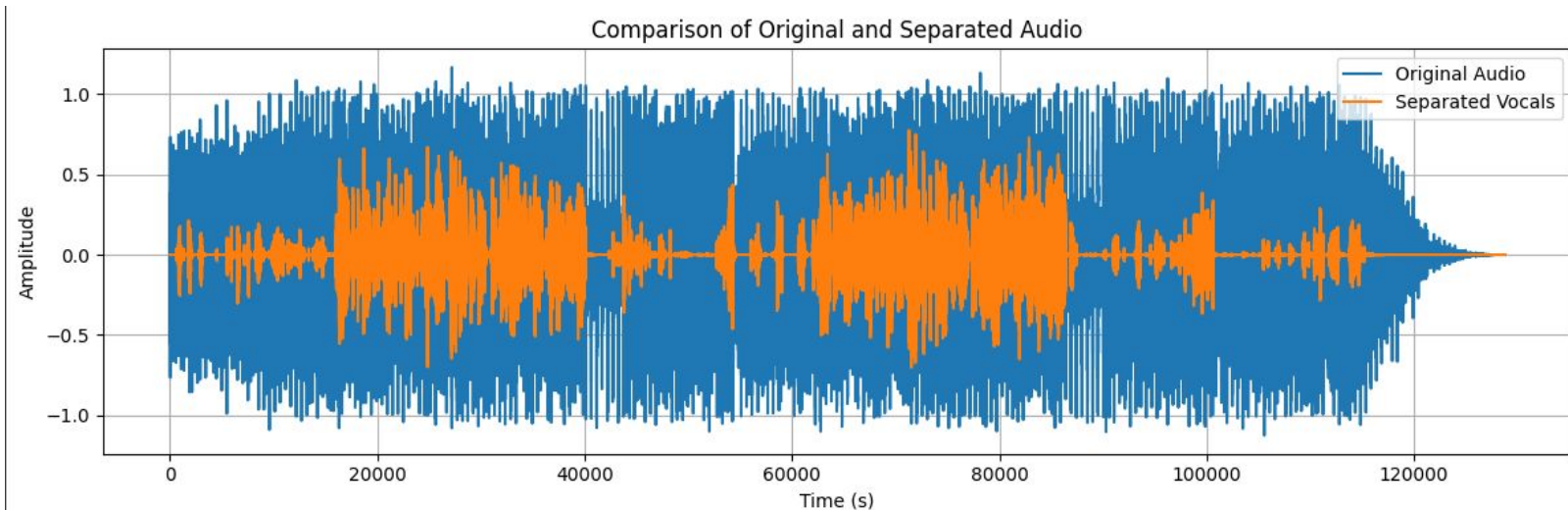


- **Prepares audio data for analysis:**
Removes silence for cleaner audio (improves accuracy).
Isolates vocals using Spleeter (deep learning model).
 - **Silence Removal:**
Identifies silent segments based on low audio energy (Short-Term Energy) & spectral changes (Spectral Flux).
Divides audio into windows for precise removal (preserves short vocal segments).
-

- **Vocal Separation with Spleeter:**

Separates vocals from accompaniment for focused analysis (timbre, pitch, singing style).

2-stem model targets vocals specifically (other models available for instruments).



Features for Singer Diarization

- **Mel-Frequency Cepstral Coefficients (MFCCs):**
Mimic human perception of sound frequencies, capturing the spectral envelope (timbre, pitch).
 - **Chroma Features:**
Represent short-term harmonic content (vocal range, singing style).
 - **Delta and Delta-Delta of MFCCs:**
Capture rate of change in spectral features over time (pitch variations).
 - **Spectral Contrast:**
Measures amplitude differences between peaks and valleys in the audio spectrum (spectral richness, vocal quality).
 - **Root Mean Square Energy (RMSE):**
Quantifies overall energy or loudness (vocal intensity).
-

Predicting Number of Singers: Clustering

Clustering:

Groups audio segments with similar features together.

Each cluster represents a potential singer.

Finding the optimal number of clusters is key to accurate singer count estimation.

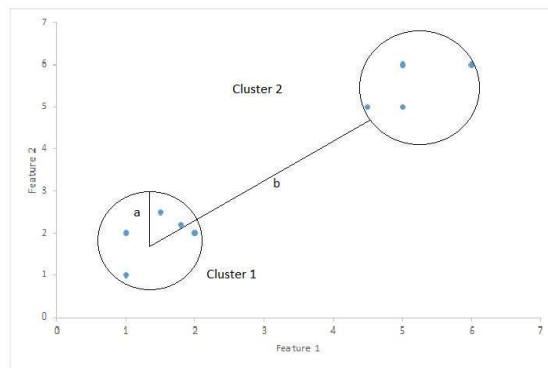
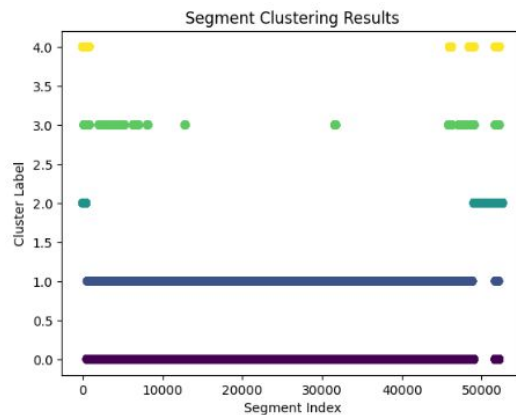
Silhouette Score:

A metric used to assess clustering quality.

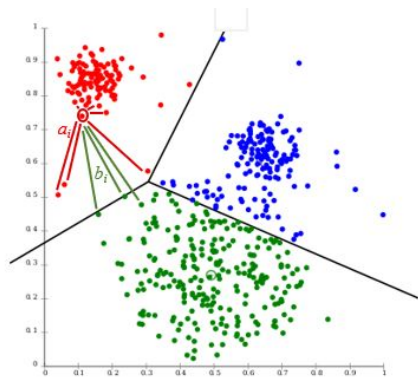
Considers two aspects -

Cohesion within clusters (similarity of features).

Separation between clusters (distinctiveness of singer profiles).



Predicting Number of Singers: Silhouette Score



Iterative Strategy: Explore a range of potential cluster counts.

Spectral Clustering: Groups audio segments based on features in each iteration.

Silhouette Score Calculation: Evaluate the quality of each clustering solution.

Plot Silhouette Scores vs. Number of Clusters: Reveals an "inflection point."

Inflection Point: The optimal number of clusters (singers).

Represents a balance between high intra-cluster cohesion and distinct inter-cluster separation.

Accurate singer count estimation is crucial for effective diarization.

Implementation Details

List of packages required:

Spleeter-2.4.0;
pyAudioAnalysis-0.3.14;
pyannote.audio 3.1.1;

Secondary dependencies:

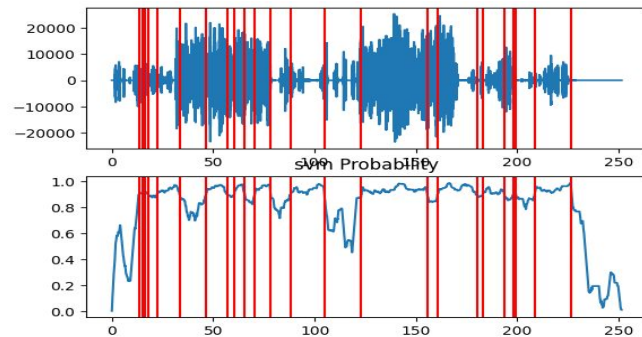
Eyed3-0.9.7; Hmmlearn-0.3.2; pydub-0.25.1; numpy 1.25.2;
Scipy 1.11.4; scikit-learn 1.2.2; tensorflow-2.15.0

Spleeter

- Pre-trained deep learning model designed for audio source separation
- Script provides the audio file path to Spleeter.
- Separator object is initialized with the chosen model configuration (e.g., "spleeter:2stems" for vocals and accompaniment separation).

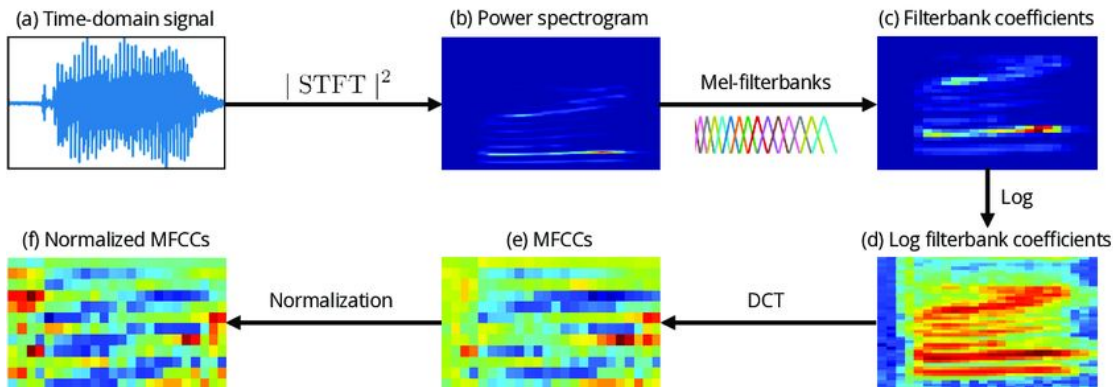
Silence Removal

- Uses pyAudioAnalysis library
- Removes silent segments from vocals
- Configurable parameters for silence detection (duration, smoothing)
- Outputs a list of segments representing non-silent vocal parts.



Extracting Informative Features

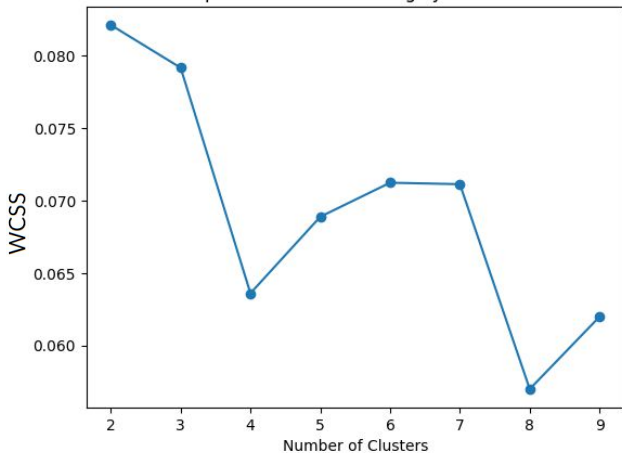
- Captures unique vocal characteristics using librosa.
- Extracts features like:
 - Mel-Frequency Cepstral Coefficients (MFCCs) - timbre, pitch
 - Chroma features - vocal range, singing style
 - Delta and Delta-Delta of MFCCs - vocal changes over time
- These features provide a comprehensive representation for singer identification.



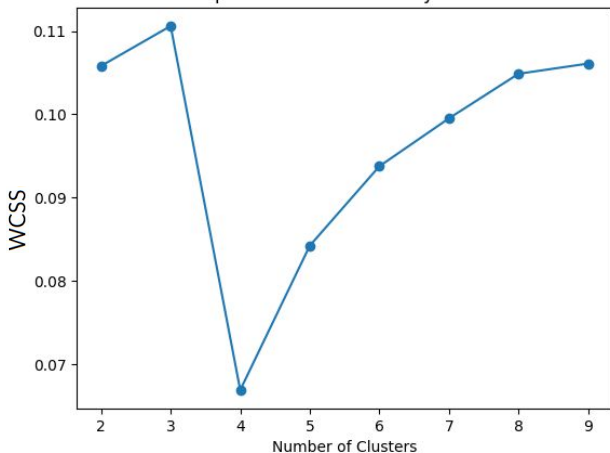
Identifying Singers (Part 1)

- **Extracts features** (MFCCs, chroma) to capture vocal characteristics.
 - **Groups similar segments** using spectral clustering for singers with shared vocal qualities.
 - **Finds optimal number of clusters** (singers):
 - Iterates through possible cluster counts.
 - Performs spectral clustering for each count.
 - Analyzes silhouette scores to identify the "elbow point" - the most likely number of singers.
-

Optimal Clusters for SingMyPleasure



Optimal Clusters for PinkySwear



Identifying Singers (Part 2)

Elbow method for optimal cluster count:

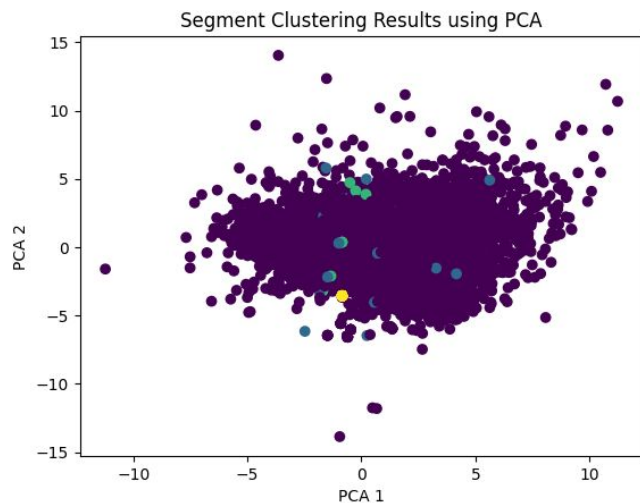
- Runs spectral clustering for varying cluster counts (potential singers).
- "Elbow point" in the WCSS (sum of squared distances) vs. cluster count plot signifies the best number of singers.
- Balances model complexity and clustering performance.

WCSS curve analysis:

- Lower WCSS indicates better data point representation within clusters.
- Flattening curve suggests minimal gain from adding more clusters (avoiding overfitting).

Segmenting Vocals and Assigning Singers

Divides vocals into overlapping windows (1-1.3s with 50% overlap) to capture singer characteristics within each segment.

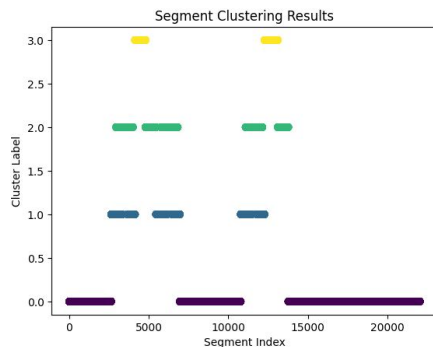
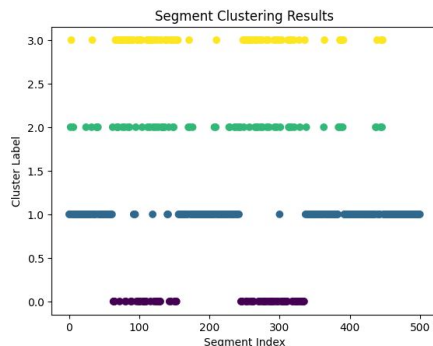
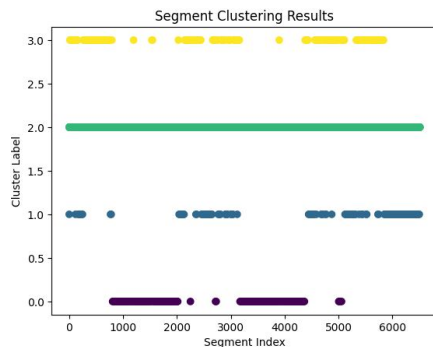


Visualizing Clusters (PCA):

- Reduces feature dimensions for visualization using Principal Component Analysis (PCA).
- Not ideal in this case - PCA captured low variance (less than 50% with 10 components).

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%



Segmenting Vocals and Assigning Singers (continued)

Spectral Clustering for Grouping:

- Groups similar segments based on features (potentially from the same singer).
- Creates a scatter plot where each point (segment) is colored by its assigned cluster.
- Segments with the same color likely belong to the same singer.

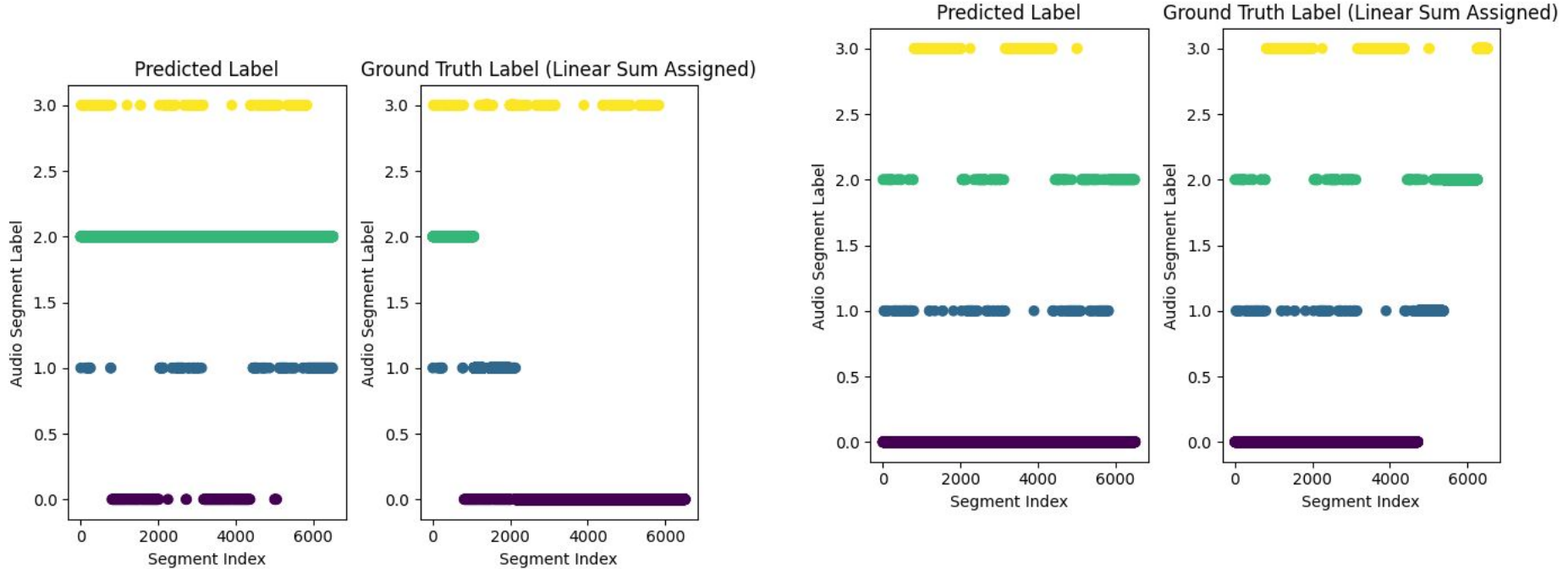
Manual Labeling:

- Human listeners identify the singer's voice in each segment.
 - Creates a reference set of labeled segments for each singer.
 - Used to evaluate the system's performance.
-

Why Hungarian Algorithm for Speaker Diarization?

- **Assigning Speaker Labels Accurately is Key:**
 - Need one-to-one correspondence between speech segments and speaker identities.
 - **Hungarian Algorithm: The Optimal Assignment Tool**
 - Solves the assignment problem: matching segments (potential speakers) to labels (identities).
 - Minimizes the total cost of assigning segments to labels
 - **Cost Function for Speaker Diarization:**
 - Considers acoustic feature similarity between segments and speaker profiles.
 - Lower cost for segments with high similarity to a specific speaker.
 - **Finding the Optimal Matching:**
 - Hungarian algorithm explores all assignments, evaluating costs.
 - Eliminates inefficient assignments and finds lower-cost configurations.
 - Reaches a minimum total cost assignment for all segments.
-

	diarization error rate	total correct		correct	false alarm	false alarm	missed detection	missed detection	confusion	confusion
	%			%		%		%		%
item										
None	24.98	5865.00	4628.00	78.91	228.00	3.89	96.00	1.64	1141.00	19.45
TOTAL	24.98	5865.00	4628.00	78.91	228.00	3.89	96.00	1.64	1141.00	19.45



	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

Metrics for Evaluation

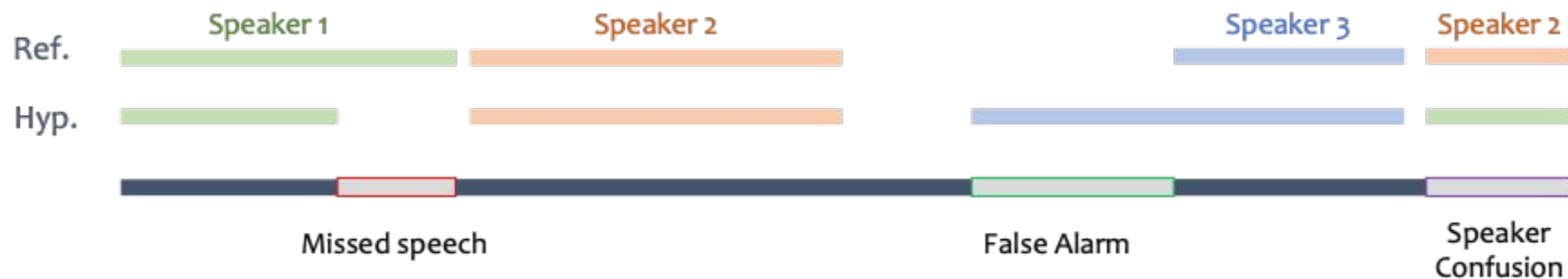
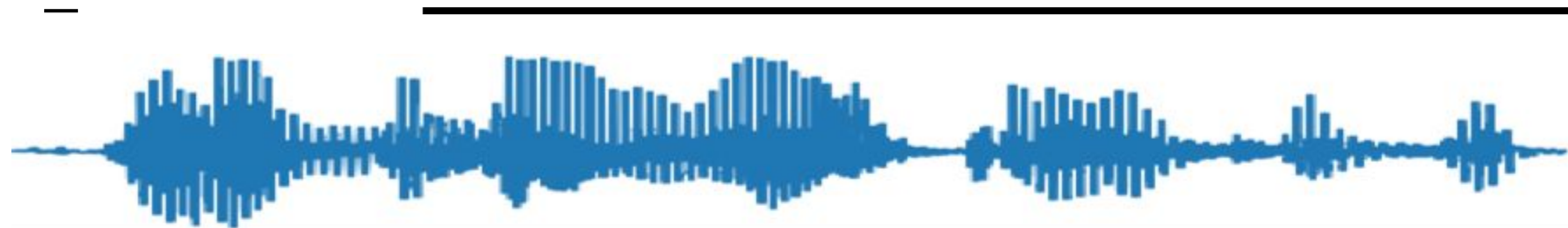
Metrics:

- **False Alarm:** System identifies silence/non-speech audio as a singer.
- **Missed Detection:** System fails to identify a speech segment (singer).
- **Speaker Confusion:** System assigns a segment to the wrong singer.

Diarization Error Rate (DER):

- Combines these errors as a percentage of total speech duration.
- Lower DER indicates better diarization performance.
- Singer diarization DER can range from 30% to 50% or higher depending on methodology and dataset.

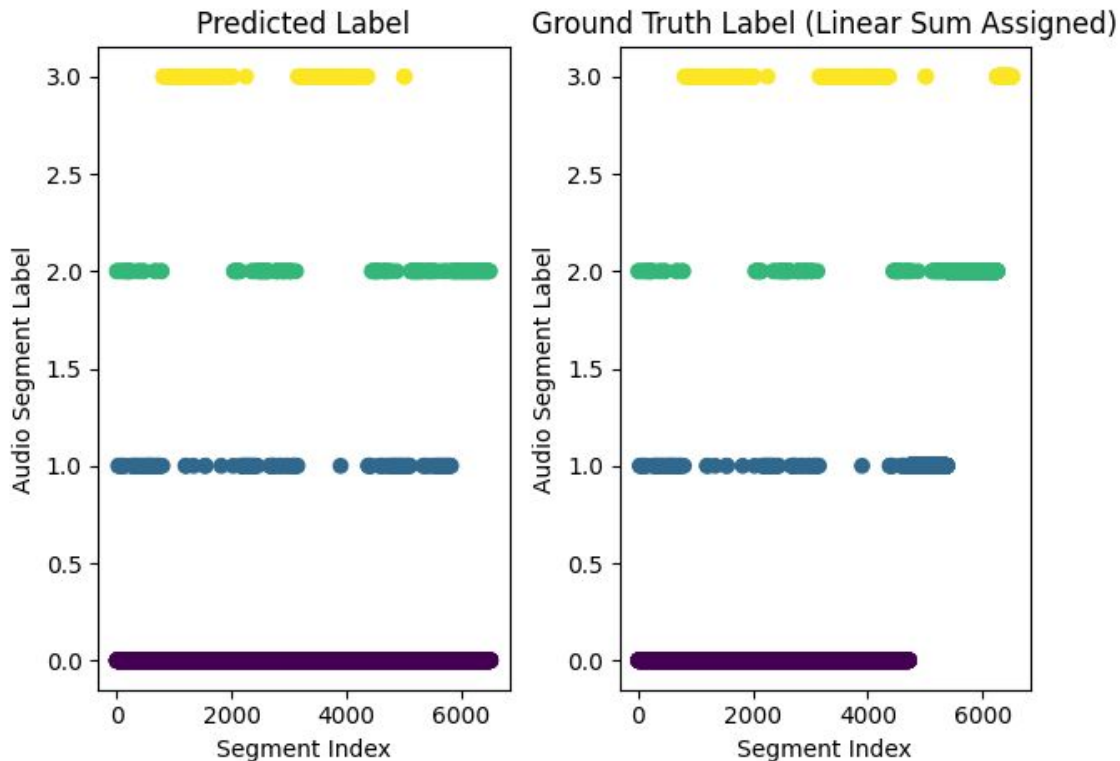
$$\text{DER} = \frac{\text{false alarm} + \text{missed detection} + \text{confusion}}{\text{total}}$$



$$\text{DER} = \frac{\text{Missed speech} + \text{False Alarm} + \text{Speaker Confusion}}{\text{Total length}}$$

Results

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 based on the DER metric.



Singer Prediction (by No. of Singers)

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%

- Songs are included from a diverse number of genres
- Instances of both overestimation and underestimation of singer count occur
- Both errors are balanced and bias is not observed

Overall Rate is 58.49%

Wide diff. In rates shows that a more detailed investigation is necessary into the reason behind it

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

Diarization Breakdown

- Due to Silence Removal and Vocal Separation, all show very low Miss Rate
- False Alarm seems to be alarming high for DCAP, which may be explained by difference in original purpose and dataset
- Singer Error is ~moderate across the board

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%

The authors in the DCAP paper have a DER of 17.8% and 43.1% on the dataset and purpose used by them.

The authors in the VAD Agglomerative Clustering paper show a DER of 52.9%, Singer Error of 10.6%, FA of 26.5%, and Miss rate of 15.8% for closest comparison of conditions

Discussion/Summary

- **Successful identifications:** System correctly identified the number of singers in 53 recordings.
 - Highlights system's ability to handle diverse scenarios.
 - **Room for improvement:** In some cases, the system overestimated (13) or underestimated (9) singers.
 - Solo and 4+ singer recordings pose challenges (low accuracy).
 - **Overall evaluation:**
 - Balanced error rates indicate improvement compared to the baselines which exist.
 - **Future directions:**
 - Classification report and singer count results offer valuable insights.
 - Further exploration and refinement can lead to more robust speaker identification.
-

Conclusions and Future Work

- **Achievements:**

- Improved Speaker Identification: Achieved lower DER compared to baseline systems through deep learning for vocal separation, feature extraction, and spectral clustering.
- Optimal Speaker Labeling: Hungarian algorithm ensured accurate label assignment based on acoustic similarity.

- **Areas for Improvement:**

- Singer Misclassification: Reduce Singer Error Rate through more sophisticated speaker profiling or alternative clustering for singers with similar voices.
- Complex Scenarios: Explore alternative features or incorporate music genre information to improve identification in solo recordings and those with many singers.

- This research presents a promising system for automatic singer identification with room for further development.
 - Addressing limitations like singer differentiation and complex scenarios can lead to a valuable tool for music analysis and classification.
 - Future work focused on speaker profiling, clustering algorithms, and music genre information can enhance the system's robustness and generalizability.
-

Confirmation of submission to SCOPUS/equivalent indexed journals

ICCIDE-2024 submission 14  Inbox x



ICCIDE-2024 <iccide2024@easychair.org>

Mon, 15 Apr, 18:20

to me ▾

Dear authors,

We received your submission to ICCIDE-2024 (3rd INTERNATIONAL
CONFERENCE ON COMPUTATIONAL INTELLIGENCE AND DATA ENGINEERING):

Authors : Advait Deochakke and Priyadarshini Jayaraju

Title : Singer Diarization In Multi Singer Audio

Number : 14

The submission was uploaded by Advait Deochakke
<advaitdeochakke@gmail.com>. You can access it via the ICCIDE-2024
EasyChair Web page


<https://easychair.org/conferences/?conf=iccide2024>

Thank you for submitting to ICCIDE-2024.

Best regards,
EasyChair for ICCIDE-2024.

Guide Approval Mail Snapshot

Regarding Approval for Final Thesis Presentation Inbox x


 **ADVAIT DEOCHAAKKE 20BCE1143** <advait.deochaakke2020@vitstudent.ac.in>
to Priyadarshini ▾



Tue, Apr 23, 3:20 PM (2 days ago) ☆ ↶ ⋮


Hello M'am,
I have made the changes requested to the ppt, and updated it accordingly.
Requesting for approval for final presentation.

Presentation is attached in this email


—
[Advait Deochakke](#)
20BCE1143
(Mobile: 7038777539)

2 Attachments • Scanned by Gmail 



 **ADVAIT DEOCHAAKKE 20BCE1143**
Maam can you please update on the same, as review is tmrw morning. Thank you

Wed, Apr 24, 8:56 PM (11 hours ago) ☆

 **Priyadarshini J**
to me ▾

8:27 AM (9 minutes ago) ☆ ↶ ⋮

Approved

References (short list)

- [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*.
-

References (contd.)

- [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 Analysis (FMA 2015) (pp. pp-124).
-