SPEAKER IDENTIFICATION AND VERIFICATION

Alli Khadga Jyoth, Manish Vutkoori, Bikram Majhi
Indian Institute of Technology
Jodhpur
{m23csa003, m23csa014, m23csa007}@iitj.ac.in
GitHub: https://github.com/KhadgaA/Speaker-Identification

ABSTRACT

The effectiveness and efficiency of a speech signal-based framework rely on the quality of its distinctive features. In the current era of research, relying just on the functioning of just one feature may not be sufficient to achieve both resilience and performance concurrently. To address this issue, researchers utilize multiple sources and employ diverse fusion approaches. The past study materials indicate that strengths and recognition rate of systems are improved by utilizing the features provided by various sources. Although these fusion techniques improve strengths and identification rate of system, they also have certain drawbacks in the system. The objective of this project is to implement a multilingual speaker system by utilizing various machine learning models and fusion techniques. The goal is to examine the benefits of different fusion strategies and their utility in constructing an efficient system for a multilingual speaker identification and verification system.

1 Introduction

Speaker identification and verification falls under the broader umbrella of speech recognition, encompassing two primary types: speech recognition and speaker recognition. Both involve extracting essential information from speech signals and leveraging machine-based techniques for identification purposes. Speaker recognition specifically aims to extract information from speech signals based on predefined criteria, whereas speech recognition focuses solely on extracting textual information from speech signals. These systems share similarities with pattern recognition methodologies. The accuracy of speaker identification systems heavily relies on the discriminative capabilities of the features utilized in the process. Feature extraction methods are typically tailored to the specific requirements of speaker identification tasks. In speaker identification and verification, computers often employ techniques such as linear prediction cepstral coefficients (LPCC) or mel-frequency cepstral coefficients (MFCC) to derive precise features that characterize the vocal information of the speaker [[1] [2] [3]].

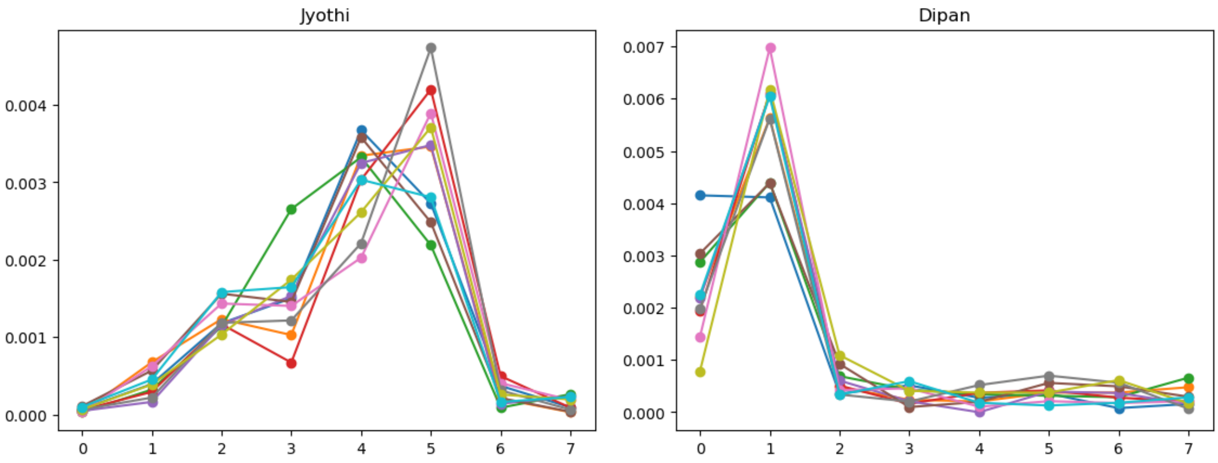
Effective feature extraction techniques that are resistant to noise are essential for the successful deployment of accurate Speaker Identification systems. An optimal feature vector should contain precise information extracted from the recorded speech signal and should be resistant to noise. Therefore, the advancement of techniques for extracting features that are resistant to noise has been a prominent focus of study in Automatic Speech Recognition (ASR) and Identification for the past fifty years [[4] [5] [6]]. Linear predictive cepstral coefficients (LPCC) , MFCC, perceptual linear prediction (PLP), and wavelets are feature extraction methods that have been proposed by scientists over the past sixty years. Among the suggested methods, MFCC is renowned for its superior accuracy in voice recognition systems. The computation overhead of MFCC is negligible and it operates effectively in clean surroundings. Nevertheless, it exhibits poor performance when confronted with the presence of additional noise. Therefore, the limited ability of MFCC to handle background noise has led to the adoption of noise resilient features. The GFCC [[4]] and BFCC [[5],[7]] are referenced. Each of the three strategies use distinct filterbanks for the purpose of feature extraction. MFCC stands for Mel-frequency cepstral coefficients. It utilizes a Mel-scale filter-bank with 13 filters. GFCC, on the other hand, uses a Gammatone filterbank with 18 filters. BFCC uses a Gammachirp filter-bank, also with 18 filters. The GFCC system is specifically designed to replicate the functioning of the human auditory system. It achieves this by utilizing a set of non-linear filters to analyze speech signals . Nevertheless, MFCC employs a set of linear filters for the purpose of analyzing speech signals. Therefore, researchers have utilized the noise characteristics of the GFCC and BFCC features to create speech recognition and identification systems [[4],[5],[7],[8], [9], [1]].

While above feature extraction technique have proven to be highly effective in speech identification and verification systems, recent research proposes the use of constant Q cepstral coefficients (CQCC) as a new approach that offers a variable time-frequency resolution. The constant Q transform (CQT) is a time-frequency analysis method that outperforms standard methods like the discrete Fourier transform (DFT) in terms of frequency resolution at lower frequencies and time resolution at higher frequencies. This leads to enhanced spectrum modeling [[10]]. Linear Frequency Cepstral Coefficients (LFCCs) are efficient in identifying high-frequency sounds because they employ a linear filter bank. This approach provides a harmonious combination of gathering high-frequency data and processing efficiency in comparison to MFCCs [[3]]. Modulation Spectral Cepstral Coefficients (MSRCCs) aim to capture the changing patterns in the frequency content of speech sounds. Adding this supplementary information to conventional cepstral coefficients can improve the discriminatory capability of speaker identification algorithms [[11]]. Nonlinear Gammatone Cepstral Coefficients enhance cepstral coefficients (NGCCs) works by integrating non-linear modifications that are influenced by the human auditory system. The objective of this approach is to extract speaker-specific information that is encoded in speech signals with greater efficiency .

The voice recognition system is speaker independent, meaning it does not rely on specific individuals. However, it requires a substantial amount of data to accurately represent phoneme-based information. In order to alleviate the complexities, individuals rely on a substantial amount of information. Tripathi et al. introduced various types of source information and integrated them with MFCC characteristics utilizing specified fusion techniques. Additionally, they have demonstrated that the integration of source information and MFCC features not only increases the accuracy rate, but also enhances the resilience of the phoneme recognition process.

1.1 Our Approach

The performance of the fusion system relies upon both the efficiency of feature extraction and the selection of fusion methods. The optimized result can be achieved by the utilization of appropriate fusion and efficient characteristics. Optimizing performance in speech processing frameworks mostly relies on the fusion technique employed, which heavily depends on the excitation source and MFCC information [[7], [8]]. The speech sample is processed to serve as input for the feature extraction stage in the pre-processing phase. The goal of the feature extraction stage is to compute the necessary features by employing diverse signal processing techniques. In feature level fusion, various features are calculated and concatenated for creating feature embeddings. The test features are calculated using a similar manner and are then utilized for matching. In case of Model based fusion, various models are created using individual feature sets. Further, the different models parameters are combined to create composite models. We have used grid search for getting the optimal hyperparamters. Finally, the comparison is created with test speech specimen and composite . In score level fusion, different characteristics are obtained from given voice signal and used to create the corresponding embeddings. During matching, the given features are matched with corresponding embeddings, and calculate individual score. These score are combined to give final score. Previous research demonstrates that instead of utilizing individual features,



(a) Female dominant freqs

(b) Male dominant freqs

Figure 1: Comparing Dominant frequencies

combining numerous features produces optimum classifications for tasks involving the identification of speech-based patterns. Furthermore, the integration of different characteristics not only strengthens the resilience but also improves the efficiency of the systems. Recent researchers have demonstrated the advantages of many characteristics in speech recognition systems, particularly for automated speech recognition and identification systems. These studies only utilize fusion-based approaches to combine features at each level. These strategies possess their own merits and drawbacks. A synergistic fusion technique might be developed by leveraging the benefits of individual combination schemes, resulting in an effective, efficient, and versatile solution for various speech processing systems. The objective is to analyze the benefits of various speaker recognition and identification systems and utilize them to enhance the efficiency of the recognition scheme. The primary discoveries of the research endeavor are as follows:

2 Methodology

Our approach introduces a hierarchical structure, first we extract the features using different methods like the Mel-frequency Cepstral Coefficients (MFCCs), BFCCs, etc. Then these features are concatinated to give a final feature vector of length $N \times d$ where N is the number of different types of featuers and d is the number of cepstral Coefficients. The different feature extraction methods are given in the next section. Cepstral Coefficents operate on spectrograms, but in conjunction with cepstrals we also utilize frequency based Dominant Frequency feature. The Dominant frequencies are the frequencies which are observed dominantly in a audio sample. From our observations as seen in the figs 1 From the figure we can see a clear difference in the dominant frequencies of male vs female.

3 Techniques of feature extraction

3.1 Mel-Frequency Cepstral Coefficients (MFCCs)

Mel-Frequency Cepstral Coefficients (MFCCs) are essential in processing speech and audio signals for tasks like speech recognition and speaker identification.[12]

The MFCC computation comprises several steps:

1. a pre-emphasis filter enhances higher frequencies in the signal.
2. the signal is divided into short frames, typically 20-40 milliseconds long, with overlap. Each frame undergoes windowing to mitigate spectral leakage.
3. the Fast Fourier Transform (FFT) converts each windowed frame into the frequency domain.
4. Mel filterbanks, spaced uniformly on the Mel scale, are applied to the power spectrum, yielding filterbank energies. These energies are logarithmically scaled to mimic human loudness perception.
5. Finally, the Discrete Cosine Transform (DCT) decorrelates the log filterbank energies, resulting in MFCCs.

$$\text{MFCCs}(n) = \sum_{m=0}^{M-1} \log \left(\sum_{k=0}^{N-1} |X(k)|^2 H_m(k) \cos \left[\frac{\pi}{N} (k + \frac{1}{2}) n \right] \right) \tag{1}$$

where:

- N is the number of samples in each frame,
- $X(k)$ is the discrete Fourier transform of the k-th frame,
- $H_m(k)$ is the m-th Mel filter,
- M is the number of Mel filters, and
- n is the index of the MFCC coefficient.

3.2 BFCC (Bandpass Filtered Cepstral Coefficients)

BFCC (Bandpass Filtered Cepstral Coefficients) is a method widely used in audio signal processing, particularly for tasks like speech recognition and music genre classification. It extends the Mel Frequency Cepstral Coefficients (MFCC) approach by incorporating bandpass filtering to better capture spectral characteristics[13].

1. **Mel-Frequency Analysis**: MFCC computation involves pre-emphasis, framing, FFT, Mel filter bank application, logarithmic compression, and DCT.
2. **Bandpass Filtering**: BFCC introduces bandpass filters designed to extract specific frequency bands relevant to the task. These filters are applied to the MFCC coefficients to obtain bandpass-filtered cepstral coefficients.
3. **Normalization and Delta Features**: BFCC coefficients are typically normalized and supplemented with delta and delta-delta coefficients to account for variations in signal amplitude and capture temporal information.

3.3 Constant-Q Chroma Cepstral Coefficients (CQCC)

CQCC is a popular technique for audio analysis, particularly in tasks like music information retrieval and speech recognition.

1. **Constant-Q Transform (CQT):** This Fourier transform variant adapts to the nonlinear frequency resolution of the human auditory system. It computes the CQT of the input signal using the equation2

$$X_c(\omega) = \sum_{n=-\infty}^{\infty} x[n] g_c[n] e^{-j\omega n} \tag{2}$$

Variables:

$X_c(\omega)$: CQT of the signal.
$x[n]$: Input signal.
$g_c[n]$: Complex-valued window function.
ω : Angular frequency.

2. **Chroma Energy Normalized Difference Vector (CEN):** CEN measures the energy difference between adjacent chroma vectors and normalizes it. It's calculated as3

$$CEN_i = \sum_{j=1}^{12} |X_i(j) - X_{i-1}(j)| \tag{3}$$

Variables:

CEN_i : CEN value for the i-th frame.
$X_i(j)$: Energy of the j-th chroma component in the i-th frame.
$X_{i-1}(j)$: Energy of the j-th chroma component in the previous frame.

3. **Cepstral Coefficients:** These coefficients, derived from the Discrete Cosine Transform (DCT) of the logarithm of the power spectrum, capture audio signal characteristics in the frequency domain. The equation for cepstral coefficients4

$$C_q = \sum_{m=0}^{M-1} \log(|X_q(m)|^2) \cos\left[\frac{\pi}{M}(m+0.5)q\right] \tag{4}$$

Variables:

C_q : The q-th cepstral coefficient.
$X_q(m)$: Magnitude spectrum of the q-th sub-band obtained from the CQT.
M : Number of cepstral coefficients.
q : Index of the cepstral coefficient.

3.4 GFCC (Gammatone Frequency Cepstral Coefficients)

GFCC (Gammatone Frequency Cepstral Coefficients) are pivotal in audio signal analysis, especially in discerning speech and music.[14]

1. Initially, a gammatone filterbank is employed, replicating the human auditory system's frequency processing. This filterbank comprises logarithmically spaced filters, each capturing specific frequency bands through convolution with the input signal.

2. A logarithmic compression step models the nonlinear response of the human ear to sound intensity, ensuring the filterbank's outputs resemble the auditory spectrum.
3. The Discrete Cosine Transform (DCT) is applied to the compressed filterbank outputs, aiding in decorrelating features and condensing information. The resulting DCT coefficients are representative of the cepstral domain, offering insights into the audio's characteristics.
4. A selection process is then conducted to pick relevant DCT coefficients, facilitating dimensionality reduction while retaining discriminative information.

These steps collectively yield GFCC feature vectors instrumental in tasks like speech recognition and music genre classification.

3.5 LFCC (Log Filterbank Energies)

LFCC (Log Filterbank Energies) features are extensively utilized in audio processing, particularly in speech recognition. These features stem from Mel-frequency cepstral coefficients (MFCCs), a prevalent tool in speech processing systems.[13]

1. **Pre-processing** involves segmenting the audio signal into frames and applying windowing functions like Hamming or Hanning to mitigate spectral leakage. Following this, the magnitude spectrum of each frame is computed using FFT (Fast Fourier Transform).
2. **Mel Filterbank** applies Mel filters to the magnitude spectrum, yielding filterbank energies representing energy within specific frequency bands.5

$$H_m(f) = \begin{cases} 0 & \text{if } f < f_m \\ \frac{f-f_{m-1}}{f_m-f_{m-1}} & \text{if } f_{m-1} \leq f \leq f_m \\ \frac{f_{m+1}-f}{f_{m+1}-f_m} & \text{if } f_m \leq f \leq f_{m+1} \\ 0 & \text{if } f > f_{m+1} \end{cases} \tag{5}$$

3. **Logarithmic Compression** is applied by taking the logarithm of the filterbank energies, simulating human perception of sound intensity.6

$$L_m = \log\left(\sum_{k=1}^{N} |X(k)|^2 H_m(f_k)\right) \tag{6}$$

4. **Discrete Cosine Transform (DCT)** is employed to decorrelate the log filterbank energies and extract a compact representation, resulting in LFCC features.7

$$C_n = \sum_{m=1}^{M} \left(L_m \cdot \cos\left[\frac{\pi}{M} \cdot \left(m - \frac{1}{2}\right) \cdot n\right]\right) \tag{7}$$

3.6 MSRCC (Minimum Squared Residue Cross-Correlation)

The MSRCC (Minimum Squared Residue Cross-Correlation) is a method primarily utilized for audio signal processing, particularly for tasks such as source separation and localization. It is commonly employed in scenarios where multiple sound sources are present and need to be separated or localized from a mixture.

MSRCC is based on minimizing the squared residue between the observed mixture and the estimated source signals. It operates on the cross-correlation matrix of the observed mixture signals.

Cross-Correlation Matrix (R):

$$R = XX^T$$

Where X is the matrix containing the observed mixture signals.

MSRCC Objective Function:

$$J(\hat{X}) = \|R - \hat{X}\hat{X}^T\|_F^2$$

Where $\hat{X}$ is the estimated source signals matrix and $\|\cdot\|_F$ denotes the Frobenius norm.

Optimization: The objective function $J(\hat{X})$ is minimized to estimate the source signals $\hat{X}$.

Usage in Audio Processing: MSRCC can be applied to various audio processing tasks, including:

- Source Separation: Given a mixture of audio signals, MSRCC can estimate the individual source signals.
- Source Localization: By analyzing the cross-correlation matrix, MSRCC can help localize the spatial positions of sound sources.

3.7 Neural Generative Contrastive Coding(NGCC)

NGCC, short for Neural Generative Contrastive Coding, is a technique rooted in contrastive learning principles, primarily used in audio signal processing tasks like audio generation, denoising, and source separation. It leverages neural networks, typically CNNs or RNNs, to encode audio data into a compact representation in a latent space, where similar audio samples cluster together. The model is trained with a contrastive loss function, such as InfoNCE, which aims to maximize agreement between positive pairs (similar samples) and minimize it between negative pairs (dissimilar samples). Negative samples are generated to facilitate efficient training, often through augmentation or sampling from different distributions.[15]

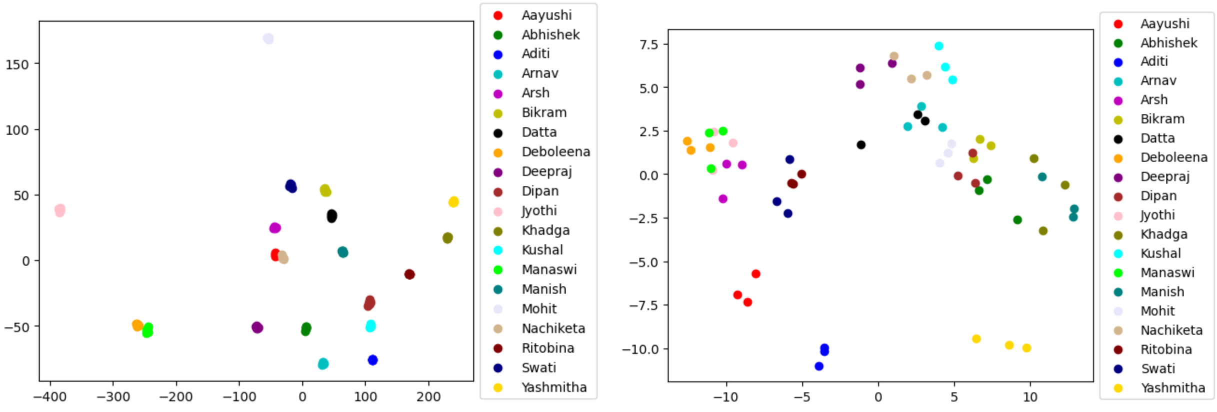
NGCC finds application in various audio tasks: generating new audio, denoising by separating clean signals from noisy ones, and source separation, distinguishing different audio sources in mixed signals. Evaluation metrics like SNR, PESQ, or MSE gauge its performance.

4 Experiments and results

We evaluate the performance of various machine learning models using precision, recall, F1-score, and accuracy metrics. The Classifier Performance table reveals a significant variation in the performance of the models.

Classifier	**Performance (at 27 samples)**			
	Precision	**Recall**	**F1-Score**	**Accuracy**
KNeighborsClassifier	0.41	0.43	0.39	0.43
LogisticRegression	0.63	0.65	0.64	0.65
DecisionTreeClassifier	0.70	0.68	0.66	0.68
SVC	0.97	0.97	0.97	0.97
NuSVC	0.97	0.97	0.97	0.97
RandomForestClassifier	0.97	0.97	0.97	0.97
AdaBoostClassifier	0.92	0.92	0.91	0.92
GradientBoostingClassifier	0.97	0.97	0.97	0.97
SGDClassifier	0.97	0.97	0.97	0.97
GaussianNB	0.89	0.82	0.82	0.82
MultinomialNB	0.98	0.98	0.98	0.98

We have created a custom multilingual dataset and extract features using seven distinct techniques. Instead of relying on a single feature extraction method, we concatenate the embeddings obtained from these techniques to create a comprehensive and rich feature representation for each speaker.



(a) LDA on Training data

(b) LDA on Test data

Figure 2: LDA feature dimensionality reduction

The rationale behind concatenating the embeddings is to leverage the strengths of each feature extraction technique and provide a more holistic representation of the speaker's characteristics. This approach allows us to capture a wide range of speaker-specific information, including spectral, prosodic, and voice quality features, which are crucial for accurate speaker identification in a multilingual context. To evaluate our feature extraction methods we show LDA plots in 2, from the figure it is evident that our feature extraction pipeline is able to effectively extract language independent and speaker specific features, which are essential for any speaker identification system.

The performance of the models is evaluated based on precision, recall, and F1-score metrics. As shown in the Classifier Performance table, most models achieve impressive results, with some models significantly outperforming others.

4.1 Comparisons and Possible Reasons for Performance Differences

Various models exhibit different performance levels due to their unique characteristics and abilities to handle the complexities of the task. Here, we provide reasons why certain models perform better than others in the context of multilingual audio detection.

1. **MultinomialNB:** The MultinomialNB model achieves the highest performance in our multilingual speaker identification task. This can be attributed to its ability to effectively handle discrete feature representations, which are obtained by concatenating embeddings from various feature extraction techniques.
2. **SVC, NuSVC, RandomForestClassifier, GradientBoostingClassifier, and SGDClassifier:** These models achieve impressive performance, with F1-scores of 0.97 and accuracies of 0.97. They are all capable of handling high-dimensional feature spaces and capturing complex patterns in the data.
3. **AdaBoostClassifier:** The AdaBoostClassifier achieves relatively high performance but lags behind the top-performing models. This model might be less effective in handling the complexities of the multilingual speaker identification task compared to other ensemble methods.
4. **DecisionTreeClassifier:** The DecisionTreeClassifier achieves moderate performance with an F1-score of 0.66 and an accuracy of 0.68. This model might capture some underlying patterns in the concatenated embeddings, but it might not be as robust or consistent as other models.
5. **LogisticRegression:** The LogisticRegression model achieves moderate performance with an F1-score of 0.64 and an accuracy of 0.65. This model might be less effective in capturing complex patterns in the high-dimensional feature space resulting from the concatenation of multiple embeddings.
6. **KNeighborsClassifier:** The KNeighborsClassifier achieves the lowest performance among the evaluated models. This model might struggle to find relevant neighbors in the high-dimensional feature space created by concatenating multiple embeddings.

5 Conclusion and Future works

The aim of the project was to distinguish different multilingual speakers from one another, using short clips of their audio samples. To achieve this goal we experimented with various feature extraction methods and then finally settled on using the concatenated features. We experimented our methods using different models as given in table 4. From this we can conclude that our method of using different feature extraction methods like MFCCs, BFCCs, etc.,, and then using a concatenated vector as our final feature is a viable option for training a ml model for fast and efficient, speaker identification system. In future we would like to expand our approach by using Deep learning models and also training on much larger and diverse dataset.

References

[1] Douglas A Reynolds. Speaker identification and verification using gaussian mixture speaker models. *Speech communication*, 17(1-2):91–108, 1995.

[2] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798, 2010.

[3] Rohan Kumar Das and SR Mahadeva Prasanna. Exploring different attributes of source information for speaker verification with limited test data. *The Journal of the Acoustical Society of America*, 140(1):184–190, 2016.

[4] Philippe Thévenaz and Heinz Hügli. Usefulness of the lpc-residue in text-independent speaker verification. *Speech Communication*, 17(1-2):145–157, 1995.

[5] Debadatta Pati and SR Mahadeva Prasanna. Subsegmental, segmental and suprasegmental processing of linear prediction residual for speaker information. *International Journal of Speech Technology*, 14:49–64, 2011.

[6] K Sri Rama Murty and Bayya Yegnanarayana. Combining evidence from residual phase and mfcc features for speaker recognition. *IEEE signal processing letters*, 13(1):52–55, 2005.

[7] Josef Kittler, Mohamad Hatef, Robert PW Duin, and Jiri Matas. On combining classifiers. *IEEE transactions on pattern analysis and machine intelligence*, 20(3):226–239, 1998.

[8] A Venturini, Leonardo Zao, and Rosângela Coelho. On speech features fusion, α-integration gaussian modeling and multi-style training for noise robust speaker classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(12):1951–1964, 2014.

[9] Joseph P Campbell. Speaker recognition: A tutorial. *Proceedings of the IEEE*, 85(9):1437–1462, 1997.

[10] Rosa González Hautamäki, Tomi Kinnunen, Ville Hautamäki, Timo Leino, and Anne-Maria Laukkanen. I-vectors meet imitators: on vulnerability of speaker verification systems against voice mimicry. In *Interspeech*, pages 930–934. Citeseer, 2013.

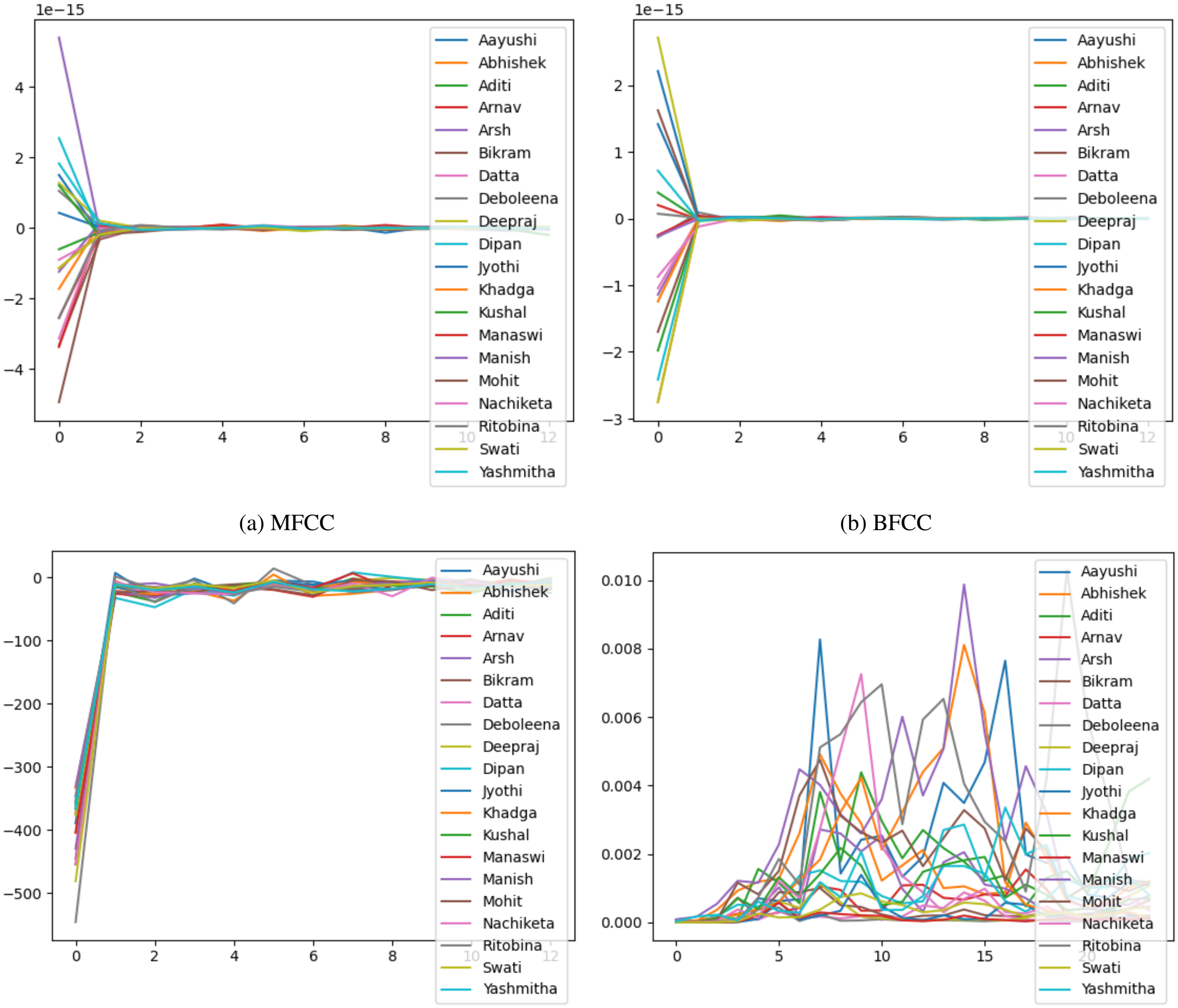
[11] Debadatta Pati and SR Mahadeva Prasanna. Speaker verification using excitation source information. *International journal of speech technology*, 15:241–257, 2012.

[12] S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366, 1980.

[13] Xuedong Huang, Alex Acero, Hsiao-Wuen Hon, and Raj Reddy. *Spoken language processing: A guide to theory, algorithm, and system development*. Prentice hall PTR, 2001.

[14] Florian Eyben, Stavros Petridis, Björn Schuller, George Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. Audiovisual classification of vocal outbursts in human conversation using long-short-term memory networks. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5844–5847, 2011.

[15] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.



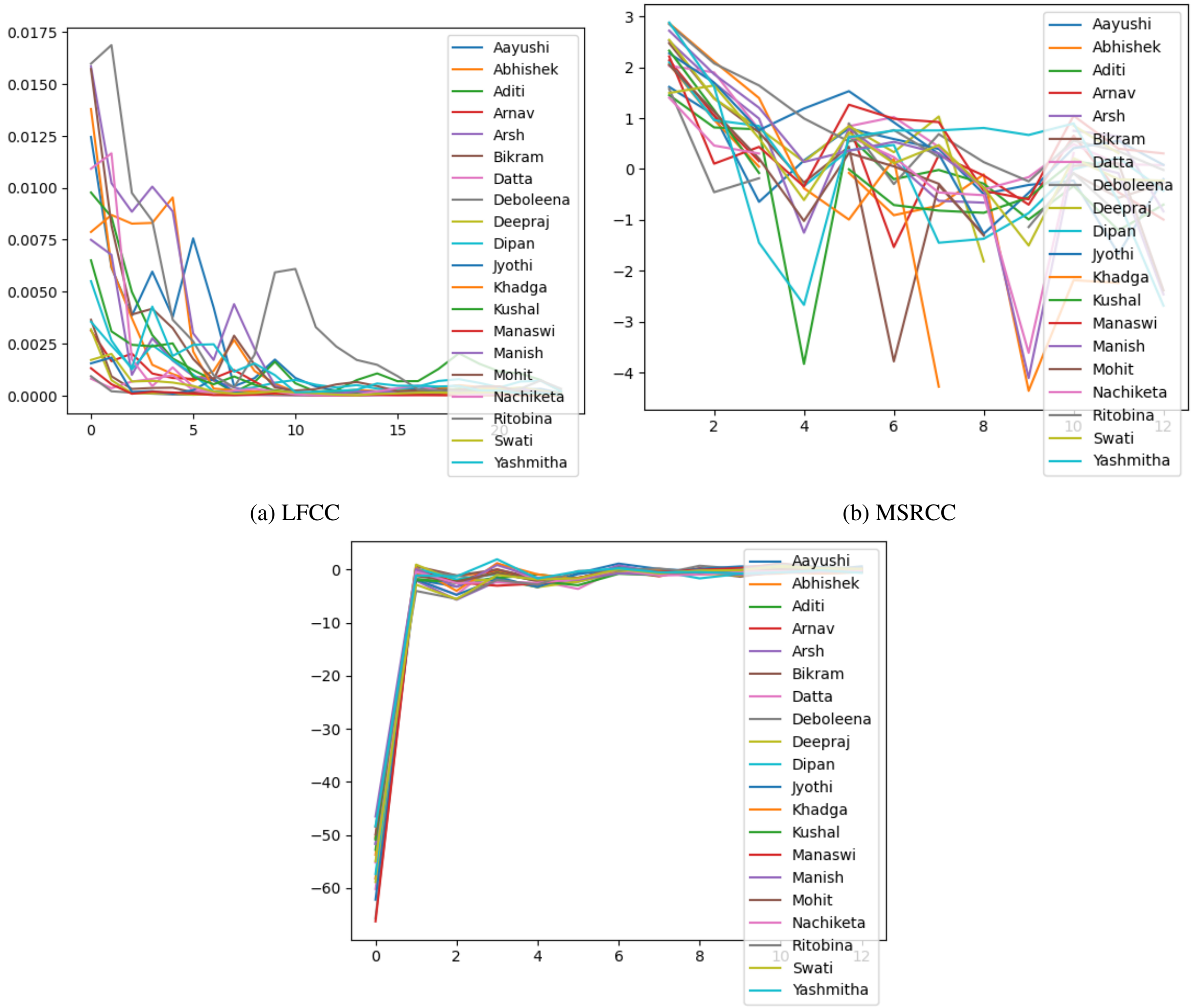
(a) MFCC

(b) BFCC

(c) CQCC

(d) GFCC

Figure 3: Feature Extractions techniques



(a) LFCC

(b) MSRCC

(c) NGCC

Figure 4: Feature Extractions techniques