Speaker Identification

Mohit Bansal
B.Tech
CSB
mohit20526@iiitd.ac.in

Bhavya Narnoli
B.Tech
CSD
bhavya21316@iiitd.ac.in

Pravar Aggarwal
BTech
ECE
pravar20229@iiitd.ac.in

Yatin
MTech
CSE
yatin23108@iiitd.ac.in

Purbasha Barik
MTech
CSE
purbasha23068@iiitd.ac.in

Abstract

Speaker identification, a fundamental challenge in signal processing, has garnered substantial attention due to its wide-ranging applications in security, forensics, and human-computer interaction. This paper presents a comprehensive review of speaker identification methodologies using classical machine learning techniques.

The two primary phases of Speaker Identification systems are feature extraction and categorization. In this article, we investigate these two components with the intention of raising a speaker's performance system for identification in noisy environments.The parameters used in the proposed system are: Mel Frequency Cepstral Coefficients (MFCC), their first and second derivatives (Deltas and DeltaDeltas).In this paper, a complete comparison of different combinations of the previous features(MFCC, Delta mfcc, delta delta mfcc,ldb,dwt) is discussed.

The accuracy rate is raised by the suggested approach. Using the Gaussian Mixture Model (GMM) classifier, the trial results yielded the best accuracy of 99.08% for speaker identification.

Keywords:Speaker Identification, Speech Feature Extraction; MFCC ,Delta MFCC,Delta Delta MFCC , LDB,DWT. Classifier:GMM

1 Existing Analysis

(Al-Rawahy et al.,2012) A text-independent speaker-identification system based on the DCT-Cepstrum Histogram and Gaussian Mixture Model GMM is implemented. The new feature was tested using speech files from the ELSDSR database and TIMIT corpus. The new feature set managed gave a accuracy of 100% on 23 speakers from the ELSDSR database, and 99% on 630 speakers from the TIMIT corpus.

Jahangir et al. (2018) employed statistical features (min, max, mean, mode, median, standard deviation, variance, and covariance) along with zero-crossing rate and RMS, extracted from speech data collected from 5 females and 5 male speakers. These features were assessed on classification algorithms, including SVM, k-NN, RF, NB, and J48. The authors introduced a hierarchical classification approach, where speaker gender was identified at the first level, followed by speaker classification at the second level. The Random Forest (RF) algorithm gave an accuracy of 96.9% for gender classification and 78% and 88.7% for male and female classification models, respectively.

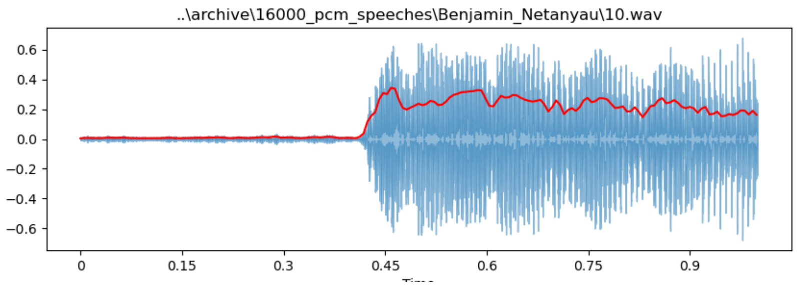
(Abdalmalak et al.,2018) extracted MFCCs, ∆MFCCs, ∆∆MFCCs, PLP, BFCC and RASTA-PLP features and tried 13 different combinations of these features. Finally a combination of three distinct classifiers: linear kernel SVM, RBF kernel SVM, and logistic regression. This approach improved the generalization and learning capabilities of a The results showed that combining the features MFCC, PLP, RASTA-PLP, and BFCC led to the most optimal performance, achieving an impressive accuracy of 98% when applied to clean speech data.

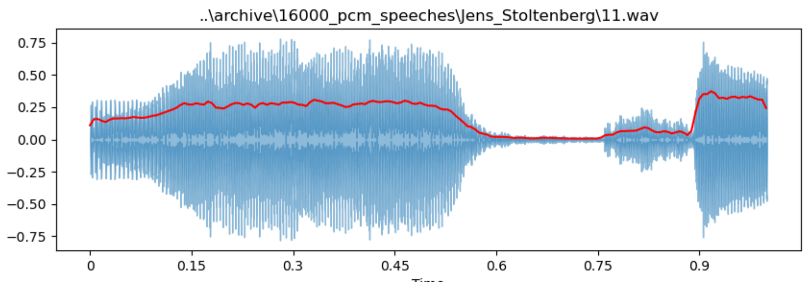
1.1 EDA

1.1.1 RMS Energy Envelope and Amplitude

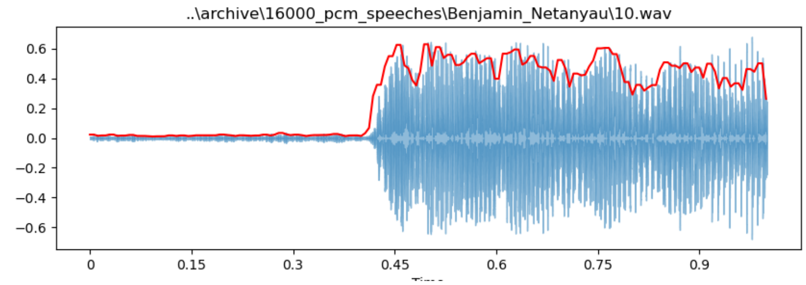
RMS (Root Mean Square) envelope of an audio signal is a representation of the signal's amplitude variations over time. It provides a smoothed measure of the signal's energy, which is especially useful for analyzing and visualizing the overall intensity or loudness of the audio.

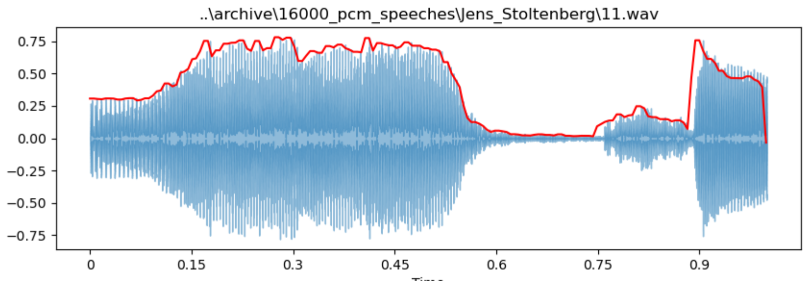
Amplitude Envelope - Loudness varies over time for a given audio. It is found by dividing the given audio into multiple frames of equal size. Then for each frame find the maximum amplitude (loudness) and plot it along with the original sound wave to get the below plots.





The rms energy can be used as a feature to train our model for identifying the speaker. Although, training with mfcc, and other features didn't boost the accuracy.

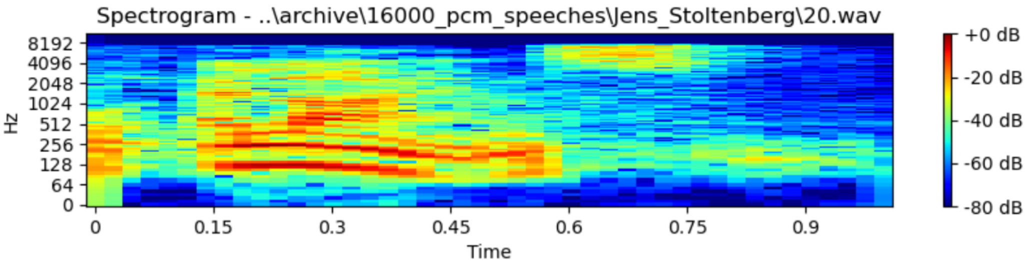




1.1.2 Spectogram

It indicates all the frequency components present at a given time in the audio and their respective dominance in that frame.

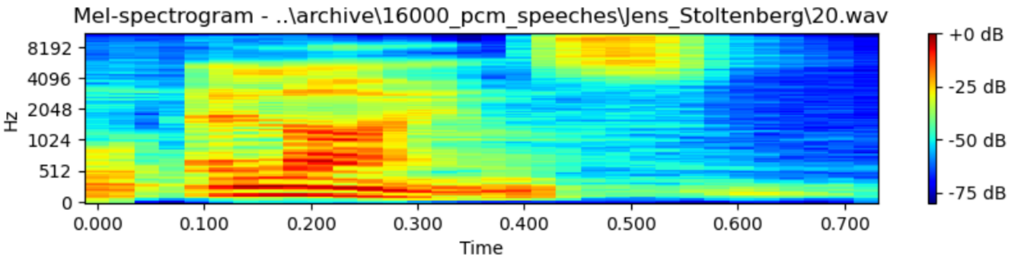
Above is the plot of the spectrogram of random audio from benjamin class. The cmap used is 'jet' to get a clear picture of the dominant frequency component at a given time in the audio.



The red part indicates the dominant frequency whereas blue represents the frequency that is least present in that frame. It also contains the information when the speaker is talking. Like in the first plot, the speaker started talking after a delay as when he spoke then a certain range of frequency components became dominant in that duration.

1.1.3 Mel-Spectogram

It is a visual representation of audio data that indicates the change in frequency of sound over time. It differs from a spectrogram as it uses a Mel scale, which emphasizes human-perceived sounds. Whereas a normal spectrogram considers a linear relationship in frequency. Above is the plot of the mel-spectrogram for random audio belonging to Benjamin and Jens class. Mel-spectrogram provides better details about the human speech sound than the spectrogram and with the help of mel-spectrogram, we can extract relevant features that can be used for classification or speaker recognition tasks.



1.1.4 Wavelet Transform

It is a mathematical tool that is used to divide a signal into frequency components. Each component is studied by matching it to its scale. The wavelet transform provides both time and frequency information, making it suitable for non-stationary signal analysis. (Note: Speech signals are non-stationary in nature. Their statistical properties like mean, variance, intensity, etc, change over time.)

$$DWT_x^{\phi}[k,m] = \frac{1}{\sqrt{2^m}} \sum_{n=-\infty}^{\infty} x[n]\phi\left(\frac{n-k2^m}{2^m}\right) \tag{1}$$

Where:
* $DWT_x^{\phi}[k,m]$ represents the discrete wavelet transform of a signal x[n].

* ψ* is the wavelet function ψ
* k and m adjust the scaling and translation of the signal
* Sum is taken over all the samples n

For different parameter values, we will have different wavelet functions. These are also called mother wavelets.

2 Methodology

2.1 Audio Augmentation

Augmentation techniques applied:

1. **Time Stretching:**
   - Time stretching is employed to alter the duration of audio signals, creating temporal variations.
2. **Background Noise:**
   - Background noises are added using the 'audiomentations' library, introducing the noises given to us inside the '_background_noises_' folder.
3. **Gaussian White Noise:**
   - Gaussian white noise is added to simulate random noise, enhancing resilience to unexpected sound conditions, and making the model robust.
4. **Other Speaker Voices as Noise:**
   - The other speaker voices are added as background noise, providing additional diversity in the dataset.

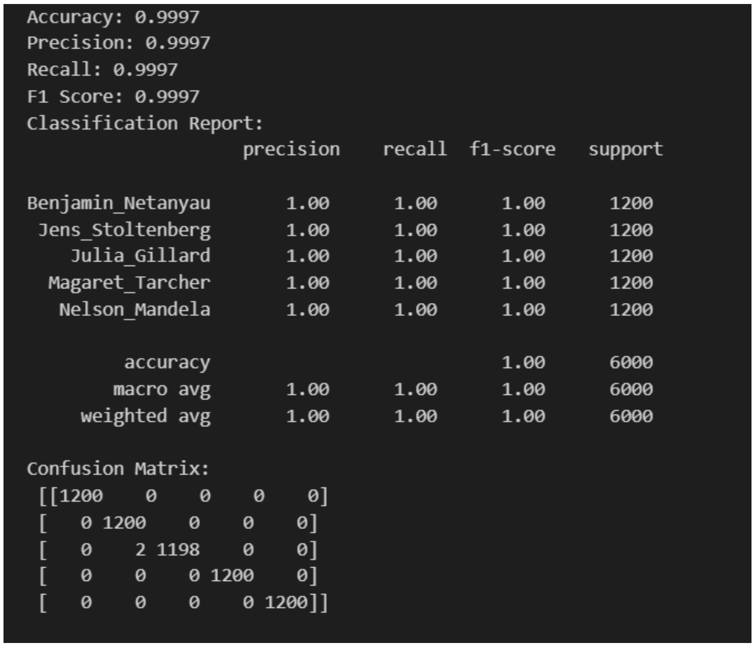
2.2 Feature Engineering

1. **Premphasis:**
   - Preemphasis amplifies the high-frequency components of a signal, boosting higher frequencies relative to lower ones, thus increasing the signal-to-noise ratio crucial for speech intelligibility.
   - Compensates for attenuation during transmission or recording, ensuring higher frequencies are pronounced after transmission or recording.
2. **MFCC:**
   - Represents the short-term power spectrum, emphasizing perceptually relevant frequency components.
   - Involves framing the signal, windowing, FFT computation, processing through a Mel filter bank, logarithm, and DCT application.
   - Provides a compact representation capturing overall spectral characteristics for analysis.
3. **Delta MFCC:**
   - Captures changes in spectral characteristics over time, offering insights into speech spectrum evolution.
   - Provides a dynamic representation of a speaker's voice by focusing on subtle changes in speech dynamics.
   - Helps differentiate unique voice features accurately for speaker recognition.
4. **Delta Delta MFCC:**
   - Complements Delta MFCC by capturing acceleration in changes to vocal characteristics.
   - Enhances system precision in capturing variations in speech of different speakers by addressing changes and accelerations.
5. **LDB:**

- Quantifies energy distribution across frequency bands, enhancing discriminatory power in speaker recognition.
- Analyzes specific frequency regions in the frequency domain.

3 Models

Model 1 : Features : MFCC, Delta mfcc, delta delta mfcc
Cross Validation accuracy on 20% split of data GMM - 99.97 %



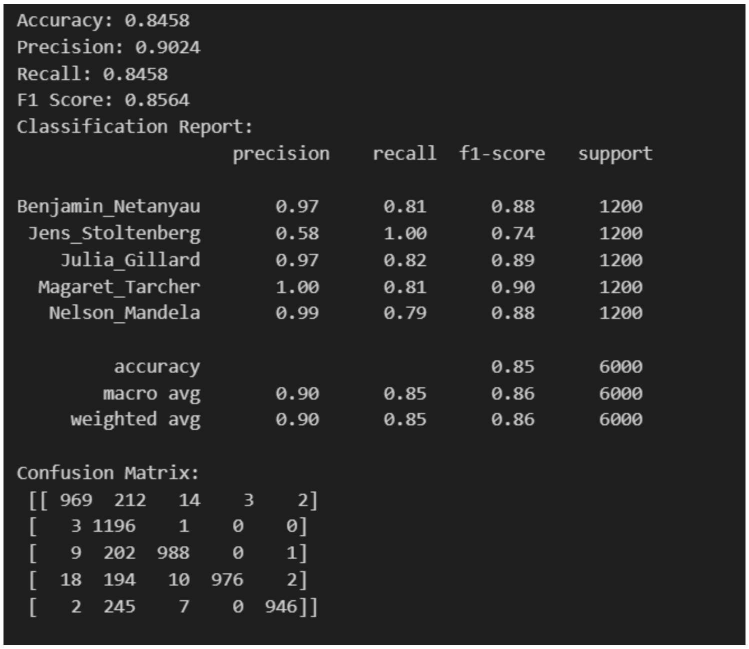
```
Accuracy: 0.9997
Precision: 0.9997
Recall: 0.9997
F1 Score: 0.9997
Classification Report:
                   precision    recall  f1-score   support

Benjamin_Netanyau       1.00      1.00      1.00      1200
 Jens_Stoltenberg       1.00      1.00      1.00      1200
    Julia_Gillard       1.00      1.00      1.00      1200
  Magaret_Tarcher       1.00      1.00      1.00      1200
   Nelson_Mandela       1.00      1.00      1.00      1200

         accuracy                           1.00      6000
        macro avg       1.00      1.00      1.00      6000
     weighted avg       1.00      1.00      1.00      6000

Confusion Matrix:
 [[1200    0    0    0    0]
 [   0 1200    0    0    0]
 [   0    2 1198    0    0]
 [   0    0    0 1200    0]
 [   0    0    0    0 1200]]
```

Model 2 : GMM for MFCC, Delta MFCC, and LDB features GMM - 84.58 % accuracy



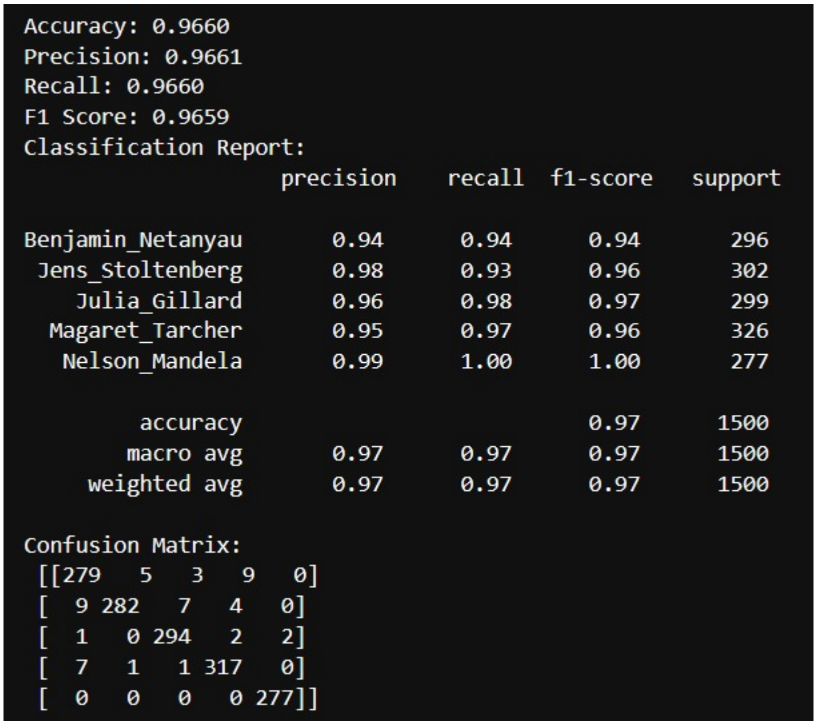
```
Accuracy: 0.8458
Precision: 0.9024
Recall: 0.8458
F1 Score: 0.8564
Classification Report:
                    precision    recall  f1-score   support

Benjamin_Netanyau       0.97      0.81      0.88      1200
 Jens_Stoltenberg       0.58      1.00      0.74      1200
    Julia_Gillard       0.97      0.82      0.89      1200
  Magaret_Tarcher       1.00      0.81      0.90      1200
   Nelson_Mandela       0.99      0.79      0.88      1200

         accuracy                           0.85      6000
        macro avg       0.90      0.85      0.86      6000
     weighted avg       0.90      0.85      0.86      6000

Confusion Matrix:
 [[ 969  212   14    3    2]
 [   3 1196    1    0    0]
 [   9  202  988    0    1]
 [  18  194   10  976    2]
 [   2  245    7    0  946]]
```

Model 3: SVM (MFCC, delta mfcc, ldb) SVM (RBF Kernel) - accuracy 96.60%



```
Accuracy: 0.9660
Precision: 0.9661
Recall: 0.9660
F1 Score: 0.9659
Classification Report:
                    precision    recall  f1-score   support

Benjamin_Netanyau       0.94      0.94      0.94       296
 Jens_Stoltenberg       0.98      0.93      0.96       302
    Julia_Gillard       0.96      0.98      0.97       299
  Magaret_Tarcher       0.95      0.97      0.96       326
   Nelson_Mandela       0.99      1.00      1.00       277

         accuracy                           0.97      1500
        macro avg       0.97      0.97      0.97      1500
     weighted avg       0.97      0.97      0.97      1500

Confusion Matrix:
 [[279   5   3   9   0]
 [  9 282   7   4   0]
 [  1   0 294   2   2]
 [  7   1   1 317   0]
 [  0   0   0   0 277]]
```

4 Analysis

Model 1: GMM with MFCC, Delta MFCC, and Delta Delta MFCC Features
Accuracy –> 99.08%
This model, trained on 80% clean and augmented data, achieved a high accuracy of 99.08% on the 20% remaining split of the data. The combination of MFCC, Delta MFCC, and Delta Delta MFCC features appears to be effective

Model 2: GMM with MFCC, Delta MFCC, and LDB Features
Accuracy –> 84.4%
This model, incorporating Local Distance-Based (LDB) features along with MFCC and Delta MFCC, achieved an accuracy of 84.4%. While not as high as the first model, it's still a respectable performance.
Model 3: SVM (RBF Kernel) with MFCC, Delta MFCC, and LDB Features
Accuracy –> 96.60%
This SVM model, utilizing an RBF kernel and features including MFCC, Delta MFCC, and LDB, achieved an accuracy of 96.60%. SVMs with RBF kernels are known for their effectiveness in capturing complex relationships in the data.

5 Result

Selected Model: GMM with MFCC, Delta MFCC, Delta Delta MFCC
Reasoning: The GMM model with the combination of MFCC, Delta MFCC, and Delta Delta MFCC features was chosen as the final model. The decision might be influenced by factors such as high accuracy (99.08%), suitability for the task, or practical considerations.