Speaker Identification using Classical Machine Learning Techniques

Aadya
IIIT-Delhi
New Delhi, India
aadya21370@iiitd.ac.in

Paridhi Mundra
IIIT-Delhi
New Delhi, India
paridhi20392@iiitd.ac.in

Prakhar Sharma
IIIT-Delhi
New Delhi, India
prakhar23060@iiitd.ac.in

Shazra Irshad
IIIT-Delhi
New Delhi, India
shazra23089@iiitd.ac.in

Abstract

In today's technology, **voice recognition** has applications in security, authentication, and virtual assistants. Understanding how to accurately identify speakers under various acoustic environments, including the challenges posed by background noise, is of paramount importance. This project addresses the persistent issue of background noise by proposing novel methodologies within the classical machine learning paradigm, employing the Gaussian Mixture Model (GMM). Our findings indicate that the GMM-based model exhibits exceptional performance, achieving exceptional accuracy on the test dataset. This success underscores the model's efficacy in mitigating the impact of background noise and highlights its potential to advance speaker recognition technologies. The insights gained from this research contribute valuable knowledge to the ongoing evolution of voice recognition systems, enhancing their robustness in real-world applications.

1 Introduction

Voice recognition is primarily categorized into two components: speaker identification and speaker verification. **Speaker identification** involves identifying which registered speaker is responsible for a given utterance from a known group of speakers. On the other hand, speaker verification assesses and decides whether a speaker's claimed identity is accepted or rejected. In this project, our primary focus lies in the realm of speaker identification [1].

When it comes to voice recognition systems, the input component serves as a fundamental prerequisite. Currently, the two most common audio formats are WAV and MP3. WAV files are the preferred choice for most researchers because they cover the entire audible frequency spectrum. In contrast, MP3 files are compressed and, as a result, may lack some of the information found in their corresponding WAV files [2]. Additionally, extracting features from WAV files is essential, as this step forms the foundation for the machine learning algorithms used in data classification. This makes WAV files a prevalent choice in audio research [3]. Maintaining a consistent sampling rate within an audio sample is crucial to ensure that the extracted coefficients accurately reflect the same underlying calculations [4].

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

2 Existing Analysis

2.1 Early Foundational Work

The historical evolution of Automatic Speech Recognition (ASR) is examined, shedding light on its pivotal components—speech recognition and speaker recognition[4]. In the early stages, the focus was primarily on speech understanding, with Davis et al's groundbreaking 1952 study laying the foundation for speaker recognition using a novel NLP-based approach [5].

2.2 Feature Extraction: The Crux of Speaker Recognition

A subsequent investigation delved into the efficacy of Mel Frequency Cepstral Coefficients (MFCC) in speaker recognition[6]. Employing techniques such as weighted MFCC, the study achieved an impressive nearly 1 per cent increase in average accuracy, demonstrating notable improvement[7]. The application of vectorization techniques, specifically using spectrograms and row mean vectors, yielded promising results, with average accuracies surpassing 90 per cent across multiple sentence variations[8].

2.3 Confronting Noise: Genetic Algorithms in Real-World Settings

In addressing the challenge of noise elimination, the integration of Genetic Algorithms into an open-set speaker recognition system showcased promising outcomes [9]. Notably, this approach achieved an average false reject rate of 14 per cent and an average false acceptance rate of 28.125 per cent, highlighting its effectiveness in handling real-world scenarios[2].

Further exploration into machine learning techniques revealed diverse performance metrics [10]. The SVM-based approach, as experimented by Fenglei et al in 2001, demonstrated effectiveness in large-scale samples, albeit with the challenge of text-independent training [11]. The Naïve Bayes model, when applied to speaker recognition, exhibited varying accuracies, with 30.39 per cent, 23.59 per cent, and 27.14 per cent using k-Nearest Neighbors (k=15), Naïve Bayes, and Conditional Random Field (CRF), respectively [11].

2.4 Navigating the Evolution

From early NLP-driven models to advanced machine learning techniques like SVM and Naïve Bayes, the quest for optimal feature extraction and classification methods persists [12]. The findings emphasize the dynamic landscape of speaker recognition technologies and the nuanced challenges faced in achieving accurate and robust results [4].

3 Dataset

The dataset that we have used is the "Prominent Leader's Speeches." It has a uniquely curated collection of one-second audio clips extracted from speeches delivered by five globally recognized leaders—Benjamin Netanyahu, Jens Stoltenberg, Julia Gillard, Margaret Thatcher, and Nelson Mandela. Each audio clip is encoded in the PCM format, ensuring a standardized and high-quality representation with a consistent sampling rate of 16kHz. To enhance the dataset's realism and account for real-world scenarios, a dedicated folder contains background noise audio files, including instances of laughter and applause. We analyzed the dataset and found that the Speaker's folders are well-balanced.

4 Exploratory Data Analysis

In the context of audio signal processing, various types of features are extracted to characterize different aspects of the signal. We've analyzed Time, Frequency, and Cepstral Domain Features.

4.1 Time Domain Features:

- Pitch: Represents the perceived frequency of a sound. In simpler terms, it is how high or low we perceive a sound to be.

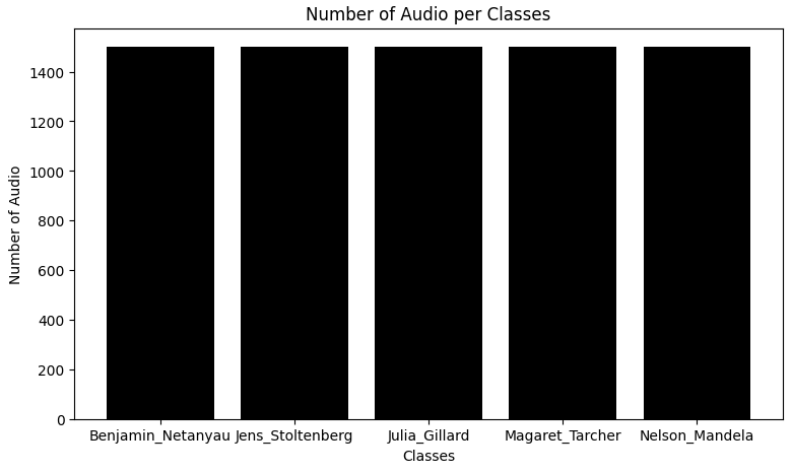


Figure 1: Well balanced dataset

```
Statistics for Benjamin_Netanyau:
Pitch: 94.99972534179688
RMSE: 0.15797200798988342
RMSE_Var: 0.007618904113769531
Amplitude_Mean: 0.12445974349975586
Amplitude_Var: 0.01710120588541031
```

Figure 2: Statistics for Benjamin Netanyau

- Amplitude: Represents the magnitude of the signal at a given point in time.
- Root Mean Square (RMS): Calculated as the square root of the mean of the squared values of the signal. It provides a measure of the signal's energy.
- Zero Crossing Rate: Indicates the rate at which the signal changes its sign, often used to characterize speech signals.

4.2 Frequency Domain Features:

- Fast Fourier Transform (FFT): Transforms a signal from the time domain to the frequency domain, providing information about the frequency components present in the signal.

4.3 Cepstral Domain Features:

- Cepstrum: Derived from the spectrum by taking the inverse Fourier transform of the logarithm of the spectrum. It helps in separating the information related to the vocal tract and excitation source in speech signals.
- Mel-Frequency Cepstral Coefficients (MFCCs): Widely used in speech and audio processing, MFCCs are coefficients representing the short-term power spectrum of a sound signal.
- Delta and Delta-Delta Coefficients: Represent the rate of change and acceleration of MFCCs, respectively. They capture dynamic information in the speech signal.

4.4 Waveplots

One of the most common and essential techniques in audio processing is the conversion of audio data into various types of waveplots. These visual representations capture the temporal characteristics of sound and provide valuable insights into audio content.

MFCC coefficient for Jens_Stoltenberg

	coeff_num	coeff_mean	coeff_var
0	0	-158.204926	21584.839844
1	1	44.822517	1879.286133
2	2	-14.626332	1692.024780
3	3	21.088486	729.215332
4	4	-2.997458	407.906647
5	5	-1.359105	176.090347
6	6	-14.453612	152.596939
7	7	-7.252954	135.594788
8	8	-15.089292	87.899498
9	9	-10.149036	96.711235
10	10	-13.737803	73.166092
11	11	-8.750462	50.412430
12	12	-14.796031	70.538666

Figure 3: Mean and Variance of MFCC Coefficients for Jens Stoltenberg

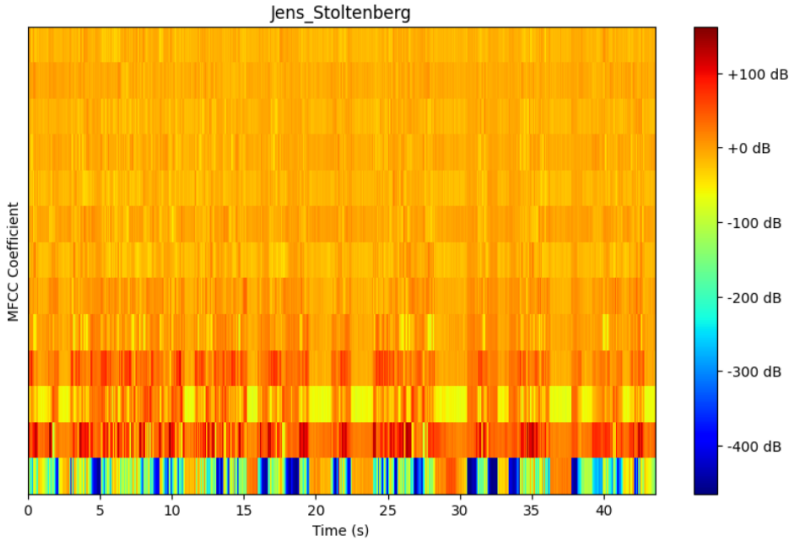


Figure 4: MFCC Plot for Jens Stoltenberg

4.4.1 Waveforms

Waveforms are fundamental in audio processing as they visually represent the amplitude of an audio signal over time. They provide an intuitive way to understand the temporal characteristics of sound, making them valuable for tasks such as speech analysis, music processing, and sound recognition.

4.4.2 Spectograms

Spectrograms offer a more detailed view of audio data. They represent how the frequency content of a signal changes over time. A spectrogram is particularly useful for identifying sound patterns, like musical notes in music or phonemes in speech, and for detecting transient events or anomalies.

4.4.3 Mel Spectograms

Mel spectrograms are a specialized form of spectrogram that emphasize the frequency components of sound as perceived by the human ear. By using the Mel scale, which is a nonlinear transformation of frequency, they align more closely with how we perceive sound. This makes Mel spectrograms a common choice in speech and audio analysis tasks where human auditory perception plays a role.

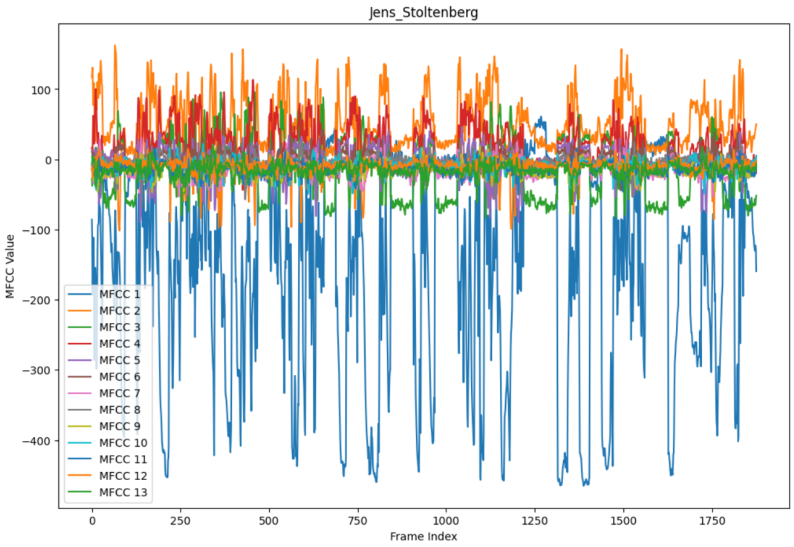


Figure 5: MFCC Index vs Value Plot for Jens Stoltenberg

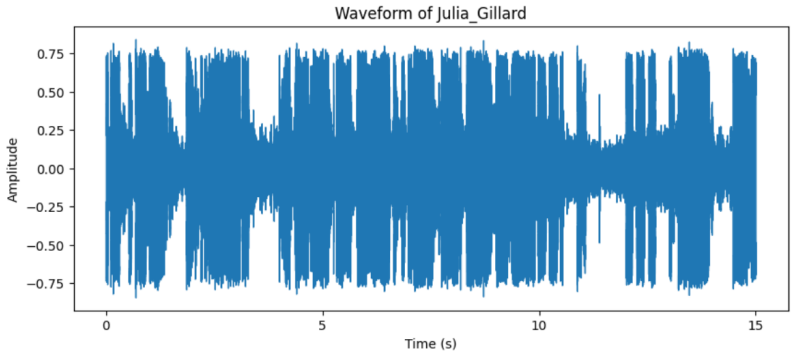


Figure 6: Waveform for Julia Gillard (15 seconds)

4.4.4 MFCC Plots

MFCC (Mel Frequency Cepstral Coefficients) plots take the concept of Mel spectrograms a step further. They condense the spectral information into a set of coefficients, which capture essential features of speech or audio, including pitch and spectral shape. MFCC plots are widely used in tasks such as speaker recognition, emotion analysis, and speech processing.

4.4.5 Fast Fourier Transform

The Fast Fourier Transform (FFT) is a mathematical algorithm widely employed in signal processing and data analysis. It's primarily used to transform a time-domain signal into its frequency-domain representation. By doing so, it reveals the underlying frequency components within a given signal. The FFT is invaluable in various fields, enabling one to extract essential information from complex signals, such as identifying distinct frequencies in audio, characterizing periodic patterns in data, and understanding the spectral content of signals in a concise manner.

4.5 Findings

In our data analysis, we ensured a balanced distribution of speaker classes but noted that noise audio had longer durations than speaker audio. To integrate noise effectively, we segmented longer noise clips into smaller segments matching speaker audio durations, maintaining dataset consistency.

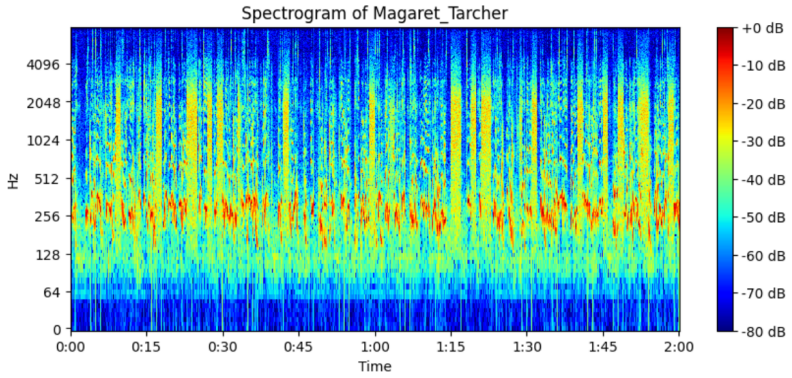


Figure 7: Spectogram for Magaret Tarcher (2 minutes)

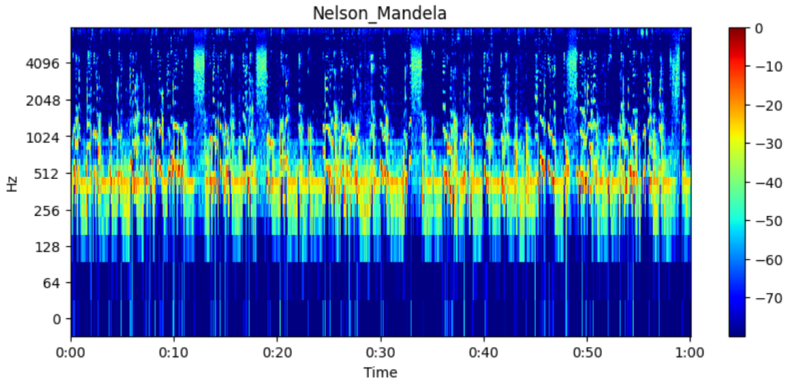


Figure 8: Mel Spectogram for Nelson Mandela

Examination of spectral features showed variations among speakers, with Stoltenberg and Gillard exhibiting more pronounced frequency content variations. Initially, we attempted to identify distinctive pitch patterns for each speaker's audio. However, this approach proved impractical given the dataset's composition of two female and three male speakers, rendering it ineffective for our purposes. We processed 6 noise files into 370 individual 1-second noise samples, aligning with our speaker audio data for further analysis and integration.

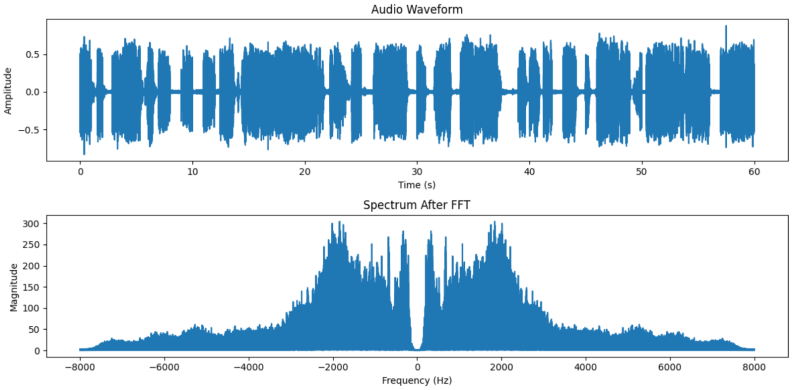


Figure 9: Before vs After doing FFT on Magaret Tarcher's Waveform

5 Methodology

5.1 Model Training

Utilizing Gaussian Mixture Models (GMM) in our speaker recognition methodology stems from their adeptness in handling the intricacies of noise within our dataset. This choice empowers us to construct a model tailored to our noise addition techniques, thereby bolstering the adaptability of our system. In particular, we have implemented five distinct GMMs, each specifically designed for a unique class of speakers. This targeted approach not only refines the accuracy of our model but also contributes to its robustness, ensuring more reliable results in the presence of varying acoustic conditions and noise.

5.2 Adding Noise

In the model, we have implemented a comprehensive methodology for adding noise to audio signals, catering to various noise sources. Initially, noise samples are loaded from the background folder and other noise directories, ensuring correct sampling rates and splitting them into manageable sections. Real noise is then added to clean audio signals through our add real noise function, where a random noise sample is selected, scaled, and superimposed on the original signal. Additionally, the code provides functions to generate synthetic white and pink noise. Our adjust noise rms function enables the adjustment of noise amplitude to achieve a specified Signal-to-Noise Ratio (SNR). This versatile approach allows for tailoring the noise augmentation process based on their specific requirements, providing a valuable tool for tasks like audio data augmentation in machine learning or simulating realistic audio environments for testing and analysis.

5.3 Testing and Evaluation

In the testing phase, the model undergoes evaluation using audio files from the pre-defined test dataset. This dataset, initially split for validation purposes, serves as a rigorous examination ground to assess the model's ability to accurately identify speakers under real-world conditions.

The evaluation stage is a crucial step in assessing the model's performance. A comprehensive set of metrics is employed to gauge the speaker recognition model's accuracy and efficacy. These include traditional measures such as accuracy, providing an overall assessment of correct predictions. A classification report is also generated, offering insights into precision, recall, and F1-score for each speaker label. The confusion matrix further delineates the model's performance by illustrating the distribution of correct and incorrect predictions across different speaker classes.

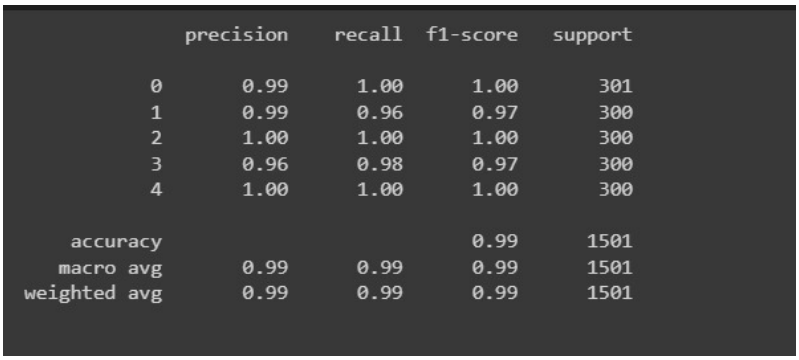
6 Results

Our model utilising Gaussian Mixture Models (GMM), showcases remarkable performance metrics. Precision, recall, and F1-score for each class with an overall accuracy of 99 percent. The robustness of the model is evidenced by its ability to effectively distinguish speakers, as supported by the macro and weighted average metrics. The integration of GMM for speaker identification aligns with established practices in the field, highlighting the paper's commitment to leveraging proven methodologies. This research contributes to the speaker recognition domain by emphasizing the efficacy of traditional machine learning techniques, specifically GMM, in achieving outstanding results, thereby providing valuable insights for further advancements in the field.

7 Conclusion

Our research employs classical machine learning models to tackle background noise challenges in speaker recognition, achieving 99 per cent accuracy on the test dataset. This success emphasizes the model's effectiveness in real-world scenarios. The methodology involves meticulous GMM training, accurate testing, and a comprehensive evaluation using metrics like accuracy, a classification report, and a confusion matrix.

The project contributes to the evolution of speaker recognition technologies, addressing challenges through innovative approaches. The "Prominent Leader's Speeches" dataset, with intentional back-



	precision	recall	f1-score	support
0	0.99	1.00	1.00	301
1	0.99	0.96	0.97	300
2	1.00	1.00	1.00	300
3	0.96	0.98	0.97	300
4	1.00	1.00	1.00	300
accuracy			0.99	1501
macro avg	0.99	0.99	0.99	1501
weighted avg	0.99	0.99	0.99	1501

Figure 10: Confusion Matrix of the GMM Model

ground noise augmentation, reflects real-world complexities. Our concise and robust methodology sets the stage for advancing speaker recognition in practical applications like security and authentication.

8 Individual Contributions

Preliminary Research was done by all.

- Aadya
  - EDA, Report, Model
- Paridhi Mundra
  - Existing Analysis, Report, Presentation
- Prakhar Sharma,
  - EDA, Model
- Shazra Irshad
  - EDA, Model

References

[1] Nishtha H. Tandel, Harshadkumar B. Prajapati, and Vipul K. Dabhi. Voice recognition and voice comparison using machine learning techniques: A survey. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pages 459–465, 2020.

[2] MOHSIN MUHAMMAD KABIR, M. F. MRIDHA, SHIN JUNGPIL, ISRAT JAHAN, and ABU QUWSAR OHI. A Survey of Speaker Recognition: Fundamental Theories, Recognition Methods and Opportunities.

[3] Automatic Speech and Speaker Recognition | Wiley Online Books.

[4] Abhishek Manoj Sharma. *Speaker Recognition Using Machine Learning Techniques*. Master of Science, San Jose State University, San Jose, CA, USA, May 2019.

[5] Sadaoki Furui. 50 Years of Progress in Speech and Speaker Recognition Research. *ECTI Transactions on Computer and Information Technology (ECTI-CIT)*, 1(2):64–74, 2005. Number: 2.

[6] Ondrej Glembek Glembek and Luka˘´s Burget. COMPARISON OF SCORING METHODS USED IN SPEAKER RECOGNITION WITH JOINT FACTOR ANALYSIS.

[7] Tumisho Billson Mokgonyane, Tshephisho Joseph Sefara, Thipe Isaiah Modipa, Mercy Mosibudi Mogale, Madimetja Jonas Manamela, and Phuti John Manamela. Automatic Speaker Recognition System based on Machine Learning Algorithms. In *2019 Southern African Universities Power Engineering Conference/Robotics and Mechatronics/Pattern Recognition Association of South Africa (SAUPEC/RobMech/PRASA)*, pages 141–146, January 2019.

[8] Sadaoki Furui. An Overview of Speaker Recognition Technology. In Chin-Hui Lee, Frank K. Soong, and Kuldip K. Paliwal, editors, *Automatic Speech and Speaker Recognition*, volume 355, pages 31–56. Springer US, Boston, MA, 1996. Series Title: The Kluwer International Series in Engineering and Computer Science.

[9] R. J. Mammone, Xiaoyu Zhang, and R. P. Ramachandran. Robust speaker recognition: a feature-based approach. *IEEE Signal Processing Magazine*, 13(5):58, September 1996.

[10] M. Faundez-Zanuy and E. Monte-Moreno. State-of-the-art in speaker recognition. *IEEE Aerospace and Electronic Systems Magazine*, 20(5):7–12, May 2005.

[11] Hou Fenglei and Wang Bingxi. Text-independent speaker recognition using support vector machine. In *2001 International Conferences on Info-Tech and Info-Net. Proceedings (Cat. No.01EX479)*, volume 3, pages 402–407 vol.3, 2001.

[12] J.P. Campbell. Speaker recognition: a tutorial. *Proceedings of the IEEE*, 85(9):1437–1462, September 1997.