Speaker Recognition Based on Timbre Analysis

By

Md. Abu Quwsar Ohi, ID:15163103017
Md. Ruhullahil Kabir, ID:15163103040
Md. Touhidul Islam, ID:15163103043

Submitted in partial fulfillment of the requirements of the degree of

Bachelor of Science in

Computer Science and Engineering

Department of Computer Science and Engineering

Bangladesh University of Business and Technology

June 2020

Contents

Declaration

We confirm that this thesis presented for the degree of Bachelor of Science, has

- been composed entirely by ourself
- been solely the result of our own work
- not been submitted for any other degree or professional qualification (except where states otherwise by reference or acknowledgment)

Md. Abu Quwsar Ohi ID: 15163103017	______________________ Signature
Md. Ruhullahil Kabir ID: 15163103040	______________________ Signature
Md. Touhidul Islam ID: 15163103043	______________________ Signature

Dedication

We would like to dedicate our work to diligent researchers for whom the modern age is adequate for robust technology devices.

Acknowledgement

We would like to express our sincere gratitude to Md. Saifur Rahman, Assistant Professor, and Dr. Muhammad Firoz Mridha, Associate Professor, without whom this research work would not exist in its present form.

Abstract

Speaker recognition is an active research area that contains notable usage in biometric security and authentication system. Currently, there exist many well-performing models in the speaker recognition domain. However, most advanced models implement deep learning that requires GPU support for real-time speech recognition, and it is not suitable for low-end devices. In this thesis, we propose a lightweight text-independent speaker recognition model based on a random forest classifier. It also introduces new features that are used for both speaker verification and identification tasks. The proposed model uses human speech based timbral properties as features that are classified using random forest. Timbre refers to the fundamental properties of sound that allow listeners to discriminate among them. The prototype uses seven most actively searched timbre properties, boominess, brightness, depth, hardness, roughness, sharpness, and warmth as features of our speaker recognition model. The experiment is carried out on speaker verification and speaker identification tasks and shows the proposed model's achievements and drawbacks. In the speaker identification phase, it achieves a maximum accuracy of 78%. On the contrary, in the speaker verification phase, the model maintains an accuracy of 80% having an equal error rate (ERR) of 0.24. The overall accuracy measures were done using LibriSpeech (clean) dataset.

List of Tables

List of Figures

Introduction

1.1 Introduction

Speaker recognition is the process of recognizing an individual by hearing a voice. Speaker recognition is an essential perspective of biometric identification and verification. Commonly, speaker recognition is considered a pattern recognition problem. The recognizer's goal is to identify a speaker (previously known) by analyzing the vocal properties of a speech. Generally, humans recognize speakers based on the previously learned timbral properties of speech. Timbral properties refer to the basic properties of speech features such as hardness, softness, roughness, etc.

Speaker recognition can be divided into two divisions based on the system's usage, speaker identification [1], and speaker verification [2]. In machine learning, the identification systems use multi-classification models, whereas the verification systems use binary-classification models. Concerning the utterance used for speaker recognition models, the model can be either text-independent or text-dependent. A text-dependent model only recognizes speakers based on the predefined keyword or passphrase that needs to be uttered by the speaker. This feature is preferred for unlocking devices or verification purposes. Microsoft implemented the text-dependent speaker verification on Windows 10 [3]. On the contrary, a text-independent model can recognize speakers based on any utterance of the speakers. At present, most state of the art speaker recognition model uses a text-independent recognition scheme.

1.2 Problem Statement

At present, the modern deep learning-based speaker recognition systems perform notably better than previous machine learning-based models. Although they perform better on famous datasets, these architectures' usage has not been introduced in industrial phases due to the drawbacks. The present deep learning-based models widely depend on embedding systems, which are mostly trained using HMM. The foremost disadvantage of the embedded system-based speaker identification models is that they often fail to generate better results while testing on the individuals' speech, which is not used in the embedding system's training. As a result, the embedding system and the classification model need to be tweaked based on the occurrence of a new individual. On the contrary, the speaker verification models are easy to train, and most of the state of the art speaker verification models do not rely on embedding vector systems. Thus, speaker verification systems are often implemented in modern techniques and devices, while the speaker identification system is not yet implemented.

1.3 Problem Background

The present field of speaker recognition systems is dominated by embedding system based models, which is not preferable for real-world implementation. Although they produce greater accuracy over most of the famous datasets available, they fail to utilize real-world speech data properly. The key challenges of speech recognition systems are,

- Segmenting speech from real-world audio data.
- Identifying speakers in noisy environments.
- Filtering speech from noisy audio data if required.
- Implementing text-independent speaker recognition systems directly based on vocal properties.

Although researchers are continuously solving these challenges, the implementations are still not accepted as industry level architecture.

1.4 Research Objectives

The research work aims to implement a speaker recognition and verification system that solves the aforementioned vital challenges, including speech segmentation and text-independent speaker recognition. The proposed architecture is tested on a challenging dataset, reviewed in Section 4.4, and the proposed architecture presents promising results.

1.5 Motivations

Speaker recognition has a wide variety of usage in the biometric authentication system, speaker diarization, speech recognition, forensics, and security [4–6]. Now, the tech giant Microsoft provides API for speaker identification and verification. Conversely, IBM Watson includes API for speaker diarization. Speaker recognition systems have a profound influence on call-centers targeting and serving the most priority customers. Applicable services include voice dialing, banking over a telephone network, telephone shopping, database access services, information and reservation services, voice mail, security control for confidential information, and remote access to computers. Another vital application of speaker recognition technology is as a forensics tool.

1.6 Flow of the Research

The research work was carried out in multiple steps. After finalizing the research topic, we first studied the basic theory of speech and sound needed to carry our research work. After the practice, we investigated the most promising state of the art speaker recognition architectures. We investigated the lackings of the proposed architectures and produced our speaker recognition architecture. After finalizing the design, we implemented the overall method. To test the proposed model, we collected a popular dataset and ran tests and evaluations on our implemented architecture. Finally, we completed our thesis writing. Figure 1.1 illustrates the overall steps of the research procedure in a flow diagram.

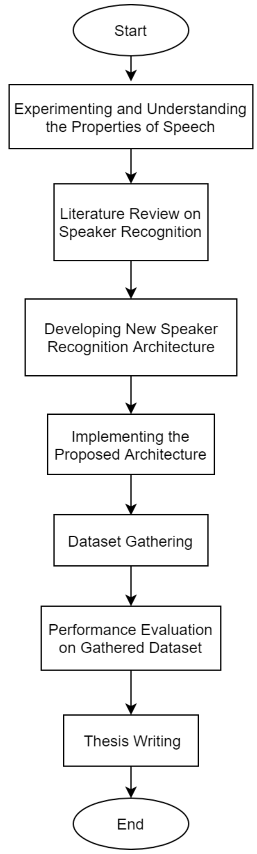

Figure 1.1. The figure illustrates the flow of the research work.

1.7 Significance of the Research

The study's findings will rebound to the benefit to the researchers that direct extraction of features from speech utterance is possible and is also promising. The course explains the extraction procedure of the most known seven feature properties, most important in identifying speakers. Also, we experiment that timbral properties can be defined with the seven extracted speech properties. This work will influence the researchers to investigate the direct extraction procedure of timbral features, without relying on embedding systems. Further research will be suitable for the present industry level implementation of a speaker recognition system ideal for the most challenging environments.

1.8 Research Contribution

The overall contribution of the research work includes,

- We introduce a speaker recognition system that identifies speakers based on the timbral properties of the speech.
- We report speech timbral properties that can be extracted from mel-frequency cepstral coefficients (MFCC) using regression.
- We experiment with a famous dataset and evaluate our proposed architecture's performance in speaker identification and verification scheme.

1.9 Thesis Organization

The thesis work is organized as follows.

Chapter 2 highlights the background and literature review on the field of the speaker recognition system.

Chapter 3 contains the proposed architecture of the speaker recognition system, along with a detailed walkthrough of the overall procedures.

Chapter 4 includes the details of the tests and evaluations that were performed to evaluate our proposed architecture.

Chapter 5 explains the standards, ethical policy, and the challenges of the proposed architecture and the overall field of our study.

Chapter 6 comprises the overall design and implementation constraints of our conducted thesis work.

Chapter 7 illustrates the time schedules that we managed while conducting the thesis.

Finally, Chapter 8 contains the overall conclusion of our thesis work.

1.10 Summary

This chapter includes a comprehensive overview of the problem that we specifically target, the objectives of our thesis work, and the motivation of the thesis work's output. This section also illustrates the overall steps on which we carried out our thesis work.

Background

2.1 Introduction

Speaker recognition is a wide field of interest among researchers, and still, research work is being carried out actively in this field. Most researchers include verification and identification systems in speaker recognition systems and carried out experiments on the two system subsets. Speaker recognition systems are into two sections depending on the working process, text-independent, and text-dependent speaker recognition system. Text-independent systems can recognize speakers based on any speech utterance of users. On the contrary, text-dependent speaker recognition systems require a specific 'password'-word to be uttered by the users, using which the system recognizes the user. In this chapter, we demonstrate the state of the art architectures that are implemented by the researchers.

2.2 Literature Review

Most of the speaker recognition models that are previously introduced use some common ideas, such as, Gaussian Mixture Model (GMM), Hidden

Markov Model (HMM) [7], Dynamic Time Wrapping (DTW), i-vectors, etc. The recent state of the art models uses Neural Network (NN) and Deep Neural Network (DNN) on the feature vectors, which gives promising results [8–20]. Mirco Ravanelli et. al. presented a model which they named SincNet [8]. It uses a new architecture named SincNet, which is used on the first layer with Convolutional Neural Network (CNN). It produces an Equal Error Rate (EER) of 0.32 in the speaker verification stage, greater than this paper's proposed model. They claimed that the DNN architecture faster convergence with epoch around 1200 to 1800, which is still time-consuming and hard to train.

Nunes et al. proposed a new approach for speaker recognition systems called AM-SincNet, based on the SincNet but uses an improved AM-softmax layer [21]. Zheli Liu et. al. used a hybrid method that adds GMM and CNN [9]. They insisted that the model recognizes speakers even on the short utterance of speech, and others also use various types of CNN [22–24] and combined with others [25]. Our proposed model also works for short uttered speech to recognize speakers. F. Richardson et al. explained in their paper that the DNN posterior technique with Mel-Frequency Coefficient Cepstral MFCC as a feature, produces a significant gain over the baseline system, but they degrade the overall performance of the system [10, 12, 13, 26, 27]. Also, there exist many papers where the tradition of using DNN remains [11]. MFCC is a widely used feature that is also commonly used in many proposed models [28, 29]. Instead of using MFCC directly, most of the models mesh it with other system or with features and produces decent results [30–32]. Seiichi Nakagawa et al. has combined MFCC and phase information of wave as features in their model, which achieved an excellent accuracy of 96.7% in speaker identification [28]. K. S. R. Murty et al. further used MFCC and residual phase information [29]. Faragallah et al. propose a robust noise automatic speaker identification (ASI) scheme named MKMFCC–SVM. It is based on the Multiple Kernel Weighted Mel Frequency Cepstral Coefficient (MKMFCC), and support vector machine (SVM)[33]which perform better in noisy condition. Safavi et al. try to use Gaussian Mixture Model - Universal Background Model (GMM-UBM), GMM - Support Vector Machine (GMM-SVM), and i-vector based approaches for better result [34]. Some new speaker recognition models use identification vectors, known as i-vectors [35–40]. It is a feature extraction method that represents the distinctive characteristics of the frame-level features' distributive pattern. Ondrej Novotny et al. declared that the model they prepared was not state of the art, but the usage of i-vector will create a reliable platform for further research [38].

Contrary to i-vectors, speaker embeddings such as x-vectors and d-vectors can leverage unlabelled utterances due to the classification loss over training speakers. Themos Stafylakis and Johan Rohdin propose to train speaker embedding extractors via reconstructing the frames of a target speech segment, given the inferred embedding of another speech segment of the same utterance [41]. Some recent models use DNN to extract d-vectors that is the averaged activation from the last hidden layer of DNN [3, 42]. D. Snyder et al. represented a model using x-vectors, which are embeddings extracted using DNN [43]. They described that their x-vector based model outperformed the i-vector models. Jesús Villalba et al. claimed that x-vectors have already become the new state of the art for speaker recognition [44–46].

Wang et al. introduced an unsupervised domain adaptation approach-domain adversarial training for speaker recognition, which overcomes the domain mismatch problem in the speaker recognition by projecting the source domain and target domain data into the same subspace [47]. This approach does not require labeled data from the target domain and applies an unsupervised domain adaptation. Ahmed et al. use a different system by introducing a Weighted-Correlation Principal Component Analysis (WCR-PCA) to efficiently transform speech features in speaker recognition [48]. Sankar Nidadavolu uses cycle-GANs; they explore domain adaptation at the acoustic feature level by learning feature mappings between domains [49].

2.3 Problem Analysis

Although the state of the art models performs well, they have shortcomings. Models that do not use DNN or ANN contain features that are difficult to extract. They also fall behind on gaining better estimation. On the contrary, the DNN or ANN-based models produce higher accuracy, but they are not suitable (or designed) for recognizing speakers on the continuous audio stream. Also, they consume an enormous amount of processing power in the training phase. As a result, they are not suitable for developing a continuous speaker recognition system.

2.4 Summary

This chapter illustrates the implementations and drawbacks of the latest speaker recognition systems. The thesis's target is to eliminate the imperfections a much as possible and design a speaker recognition system that is stable and suitable for implementing in a real-world situation.

Proposed Model

3.1 Introduction

In this chapter, we discuss the feasibility analysis of the speaker recognition system and the requirements demanded in this model. Finally, this chapter explains the model's overall architecture, which is given by a detailed walkthrough.

3.2 Feasibility Analysis

The thesis work required three researchers with one supervisor and took eight months to be executed. The research work required technical support including, hardware and software. The research work also required dataset generation and evaluation process that is also performed by the researchers. The comprehensive data collection of the thesis work is executed, considering the legal feasibility of the dataset. Also, the thesis work did not require any financial support from the institution and supervisor.

3.3 Requirement Analysis

To conduct the proposed architecture of the overall requirements include,

- High-performance computing device.
- Audio input device.
- Opensource software libraries for scientific computations.
- Opensource software libraries to implement the machine learning model.

3.4 Research Methodology

In this section, the methodology of the proposed model is presented. This section is sub-sectioned into four segments. The sub-sections are sorted from input to output phase of the model consecutively with detailed explanations. Moreover, Figure 3.1 presents the overall workflow of the architecture.

3.4.1 Input Processing

Inputs passed to the model are clean and noise-free audio streams, which may contain silence streams. Each of the audio streams is scaled using the following formula,

$$S(x) = \frac{x_i}{max(abs(x_1, x_2, ..., x_n))} \tag{3.1}$$

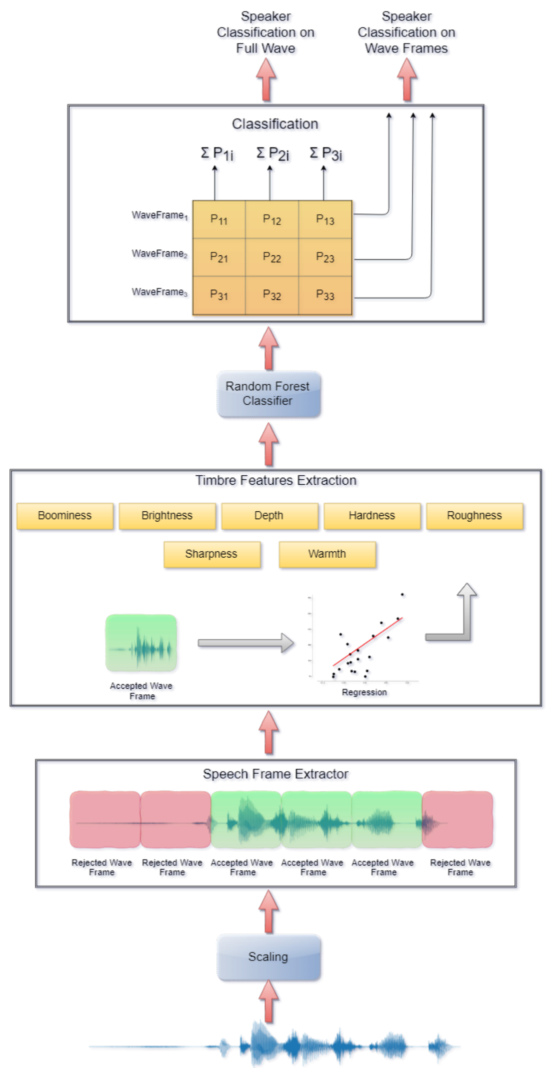

Figure 3.1. The figure illustrates the workflow of the proposed architecture (from bottom to top). The continuous raw waves are first scaled and separated on multiple wave frames. The silence wave frames are filtered out, and the timbral features are extracted using a random forest regressor. The timbral features are further classified using a random forest classifier.

The scaled audio stream further removes the silenced audio frames and the extracted features to be more accurate.

3.4.2 Speech Frame Extractor

The audio stream is further partitioned into audio segments. At first, this phase partitions every 0.3 second consecutive stream of the audio as frames. Each of the wave frames is further passed through the mean calculation function defined as follows,

$$AcceptedFrames = \{S|f(S) = \frac{\sum_{i=1}^{n} s_i}{n}\} \tag{3.2}$$

Here, a frame is rejected if the mean of each wave frame's amplitudes is less than the threshold value that is set to 0.05. This threshold value helps to eliminate the silence parts of the audio streams, which are unnecessary.

3.4.3 Timbre Features Extraction

To extract the timbre properties of sound, the model uses random forest regression. As parameters for regression, a weighted sum of MFCC spectrogram and frequency spectrogram as features. The weighted sum is derived as follows,

$$Sum_{weighted} = \sum_{i=1,j=1}^{n,m} f(i) \times t(j) \times spec(i,j) \tag{3.3}$$

$Where,$

$f(i) =$ Frequency of the i'th index

$t(j) =$ Time of the j'th index

$spec(i,j) =$ Intensity of the sound on f(i) frequency, at time t(j)

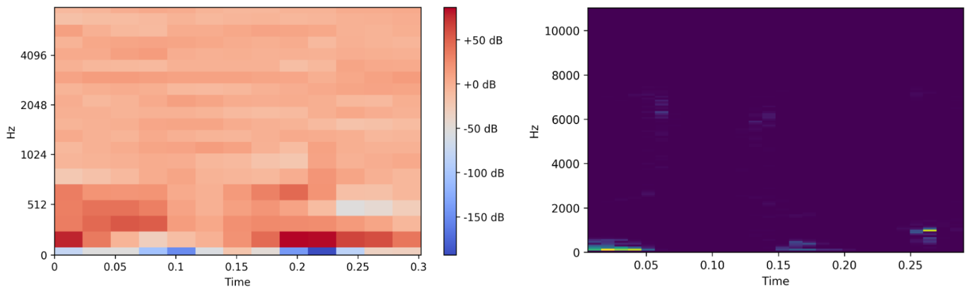

Figure 3.2. An illustration of the MFCC-spectrogram and frequency spectrogram of a 0.3-second speech frame, having weighted sum values of 11818.91, and 0.61 respectively.

The regressor is trained with the prepared dataset containing 400 wave frames and seven timbral properties. For each 0.3-second audio frame, the weighted sum is generated, and the seven timbral properties are trained individually with seven individual random forest regressors.

A short description of the seven extracted speech features is presented below.

Boominess: Booming refers to the deep and resonant sounds. The boominess of sound can also be extracted using Hatano and Hashimoto's boominess index [50].

Brightness: Brightness refers to the higher frequency of sound.

Depth: The term depth can be related to the spreading feel of sound concerning the loudness of sound.

Hardness: This refers to the unbalanced and noisy tone of the sound.

Roughness: This refers to the rapid modulation of sound.

Sharpness: This refers to the amount of high-frequency sound concerning the amount of low frequency of sound. The sharpness of a sound also can be found using the Fastl sharpness algorithm [51].

Warmth: Warmth is the opposite of the brightness of the sound.

3.4.4 Speaker Classification

Each of the features is fed to the Random Forest classifier. The Gini impurity measure is used to measure the quality of a split, which can be stated as,

$$G = \sum_{i=1}^{C} p(i) \times (1 - p(i)) \tag{3.4}$$

The features of each accepted wave frame are processed separately in train and test sessions. In the test session, the classifier outputs each speech wave frame's probabilities uttered from a particular person.

The classification of this model can be for each wave frame or of the full audio stream. To classify each wave frame, the probability vector passed that is the output of the random forest classifier, is passed through the arguments of maxima that can be stated as,

$$\arg\max_{x} f(x) = \{x | f(x) = \max_{x'} f(x')\} \tag{3.5}$$

The probability vectors of the individual wave frames are gathered and produced as a probability matrix to classify the speaker of the full input audio stream. The matrix is then converted to a probability vector defined as,

$$P_i = \sum_{j}^{n} p_{ij} \quad (3.6)$$

The generated probability vector is passed through the maxima function's arguments stated in equation 3.5 to calculate the final classification for the full audio stream.

3.5 Design, Implementation, and Simulation

The overall workflow of the proposed architecture is illustrated in Figure 3.1. All the mentioned steps of the prototype are implemented using Python [52]. The random forest classifier and regressor models are implemented using scikit-learn [53]. Also, for additional calculation, implementation, and support, Numpy [54] and librosa [55] are used. The visual evaluation reports are generated using Matplotlib [56]. The dataset used to test the architecture is directly inserted, and no variations or selections were made while testing the architecture.

3.6 Summary

This section explains the architecture of the proposed timbre-based speaker recognition method. The overall architecture uses the random forest as the base classifier and regressor of the features as well.

Implementation, Testing, and Result Analysis

4.1 Introduction

In this chapter, the proposed architecture is tested and analyzed. This section contains the system setup that was carried out. Simultaneously, this section explains the evaluation metrics used to measure the result accuracy and a detailed analysis of the result.

4.2 System Setup

For training and evaluation, the LibriSpeech corpus is used [57]. It contains speech audios that are labeled based on the 40 speakers. The dataset comprises silenced segments that were not stripped, and our proposed architecture extracts speaker information by directly using the raw audio data.

The model performs regression to extract the timbre properties from speech audio. As there is almost no proper estimation and research done on vocal timbral properties, the dataset generation for timbral properties extraction was cumbersome. We found one tool developed by AudioCommons [58], which could extract all the seven features used in the model. Yet the device produced erroneous outputs for some vocal speech. Therefore, we created a small dataset that contains speech audios and the seven verbal timbral properties, boominess, brightness, depth, hardness, roughness, sharpness, and warmth for each speech audio. The dataset contains 400 samples of 0.3-seconds length audio speech with each audio speech's seven timbral properties. The timbral features for each audio were firstly generated from the tool produced by AudioCommons and then filtered by human intuition. The 400 short audio speeches were randomly selected from LibriSpeech clean dataset. This dataset was used to train the seven individual feature extractor regressors.

4.3 Evaluation

Relative and sharable performance measures are required to estimate how superior an algorithm or approach is. The major problem for evaluating any method is adopting training and testing sets, which can introduce an inconsistency in model performance. Most of the performance metrics are based upon the confusion matrix, which consists of true positive (TP), true negative (TN), false positive (FP), and false-negative (FN) [59] values. The significance of these elements can vary on how the performance evaluation is done.

The term 'recognition' can be classified into two separate operations, identification and verification. The identification system seeks persons' identity, whereas the verification systems only check if the person is the one whom it is expected to be. The proposed approach is tested both of the scenarios, and evaluation data are presented in this section.

The accuracy of an identification system can be defined by how many correct guesses the model estimates from the model's total estimations. The accuracy is measured as,

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

To evaluate the verification system, the Receiver Operating Characteristics Curve (ROC) and Equal Error Rate EER are calculated. The ROC curve is a well-known non-parametric estimation method in biometric authentication and verification systems [32]. The ROC curve generates a visual of the probability of correct detection (True Positive Rate or TPR) versus the possibility of false alarm (False Positive Rate or FPR). The area generated by the ROC curve is known as the area under the curve (AUC). A higher value of AUC ensures the robustness of the verification system. EER can be evaluated from the ROC curve, by pointing the position where TPR is higher than FPR and TPR + FPR = 1. Lower EER value confirms the robustness of a verification system.

4.4 Results and Discussion

4.4.1 Speaker Identification

Speaker identification is the process of targeting a speaker by hearing the voice. In terms of machine learning, speaker identification is a multiclass classification problem. Figure 4.1 represents the identification accuracy of

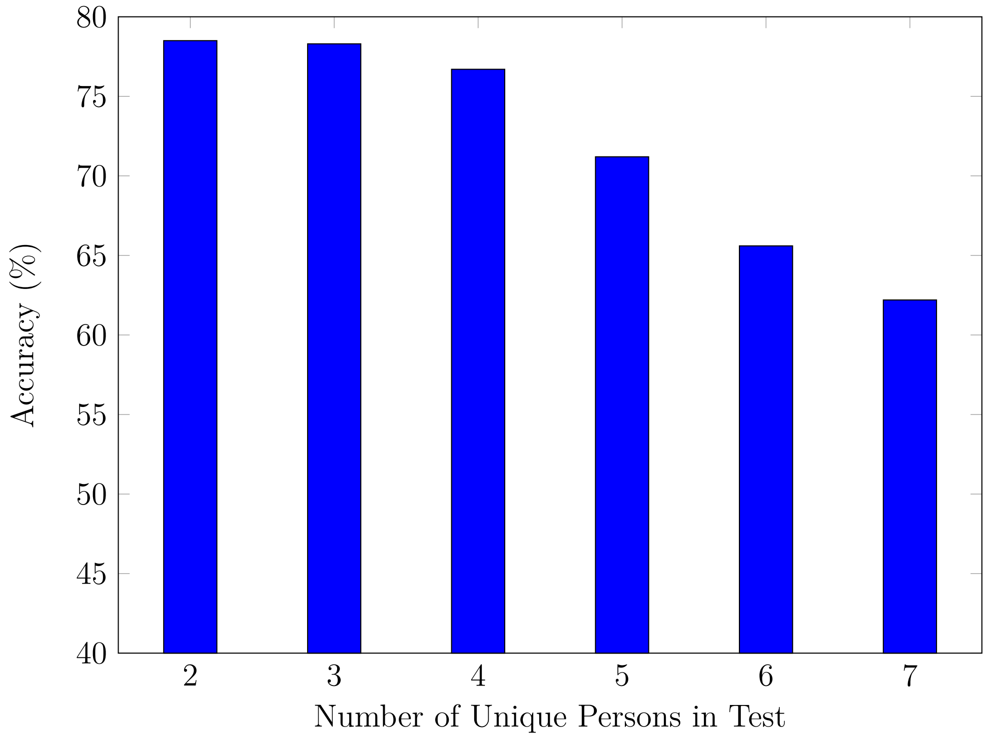

Figure 4.1. The graph illustrates the accuracy score of the speaker identification phase of the proposed architecture. The vertical axis represents the accuracy scale, whereas the horizontal scale represents the number of unique persons introduced in the identification phase.

the proposed architecture while presenting a different number of persons. The prototype's performance degrades concerning the increasing number of individual persons. The degradation points to the characteristics of the features. The features which are extracted and used in our model are densely associated with each other. Therefore, the classifier fails to fit on training data appropriately. This degradation points out that the model can only be used for a small group of individuals for identification purposes.

4.4.2 Speaker Verification

Speaker verification is the method of confirming if the voice is of a specific person. Aside from the model's identification score's unbalanced accuracy, it presents a better performance in speaker verification. In terms of machine learning, speaker verification is stated as a binary classification problem. Figure 4.2 illustrates the accuracy scores of the model, including a different number of individuals in the verification phase. The proposed model generates a satisfactory score in the speaker verification phase. It shows accuracy above 80% in most of the tested environments. The model continuously provided stable accuracy while increasing the number of unique persons.

Figure 4.3 represents the ROC curve of the proposed model that is tested on a random individual. The proposed model gives an equal error rate (EER) of 0.24, while the area under the curve (AUC) being 0.84. The equal error rate represents that the model generates its best result in verifying an individual from a continuous stream of audio.

4.5 Summary

From the evaluation reports, it is evident that this architecture performs most satisfying on speaker verification tasks. Although the method is also suitable for speaker identification tasks, increasing the number of unique people reduces the model's accuracy.

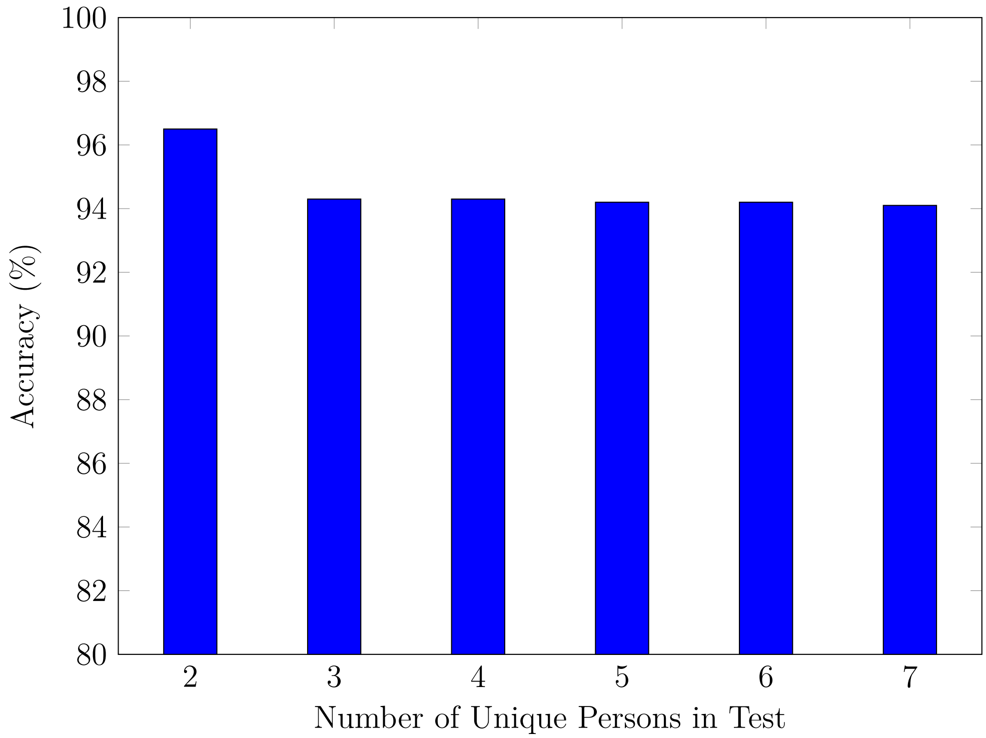

Figure 4.2. The graph illustrates the accuracy score of the speaker verification phase of the proposed architecture. The vertical axis represents the accuracy scale, whereas the horizontal scale represents the number of unique persons introduced in the identification phase.

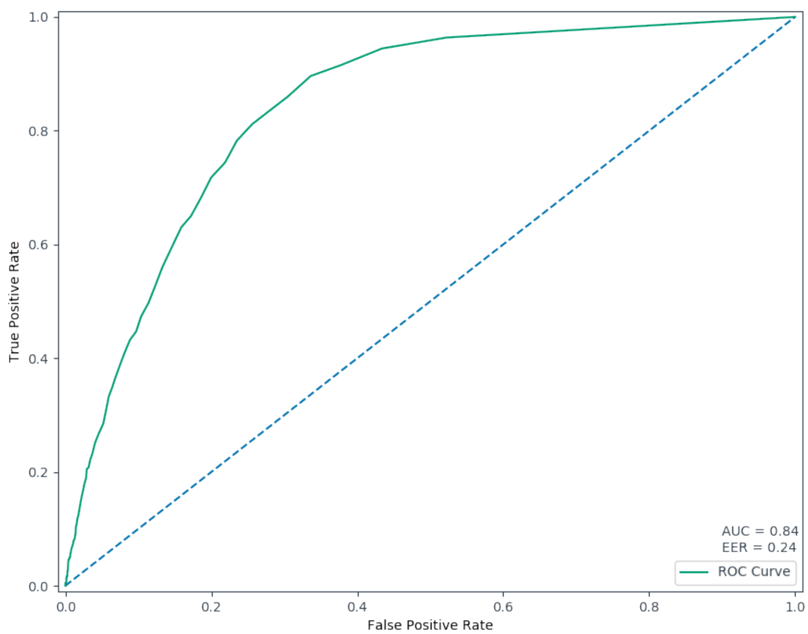

Figure 4.3. he figure represents a ROC curve of the model. The curve is generated based on identifying a random individual from the dataset. The model generates an EER of 0.24, while the AUC is 0.84.

Standards, Impacts, Ethics, and Challenges

5.1 Impacts on Society

The speaker recognition system has a wide area of impact on the usage of the system. The speaker recognition system can be used as a speech diarization system to auto-generate speech into dialogue form. As a result, this can be implemented at a national level conference or meetings to retain speech proof. The system can also be implemented to gather information about any fraud voice calls if speech data is kept nationally. Implementing speaker recognition systems with speech recognition systems will also benefit automated robot industries as it is possible to pinpoint user commands with the assistance of speaker recognition systems.

5.2 Ethics

The speaker recognition system has a broader usage level, depending on the data that is applied to prepare the model. The usage of speaker recognition systems must maintain individuals' privacy concerns and should not be used for any purpose that raises a national or social security threat. The usage, along with the dataset gathering, must be performed under the code of moral principles.

5.3 Challenges

Although modern speaker recognition technologies are evolving rapidly, the companies developing such technologies still face information security challenges. This thesis work has clearly shown that the currently used voice authentication systems are close to real-world implementation. Frequently, speaker recognition and authentication systems are protected against hacker attacks, including voice cloning attacks.

5.4 Summary

Nevertheless, it should be noted that, despite their authentication problems, speaker recognition technologies can be a well-suited supplement to other biometric methods, such as fingerprints, face recognition, and iris recognition. Authentication systems relying on the identification of several biometric characteristics are also known as multimodal biometric systems. Recent studies indicate that multimodal biometric systems are more secure than biometric systems depending on one biometric method.

Constraints and Alternatives

6.1 Design Constraints

The overall structure of the proposed architecture can be implemented based on continuous or segmented audio frames. The model requires devices with high processing capability to perform simultaneous speaker recognition. The model does not require any GPU support.

6.2 Component Constraints

The component requirements of the proposed architecture include,

- Minimum Processor Requirement: Intel i3 (7th Gen, 3GHz)
- Minimum Memory Requirement: 4GB (DDR3, 1600 bus)
- Audio Input: HD Audio Input Device

6.3 Budget Constraints

The estimated budget is to be calculated by the current market price of the component requirements.

6.4 Summary

The avoidance of deep learning in the proposed speaker recognition architecture can be implemented in lightweight devices that are cost-efficient and available.

Schedules, Tasks, and Milestones

7.1 Timeline

The overall timeline of the thesis work can be segmented into three divisions based on the three semesters of our supervisor's work execution procedure. The first-semester work process contains the planning and reviewing of the related works of the thesis work. The second-semester work process includes collaborative work of prototype designing and analysis of the prototype. In the third semester, we implemented and tested the overall architecture and reported the overall workflow.

7.2 Gantt Chart

Figure 7.1 contains the Gantt chart describing the work execution process of the thesis work. The thesis work's overall execution is three semesters long, where each semester is twelve weeks long.

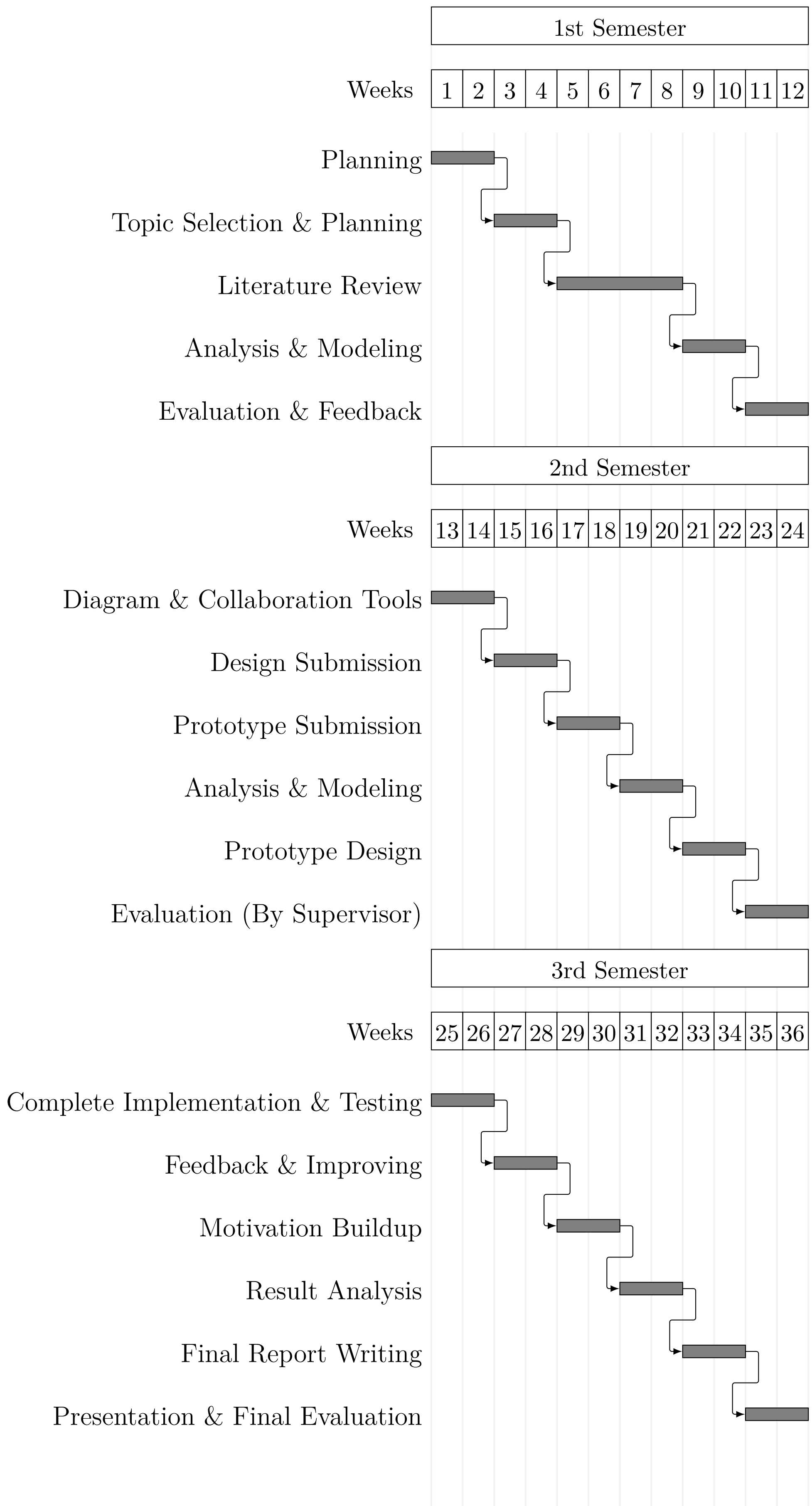

Figure 7.1. Gantt chart of the work execution process.

Conclusion

8.1 Introduction

In this thesis, we proposed a model that uses the timbral properties of voice, that is hardly used in any other research endeavors. The model is tested against a real-world continuous stream of audio, without any modification. Although the model almost fails in the speaker identification phase, it achieves a marginal score in the speaker verification phase. The model's accuracy can be improved if the scaling of the features is estimated more accurately. As the paper introduces new speech properties, further studying these features illustrated in this paper, the speaker recognition system researchers will be motivated to try out the vocal sound properties rather than only using sound waves or identity vectors as features. Therefore, we believe this research effort will influence the research to explore new speech properties that may invent more robust and lightweight architectures.

8.2 Future Works and Limitations

Although the speaker recognition system performs excellently as a speaker verification system, the main limitation of speaker recognition systems is the decrease of accuracy concerning the increasing number of individuals on which the identification is processed. We tend to solve this particular challenge of the proposed architecture. We would also develop a speech synthesis system that would synthesize speech from noisy environments that would be joined with this architecture and improve the overall performance of the proposed architecture.

References

[1] Douglas A Reynolds. An overview of automatic speaker recognition technology. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages IV–4072. IEEE, 2002.

[2] Douglas A Reynolds, Thomas F Quatieri, and Robert B Dunn. Speaker verification using adapted gaussian mixture models. *Digital signal processing*, 10(1-3):19–41, 2000.

[3] Shi-Xiong Zhang, Zhuo Chen, Yong Zhao, Jinyu Li, and Yifan Gong. End-to-end attention based text-dependent speaker verification. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 171–178. IEEE, 2016.

[4] Homayoon Beigi. Speaker recognition. In *Fundamentals of Speaker Recognition*, pages 543–559. Springer, 2011.

[5] Sadaoki Furui. Speaker-independent and speaker-adaptive recognition techniques. *Advances in Speech signal processing*, pages 597–622, 1992.

[6] Sadaoki Furui. Recent advances in speaker recognition. *Pattern recognition letters*, 18(9):859–872, 1997.

[7] Mireia Diez, Lukas Burget, and Pavel Matejka. Speaker diarization based on bayesian hmm with eigenvoice priors. pages 147–154, 06 2018.

[8] Mirco Ravanelli and Yoshua Bengio. Speaker recognition from raw waveform with sincnet. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 1021–1028. IEEE, 2018.

[9] Zheli Liu, Zhendong Wu, Tong Li, Jin Li, and Chao Shen. Gmm and cnn hybrid method for short utterance speaker recognition. *IEEE Transactions on Industrial informatics*, 14(7):3244–3252, 2018.

[10] Fred Richardson, Douglas Reynolds, and Najim Dehak. Deep neural network approaches to speaker and language recognition. *IEEE signal processing letters*, 22(10):1671–1675, 2015.

[11] Mitchell McLaren, Yun Lei, and Luciana Ferrer. Advances in deep neural network approaches to speaker recognition. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4814–4818. IEEE, 2015.

[12] George E Dahl, Dong Yu, Li Deng, and Alex Acero. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on audio, speech, and language processing*, 20(1):30–42, 2011.

[13] Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez Moreno, and Javier Gonzalez-Dominguez. Deep neural networks for small footprint text-dependent speaker verification. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4052–4056. IEEE, 2014.

[14] Shane Settle, Jonathan Le Roux, Takaaki Hori, Shinji Watanabe, and John R Hershey. End-to-end multi-speaker speech recognition. In

2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 4819–4823. IEEE, 2018.

[15] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5329–5333. IEEE, 2018.

[16] Md Hafizur Rahman, Ivan Himawan, Sridha Sridharan, and Clinton Fookes. Investigating domain sensitivity of dnn embeddings for speaker recognition systems. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5811–5815. IEEE, 2019.

[17] Prashant Anand, Ajeet Kumar Singh, Siddharth Srivastava, and Brejesh Lall. Few shot speaker recognition using deep neural networks. *arXiv preprint arXiv:1904.08775*, 2019.

[18] Xin Fang, Liang Zou, Jin Li, Lei Sun, and Zhen-Hua Ling. Channel adversarial training for cross-channel text-independent speaker recognition. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6221–6225. IEEE, 2019.

[19] Johan Rohdin, Themos Stafylakis, Anna Silnova, Hossein Zeinali, Lukáš Burget, and Oldřich Plchot. Speaker verification using end-to-end adversarial language adaptation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6006–6010. IEEE, 2019.

[20] Xu Xiang, Shuai Wang, Houjun Huang, Yanmin Qian, and Kai Yu.

Margin matters: Towards more discriminative deep neural network embeddings for speaker recognition. *arXiv preprint arXiv:1906.07317*, 2019.

[21] João Antônio Chagas Nunes, David Macêdo, and Cleber Zanchettin. Additive margin sincnet for speaker recognition. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–5. IEEE, 2019.

[22] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018.

[23] Hannah Muckenhirn, Mathew Magimai-Doss, and Sebastien Marcell. Towards directly modeling raw speech signal for speaker verification using cnns. pages 4884–4888, 04 2018.

[24] Amirsina Torfi, Jeremy Dawson, and Nasser M Nasrabadi. Text-independent speaker verification using 3d convolutional neural networks. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2018.

[25] Wayne Xiong, Lingfeng Wu, Fil Alleva, Jasha Droppo, Xuedong Huang, and Andreas Stolcke. The microsoft 2017 conversational speech recognition system. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5934–5938. IEEE, 2018.

[26] Rubén San-Segundo, Juan Manuel Montero, Roberto Barra-Chicote, Fernando Fernández, and José Manuel Pardo. Feature extraction from smartphone inertial signals for human activity segmentation. *Signal Processing*, 120:359–372, 2016.

[27] I Potamifis, Nikos Fakotakis, and G Kokkinakis. Improving the robustness of noisy mfcc features using minimal recurrent neural networks. In *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*, volume 5, pages 271–276. IEEE, 2000.

[28] Seiichi Nakagawa, Kouhei Asakawa, and Longbiao Wang. Speaker recognition by combining mfcc and phase information. In *Eighth annual conference of the international speech communication association*, 2007.

[29] 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.

[30] Todor Ganchev, Nikos Fakotakis, and George Kokkinakis. Comparative evaluation of various mfcc implementations on the speaker verification task. In *Proceedings of the SPECOM*, volume 1, pages 191–194, 2005.

[31] Vibha Tiwari. Mfcc and its applications in speaker recognition. *International journal on emerging technologies*, 1(1):19–22, 2010.

[32] Nivedita Palia, Shri Kant, and Amita Dev. Performance evaluation of speaker recognition system. *Journal of Discrete Mathematical Sciences and Cryptography*, 22(2):203–218, 2019.

[33] Osama S Faragallah. Robust noise mkmfcc–svm automatic speaker identification. *International Journal of Speech Technology*, 21(2):185–192, 2018.

[34] Saeid Safavi, Martin Russell, and Peter Jancovic. Automatic speaker, age-group and gender identification from children's speech. *Computer Speech and Language*, 50, 01 2018.

[35] Mitchell McLaren and David Van Leeuwen. Source-normalised-and-weighted lda for robust speaker recognition using i-vectors. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5456–5459. IEEE, 2011.

[36] Ondřej Glembek, Lukáš Burget, Niko Brümmer, Oldřich Plchot, and Pavel Matějka. Discriminatively trained i-vector extractor for speaker verification. In *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[37] Johan Rohdin, Anna Silnova, Mireia Diez, Oldřich Plchot, Pavel Matějka, Lukáš Burget, and Ondřej Glembek. End-to-end dnn based text-independent speaker recognition for long and short utterances. *Computer Speech and Language*, 59:22–35, 2020.

[38] Ondřej Novotnỳ, Oldřich Plchot, Ondřej Glembek, Lukáš Burget, and Pavel Matějka. Discriminatively re-trained i-vector extractor for speaker recognition. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6031–6035. IEEE, 2019.

[39] Longting Xu, Kong Aik Lee, Haizhou Li, and Zhen Yang. Generalizing i-vector estimation for rapid speaker recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(4):749–759, 2018.

[40] Xingyu Zhang, Xia Zou, Meng Sun, Thomas Fang Zheng, Chong Jia, and Yimin Wang. Noise robust speaker recognition based on adaptive frame weighting in gmm for i-vector extraction. *IEEE Access*, 7:27874–27882, 2019.

[41] Themos Stafylakis, Johan Rohdin, Oldrich Plchot, Petr Mizera, and Lukas Burget. Self-supervised speaker embeddings. *arXiv preprint arXiv:1904.03486*, 2019.

[42] Aonan Zhang, Quan Wang, Zhenyao Zhu, John Paisley, and Chong Wang. Fully supervised speaker diarization. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6301–6305. IEEE, 2019.

[43] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5329–5333. IEEE, 2018.

[44] Jesús Villalba, Nanxin Chen, David Snyder, Daniel Garcia-Romero, Alan McCree, Gregory Sell, Jonas Borgstrom, Leibny Paola García-Perera, Fred Richardson, Réda Dehak, et al. State-of-the-art speaker recognition with neural network embeddings in nist sre18 and speakers in the wild evaluations. *Computer Speech and Language*, 60:101026, 2020.

[45] Zili Huang, Shuai Wang, and Kai Yu. Angular softmax for short-duration text-independent speaker verification. In *Proc. Interspeech 2018*, pages 3623–3627, 2018.

[46] Ya-Qi Yu, Lei Fan, and Wu-Jun Li. Ensemble additive margin softmax for speaker verification. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6046–6050. IEEE, 2019.

[47] Qing Wang, Wei Rao, Sining Sun, Leib Xie, Eng Siong Chng, and Haizhou Li. Unsupervised domain adaptation via domain adversarial training for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4889–4893. IEEE, 2018.

[48] Ahmed Isam Ahmed, John P Chiverton, David L Ndzi, and Victor M Becerra. Speaker recognition using pca-based feature transformation. *Speech Communication*, 110:33–46, 2019.

[49] Phani Sankar Nidadavolu, Saurabh Kataria, Jesús Villalba, and Najim Dehak. Low-resource domain adaptation for speaker recognition using cycle-gans. *arXiv preprint arXiv:1910.11909*, 2019.

[50] Sung-Hwan Shin, Jeong-Guon Ih, Takeo Hashimoto, and Shigeko Hatano. Sound quality evaluation of the booming sensation for passenger cars. *Applied acoustics*, 70(2):309–320, 2009.

[51] Hugo Fastl. Psycho-acoustics and sound quality. In *Communication acoustics*, pages 139–162. Springer, 2005.

[52] Travis E Oliphant. Python for scientific computing. *Computing in Science & Engineering*, 9(3):10–20, 2007.

[53] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter

Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[54] Stéfan van der Walt, S Chris Colbert, and Gael Varoquaux. The numpy array: a structure for efficient numerical computation. *Computing in Science and Engineering*, 13(2):22–30, 2011.

[55] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8, 2015.

[56] John D Hunter. Matplotlib: A 2d graphics environment. *Computing in science and engineering*, 9(3):90–95, 2007.

[57] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210. IEEE, 2015.

[58] AudioCommons. *AudioCommons Materials*, Accessed September 26, 2020.

[59] Jake Lever, Martin Krzywinski, and Naomi Altman. Points of significance: classification evaluation, 2016.

Appendices