School of Information Technology and
Engineering at the
ADA University

School of Engineering and Applied
Science at the
George Washington University

TEXT-DEPENDENT SPEAKER IDENTIFICATION

A Thesis
Presented to the Graduate Program of Computer Science and Data Analytics
of the School of Information Technology and Engineering
ADA University

In Partial Fulfillment
of the Requirements for the Degree
Master of Science in Computer Science and Data Analytics
ADA University

By
Natavan Akhundova

April, 2022

THESIS ACCEPTANCE

This Thesis by: Natavan Akhundova
Entitled: *Text-Dependent Speaker Identification*

has been approved as meeting the requirement for the Degree of Master of Science in Computer Science and Data Analytics of the School of Information Technology and Engineering, ADA University.

Approved:

Dr.Samir Rustamov	Rustamov S	28.04.2022
(Adviser)		(Date)
Dr. Abzatdin Adamov		28.04.2022
(Program Director)		(Date)
Dr. Sencer Yeralan		28.04.2022
(Dean)		(Date)

ABSTRACT

Speaker identification is a process of identifying a person who is speaking and is very useful in applications such as customer service or even in investigations and reporting forensic evidence. This study focuses on finding the relationship between the latest state-of-art technology in speaker recognition which is x-vectors, and the uttered text within audio signals, as well as, the duration of them. In order to accomplish that, three different datasets are used: two relatively small digits datasets in English and Azerbaijani, and one larger dataset of digits and commands in Azerbaijani. The hypotheses tested in this research are as following: 1) x-vectors hold the information about the text in audio recordings, and the accuracy of the model changes as the text is changed; 2) x-vectors show better accuracy with longer audio recordings than shorter ones. All three datasets were trained to test the first hypothesis and the findings show that when the models are given audio samples in which a new unseen text is uttered, the accuracy decreases drastically. The last dataset was used to test the second hypothesis. Indeed, x-vectors are data-hungry and more speech samples together with longer duration of recordings gave the best results. Although, most of the experiments are conducted in the Azerbaijani language, it is believed that the results are not related to the specific language. Moreover, testing these hypotheses with a dataset of another language will yield the same results, as proved with the English dataset in this study.

Additional Keywords and Phrases: speaker recognition, speaker identification, TDNN, x-vectors, text-dependent, duration

TABLE OF CONTENTS

ACKLOWDEGMENT

This work is conducted with the help of ADA University, Baku, Azerbaijan, ATL Tech company, Baku, Azerbaijan and the George Washington University, Washington DC, US.

LIST OF FIGURES

LIST OF TABLES

Figure Caption Page

LIST OF ABBREVIATIONS

Abbreviation	Explanation
ANN	Artificial Neural Networks
CNN	Convolutional Neural Networks
DAC	Digits And Commands
DKLT	Discrete Karhunen-Love Transform
DNN	Deep Neural Networks
EER	Equal Error Rate
GMM	Gaussian Mixture Model
GMM-HMM	Gaussian Mixture Model - Hidden Markov Model
GMM-SVM	Gaussian Mixture Model - Support Vector Machine
GMM-UBM	Gaussian Mixture Model - Universal Background Model
HMM	Hidden Markov Model
IVR	Interactive Voice Response
JFA	Joint Factor Analysis
KNN	K-Nearest Neighbors
LPC	Linear Predictive Coefficients
LPCC	Linear Predictive Cepstral Coefficients
MFCC	Mel Frequency Cepstral Coefficients
MNIST	Modified National Institute of Standards and Technology
NFCC	Nonuniform-Frequency Cepstral Cefficients
PCA	Principal Component Analysis
PLDA	Probabilistic Linear Discriminant Analysis
PLP	Perceptual Linear Predictive
PLPCC	Perceptual Linear Predictive Cepstral Coefficients
RCC	Real Cepstral Coefficients
ReLU	Rectified Linear Unit
RMS	Root Mean Square
SVM	Support Vector Machine
TDNN	Time Delay Neural Networks
UBM	Universal Background Model
UFCC	Uniform-Frequency Cepstral Coefficients
VAD	Voice Activation Detection
VQ	Vector Quantization
WCNN	Within-Class Covariance Normalization
ZCR	Zero crossing

1 INTRODUCTION

This section gives an introduction about the topic, and below subsections subsequently provide the definition of the problem, objective of the study, significance of the problem, a review of significant research, and last but not the least, assumptions and limitations.

1.1 Definition of the Problem

Identifying a voice is a very basic understanding of a human being on sounds. Even from a very little age, a child can identify if a person calling him/her from another room is his/her mother, father, grandmother or grandfather without seeing the caller. We start forming this function of our brain at an early age. For example, when it comes to recognizing a mother's voice, babies are able to distinguish their mother's voice from other women starting from the very first day [1]. However, it is still unknown how our brain works and performs its duty in identifying voice and speech. In traditional understanding, speaker and speech recognition are two different functions where the first one occurs in the right hemisphere of the brain, whereas, the latter recognition occurs in the left one [2]. However, recently in 2010, it was discovered that the left hemisphere also encodes information about identifying speakers and these two functions do not work separately and are not isolated, but rather it is a network of parts of the brain that work together [2]. The area of the brain responsible for speaker and speech recognition was identified, however, it seems like the process regarding how our brain works on speech and speaker recognition is still an ongoing research and yet, a mystery for researchers.

Nevertheless, despite the exact process of identifying speakers being unknown, there is plenty of research going on to accomplish this task with the lowest error possible. Over the years people have tried to identify people based on their voices for a plethora of reasons. Identifying voice accurately is most importantly needed in crime investigation to investigate the subject and in law to derive correct judgment on innocence or guilt. Addiotionally, banks also make use of information about a speaker currently on the phone for better targeted service or for protection from fraudulent calls. However, it is not a simple and straightforward task to accomplish as it appears in spy movies.

It is true that a human voice can act as an identification method since our voices are unique due to many factors. Our throat, nose, nasal cavity and oral cavity shape the buzzing sound coming from our vocal cords and make a unique voice. Because of glottis size, vocal tract shape, as well as, the shapes and sizes of other voice producing organs, no two human-beings share identical voices. However, voiceprint identification is not as reliable as fingerprint or eye retina print. There are many reasons contributing to it. One of the reasons is that our voices change depending on our mood, health and lastly on age. A person in a happy mood may sound different than in a sad mood. In addition, cold, flu, or a dentist appointment can also affect our voices. Another reason is that sounds rarely are without background noise. Be it street noises, office noises or any other environmental sound, the variable feature of background makes the task of identifying speakers harder for humans, as well as, computers. The key point to understand is we, humans, also make mistakes in recognizing voices. Our brain can mistakenly think of two different voices as similar and belonging to the same person even though they are different, or it also happens that we cannot identify the voice on the phone when the person talking is ill. Additionally, if we have only one way of conversing with a person, say with a normal tone, we probably mistakenly recognize their voice when they sing, laugh or shout. When humans make erroneous assumptions like these in their usual lives, expecting a computer to perform in one hundred percent accuracy in recognizing speakers is not adequate. Therefore, most of the research is conducted with predefined text or text independently, but with many thousands hours of speech to perform relatively well.

Although there are many difficulties in the path, the research on speaker recognition is going forwards because of the need in governmental agencies and businesses. The future of speaker recognition promises very exciting capabilities such as using your car with your voice or signing a transaction without long and complicated steps.

1.2 Objective of the Study

The role of machine learning algorithms in accomplishing the task of speaker identification is immaculate. Since 1960, algorithms have been developed to classify recordings and recognize speakers. The earlier algorithms focused on how to differentiate separate recordings using their variability in channels and speakers, later the focus was more on how to represent each recording in a vector space. The later approach proved to be better than saving variabilities of each recording. The vector space not only represents utterances more mathematically, but also enables us to observe how similar or different utterances are in the space relatively well.

The latest state-of-art methodology in this area was developed in 2018 and is called x-vectors [3]. These vectors represent recordings and are created using deep neural networks. Their purpose is to embed utterances into numbers in vectors in order to move from our typical understanding of audio with signals and frequencies to the mathematical world. After, vector comparison and distance finding methods help to differentiate those vectors easily and help us derive conclusions on the task of speaker recognition. The authors of this embedding method claim it better represents utterances and achieve high accuracy in verifying them.

The technology discussed earlier is the main method of developing and experimenting speaker identification models in this paper. This research focuses on developing a speaker identification model using the latest state-of-art methodology to further test the relationship between identification and uttered texts and also the relationship between identification and duration of utterances. To accomplish these tasks, datasets in English and Azerbaijani languages were used. The English dataset was used to test the relationship of identification with uttered texts, and the Azerbaijani datasets were used to test both text relationship and duration relationship.

The tested hypotheses are that a speaker identification model developed with x-vectors will hold information about the text uttered in recordings and the accuracy possibly increase or decrease depending on the similarities in text in the test set. There is a similar study which tests the performance of separate speech and speaker recognition models and a combined speech and speaker recognition model [4]. According to the results, the latter correlated model performed better than models which extracted information only related to one task. This leads to the idea that speech and speaker recognition, in other words, utterance and spoken text are deeply correlated. Therefore, this research is aimed to test the hypothesis that a speaker identification model will have information about the uttered text without specifically underlying it or giving it as an input. The experiment consists of training a model with utterances of speakers containing a specified text and testing the same speakers with another text and with a combination of seen and unseen texts.

The next hypothesis is, as the duration of recordings for the enrollment of speakers increases, the accuracy of correctly identifying speakers should increase and vice-versa. Similar situation with us, human-beings, is that the more we hear a conversation of one person, it is easier for us later to distinguish this person's voice from others compared to the person we heard only once. The longer and more the recordings are, more information can be extracted from them, thus the model will differentiate speakers more easily. The paper experiments with the duration of the speech by making it longer and observes the relationship between duration of the speech and accuracy of identifying a speaker.

1.3 Significance of the Problem

Advancements in deep neural networks in recent years contributed to further development of algorithms on voice and speech to be more accurate and faster. It opens doors for more studies, research and experiments since the updated algorithms become widely available and take less time. Applications of voice and speech also increased in recent years. Considering the fact that speaking is a lot faster than typing, many assistance tasks are conducted via sound. Examples for these can be virtual assistants, voice activations and controlling in cars and Interactive Voice Response (IVR) systems.

One part of voice processing is speaker identification. The purpose of speaker identification is to recognize the speaker and answer the question "Who is speaking?". This information is very useful in many settings, such as in banks. The use cases of it can be instantly identifying users and personalizing the interaction. Also, a natural login process for chatbots and virtual assistants or in general digital channels also can be implemented. Furthermore, nowadays, two-factor authentication is very popular and demanding. Almost every website or app having a login process implements it to secure the interaction from malicious attempts. Thus, the second part of two-factor authentication can be conducted with speaker identification. Another purpose of speaker identification is criminal investigations. It dates back to the 1660s, where the trial about the death of Charles I in Britain applied the identification by voice for the first time in the history [4]. The voice of the subject in the court can be identified and further verdict could be given based on the findings.

To further underline the significance of this research, it can be said that no study has been conducted in speaker identification in the Azerbaijani language. There is only one research study related to the speaker identification task, which has been conducted to collect a dataset for speaker identification [5]. But no architecture of this field was put under the test and experiments were not conducted. Considering similar languages, speaker identification has been conducted in Turkish, but with old technology [6]. Furthermore, most of the studies in speaker recognition tests the technologies or architectures with a large dataset named VoxCeleb [7], which has 7 million utterances, and very few focus on their performance on smaller datasets. Also, most of the studies are text-independent, which creates a gap for text-dependent studies to fill in.

This study experiments with the latest state-of-art technology in speaker identification using relatively smaller datasets both in English and Azerbaijani and further tests two hypotheses about the technology: whether it has relationships with a given text in utterances and duration of utterances so that the change of one will impact the accuracy.

1.4 Review of Significant Research

The technologies developed to identify speakers based on their voice have evolved a lot during the last 60 years. In the last century two embedding methods, namely i-vectors and x-vectors, have been developed to identify speakers.

The study in [8] implements an approach consisting of i-vectors with probabilistic linear discriminant analysis (PLDA) as a backend technique to test its performance with various features. The tested features are Perceptual Linear Predictive (PLP), Mel Frequency Cepstral Coefficients (MFCC) and a combination of them. Using the conversational datasets of, such as NIST SRE 2004-2010 and the Switchboard corpora, the authors found that fusion of PLP and MFCC gave better results than MFCC alone.

Yet another interesting research is conducted by N. Ibrahim and D. Ramli [9]. They applied a speaker recognition task on the voices of frog species with i-vector methodology. They experimented with 3 sizes of Gaussians and 3 dimensions of vectors and found out that the smallest size of Gaussian combined with the largest dimension of i-vectors gave the most accurate results in recognizing and classifying sounds of frogs.

The main technological method used in this study is x-vectors [3]. The authors compared their new approach with a popular method of i-vectors to test how well the architecture performs with larger datasets. The datasets are Speakers in the Wild and NIST SRE 2016 Cantonese. The authors concluded that with the help of augmentation, such as noise and reverberation, x-vectors derived from the deep neural networks performed significantly better than i-vectors.

D. Raj, D. Synder, D. Povey and S. Khudanpur, the common authors from the previous study, have experimented more with x-vectors to find out what additional information is embedded in them [10]. They have taken into account several information pieces, such as speaker, uttered text, channel, duration of a recording and augmentation type. Moreover, the findings were compared also with i-vectors and the information they hold inside over various vector dimension sizes, like 128, 256, 512, and 768. The tested dataset was RedDots dataset, and the authors found that without augmentation for x-vectors, both embedding achieve similar results in distinguishing and classifying various information about speech samples, however, when augmentation is added to DNN, the performance of x-vectors drastically increase.

Another research about x-vectors from different authors was conducted [11]. In this research L. Gerlach, F. Kelly and A. Alexander, tested the newest state-of-art architecture with a British-English dataset with 6000 telephone recordings from 600 speakers for speaker profiling. This dataset is additionally separated into two categories in terms of recording conditions, which are landline and mobile. In all cases, x-vectors achieved better results, even for the case of mobile conversations. While i-vectors scored with 15% Equal Error Rate (EER) in speaker profiling, for x-vectors this number was 1-3%.

Next study in the field of speaker recognition informs about the caution which concerns how optimistic researchers are about the certainty of the speaker recognition task [12]. The authors list several conditions in which such systems can perform ideally which are: recording processes are controlled, speakers are not trying to fool the system when recording their voice by changing it, speech from test and train sets should not differ much in recording conditions. Additionally, prohibiting speech synthesis tools and making a task from text-independent to text-dependent will improve results.

In [13], the authors introduced the advantages of time delay neural networks (TDNN) over ordinary deep neural networks (DNN) in long temporal contexts by their speed and accuracy. The authors explained that ordinary DNN starts working on long contexts from the first layer, while TDNN starts with short context and later layers explore the information hidden in longer temporal contexts. By experimenting with a popular Switchboard dataset with a speech recognition task, they also found that the performance of TDNN is a lot better than DNN by 2-6%.

An extended version of TDNN was introduced in 2020, which claims to surpass the abilities of vanilla TDNN [14]. This methodology is named as ECAPA-TDNN and adds the emphasized channel attention, propagation and aggregation to TDNN which is where the name comes from. By utilizing the attention mechanism, ECAPA-TDNN showed improvement in the task of speaker verification with ResNet datasets. This methodology also utilizes well in other domains and settings.

Another study discusses the challenges of speaker recognition and compares the existing methods [15]. The compared technologies are template matching, nearest neighbor, hidden markov models and neural networks. For example, the challenges of neural networks are described as computationally expensive and not guaranteed to generalize. Additionally, feature extraction techniques are also discussed such as Linear Predictive Coefficients (LPC), Linear Predictive Cepstral Coefficients (LPCC), Mel Cepstral Coefficients (MFCC), Perceptual Linear Predictive Cepstral Coefficients (PLPCC) and Real Cepstral Coefficients (RCC). The paper also underlines the applications of speaker recognition, gives factors that affect the accuracy and lists 9 features to be satisfied for speaker identification to be a biometric system. The authors emphasized that speaker identification can be a biometric system, but limitations still exist and need to be solved in the future.

A similar article focused on challenges and difficulties, as well as, ideal cases of speaker recognition alongside with comparing methodologies [16]. Constant referral to the human brain, how it functions and acts lets the readers understand and draw a line of difference between a human and a machine processing and distinguishing voice.

The next study named "AZ-SRDAT - A Speech Database for Azerbaijani Language" has contributed to the Azerbaijani community by collecting voice samples in the Azerbaijani language for the task of speaker recognition [5]. 86 speakers with a female to male ratio 75:25, have uttered digits, isolated words, combinations of digits and a paragraph. The authors did not define the portion of utterances for enrollment and the other for testing, to let researchers choose which part of the dataset is more suitable for their studies, short utterances or long paragraph utterance. The recordings are at 16 kHz.

Similar to this research, a text-dependent speaker recognition system has been developed in the Turkish language [6]. The system performs speaker verification using Gaussian Mixture Model - Universal Background Model (GMM-UBM), a little bit older technology, to identify Equal Error Rate (EER) with a Turkish dataset consisting of 46 speakers. The EER was found to be 5.73% and further studies to increase the accuracy is needed as it is noted by the authors.

Some studies attempted to compare and define the best feature extraction methods and filter banks. One of such studies has been conducted with uniform-frequency cepstral coefficients (UFCC), nonuniform-frequency cepstral coefficients (NFCC) and mel-frequency cepstral coefficients (MFCC), together with filter banks as Mel-scale, uniform and non-uniform filter banks [17]. The dataset used to test the feature extraction methods is a popular TIMIT dataset, and the architecture for speaker recognition is GMM. The experiments have also been conducted with compressed speech which adds a significant insight into the features, in general. Additionally, real time recognition and identification in tv series like "Friends" are conducted. As an outcome, the author has concluded that UFCC performed the best in these experiments which means that high-frequency samples hold very important speaker information. However, the experiments with compressed speech resulted in favor of MFCC.

A similar study was conducted with not only different feature extraction techniques, but also combination of them with various architectures for speaker identification [18]. The tested techniques are mel-frequency cepstral coefficients (MFCC), linear predictive cepstral coefficients (LPCC) and perceptual linear prediction (PLP). The architectures are decision tree algorithms, Support Vector Machine (SVM), k-nearest neighbors (KNN) and neural networks. In addition, the study implemented Principal Component Analysis (PCA) and t-SNE for dimensionality reduction. The dataset is a very small one with only 15 speakers and each having 3 recordings. Nevertheless, the authors found that the best technique depends on the size of data, where with a small dataset MFCC and weighted KNN performed the best, while with a larger one MFCC together with PLP and weighted KNN gave the best results.

Yet another interesting study has been conducted with a speaker identification task on identical twins [19]. As we know, vocal tract and other voice organs define the voice, and if two people have very similar vocal tract shapes and sizes, the accuracy of speaker identification becomes questionable and challenging. The dataset in the experiments consisted of one read and one random sentence uttered by 9 male and 26 female twins. According to Hermann Kunzel, the author, although twin voice recognition is not a standard task, twins are not the exact copies of each other and inter-speaker differences can be found.

In [20], the speaker identification system for differentiating speakers with only isolated words is implemented with two settings: text-dependent and text-independent. The feature extraction is done with MFCC and UMRT, which is a transform used for image compression. The technology used in the study to train speaker recognition is neural networks. The dataset consists of 15 speakers each uttering 7 commands, such as "up", "down", "left", "right", "start", "pause" and "stop". Depending on the setting of the experiment, whether it is text-dependent or text-independent, either all commands are used in training, or some commands are excluded. The study concluded that MFCC together with UMRT performs better than MFCC alone. The best accuracy achieved for the text-dependent system is 97.91%, and for the text-independent system the accuracy is 94.44%.

The study conducted by O. Orman and L. Arslan attempted to identify the subbands in frequencies that yield better discrimination of speakers [21]. Utilizing Vector Ranking criteria, the authors found that 0−1000 Hz and 3000−4500 Hz are more significant to automatic speaker recognition systems than any other subbands and result in better performance when identifying speakers.

A book written by Sadaoki Furui gives a deep understanding of speech and speaker recognition [22]. The novel topic discussed in the book about speaker identification appears to be normalization and adaptation techniques which are not discussed in the majority of papers. The author introduces 2 techniques for normalization which is parameter-domain and likelihood normalization. The first method is effective for long utterances in text-dependent tasks, and performs well in reducing linear channel effects. However, due to its averaging methods over the entire recording, it inevitably removes some useful speaker and text information. Hence, this normalization technique does not perform well on short utterances. The second normalization technique uses probabilities as its main method to normalize utterances. The original idea of likelihood normalization is unreal to implement since it requires condition probabilities of all speakers. Thus, approximation methods have been developed to make use of this technique. One of such approximations is to treat some speakers as "core" speakers that represent the population and calculate only their conditional probabilities. Also, choosing randomly selected gender-balanced speakers to calculate probabilities is another approach.

Another research discusses speaker identification techniques and compares them on features that have been converted to vectors using Vector Quantization [23]. Those features are MFCC, LPCC and both of them combined. Traditional methodologies such as Hidden Markov Model (HMM), artificial neural networks (ANN) are put under test, together with rarely used techniques for speaker identification such as Principal Component Analysis (PCA) and Histograms. The results show that histogram based approach performed better with Vector Quantized features.

In [24], a speaker recognition task has been solved on embedded systems [24]. The stuy experimented with MFCC features and developed an architecture of GMM to accomplish the task of speaker recognition in embedded systems for home security. The authors concluded that such architecture is not very robust to noise and depends heavily on the environment. Additionally, in order to achieve higher accuracy values, it is necessary to use a state-of-art methodologies as stated by the authors.

The authors of [25] also conducted experiments to identify the best feature extraction method on a small database with a text-dependent setting. The tested feature extraction techniques were MFCC, LPC and a combination of them. The database is a very small set of speakers, particularly 20, 9 males and 11 females. The experiment tested the architecture on both speaker identification and verification. Surprisingly, LPC performed better than MFCC and also better than the combined version of features. It also stated to be more robust and stable.

Dr. Ghahabi in the study named "Deep Learning for i-Vector Speaker and Language Recognition" has proposed a deep learning as a backend module for the architecture of i-vectors to recognize speaker and the language they are speaking [26]. The author also combined probabilistic linear discriminant analysis (PLDA) which gave better performance. The main idea behind creating this architecture, according to the author, is that to eliminate phonetic or speaker labels as much as possible. Thus, a new vector representation is proposed and named as GMM-RBM, however, after testing with the NIST SRE 2010 dataset, the author concludes that the i-vectors perform better than GMM-RBM.

Another research also focuses on language identification using speaker recognition technologies. The study "Native Language Detection Using the I-Vector Framework" is a quite interesting study which aims to predict the native language of a speaker talking in the second language, particularly English [27]. For this purpose, the architecture of i-vectors has been utilized, and in addition, the modification of i-vectors has also been tested. This modification is

In [28], the authors attempted to reduce the computational cost of GMM-UBM model for the speaker verification task and the outcome was successful. The dataset consisted of 42 speakers and each speaker uttered in total 3 minutes. The authors claim that increasing vector shifts in MFCC features not only reduce the computational cost, but also increases the accuracy of speaker recognition task. The authors also experimented with the number of speakers dedicated to train universal background model and GMM mixtures, however, these experiments did not yield a difference in performance or cost.

As mentioned earlier, speaker recognition is a necessary technology that can be used in courts and criminal investigations. The next research is totally focused on such cases, where forensic evidence is reported using 3 different speaker recognition methodologies: GMM-UBM, Joint Factor Analysis (JFA) and i-vectors [29]. The methodologies are applied to Algerian Arabic dialect. The difference of this research from others is that the data collection of forensic evidence is not a controlled environment, meaning no one supervises the recording process, but the speech is acquired from other sources, such as telephone conversations. Considering the nature of this problem, GMM-UBM would need feature normalization and model transformation since the naïve version of the architecture will give unreliable outcomes. For evaluation techniques, Half Total Error Rates are used, and it is found that GMM-UBM performed the worst out of these three architectures, and JFA and i-vectors are more robust in this setting.

Some studies focus on real-time speaker identification and verification and therefore, experiment with the ways to efficiently optimize the process. One such study is [30] that researches ways to implement a real-time speaker recognition system. The authors define the challenge of speaker identification task in terms of time complexity which is the fact that time taken for speaker identification to complete its process flow depends on the likelihood computations of the given audio signal over the database of speaker models. Hence, the number of speakers, complexity of a speaker model, dimensionality and feature vectors contribute to the overall time of the pipeline. The authors have utilized Vector Quantization (VQ) in testing where, the quantization is applied in test sequence before matching. Moreover, the technique of pruning out the least probable speaker models also implemented. Together with GMM models, the authors have achieved sixteen times increase in the speed.

Similar to one of the objectives of this study, "Performance comparison of speaker recognition systems in presence of duration variability" experiments with the accuracy of the speaker recognition system while the duration is changed [31]. The experiments are conducted with GMM-UBM and i-vectors. It is known that speaker recognition systems heavily depend on the available speech data from each speaker. The number of datasets used in this study is two and they both belong to the NIST datasets. The first dataset is NIST SRE 2008 and the second one is NIST SRE 2010. Different cases of length of speech have been investigated, and the authors concluded that i-vectors become better when the length of test utterances increase. Additionally, the authors found that for short utterances if there is enough speech data for the model, GMM-UBM outperforms the i-vectors, which is a surprising finding since GMM-UBM is an older method and not a state-of-art technology.

Other authors also tackled the challenge of short utterances as an input to speaker recognition systems. The paper "Speaker Identification with Short Sequences of Speech Frame" focuses on short utterances, and even on utterances that are shorter than 1 second [32]. This study compares mel-frequency cepstral coefficients and its performance on short utterances, with the discrete Karhunen-Love transform (DKLT). The dataset that is worked on is in the Italian language and consists of 5 different audiobooks voiced by two female and three male speakers. The architecture for training a speaker identification model is chosen to be GMM. The study concluded that the DKLT performed better than MFCC, due to the fact that MFCC is great at speech recognition but has its own limitations and drawback when it comes to speaker identification.

After discussing the speaker identification problem with short length utterances, large-scale databases and challenges of speaker recognition with large databases can be investigated. The authors, L. Schmidt and M. Moreno, in the paper "Large-Scale Speaker Identification" have touched the problem that could arise in such systems, which is speaker identification needs to perform fast search on the large set of speakers [33]. Additionally, the larger the number of speakers exist in the dataset, the more is the probability for a trained model to make a mistake, thus, the accuracy values for models trained with large-scale datasets are generally low. The architecture examined is i-vectors and together with the given method, the authors also developed a searching algorithm. The algorithm is named as locality sensitive hashing and it finds the nearest neighbor in high dimensions very fast. The way L. Schmidt and M. Moreno connected i-vectors and hashing function is through cosine distance. While cosine distance compares i-vectors, locality sensitive hashing approximates the cosine distance in a quick manner. The dataset the authors used is acquired from Youtube with about one thousand speakers. It consists of more than 40 thousand 10-second, more than 40 thousand 20-second, and more than 20 thousand 60-second audio recordings. The study concludes that without changing or decreasing the accuracy of the speaker identification method, this hashing function improved the speed of the searching from one to two magnitude order.

Another research has been conducted to better represent audio signals and provide robust features since feature vectors are the most crucial part of speaker identification. The performance of the model heavily depends on how a method or technique represents the audio frames and what useful information it derives. Hence, the authors of the study [34] developed a new approach for a purpose of making audio features more robust. This new method is named hierarchical classification approach. It consisted of several layers that capture different information. For example, the first layer holds information about the gender of the speaker, and the next layer holds information about the characteristics of the speaker's voice. According to the study, other feature extraction methods also achieved good results, but the new architecture also reduces the computational time. The architecture used in this paper is simple random forest. The result is that for male speakers the identification model scored 78%, while for female speakers the model scored higher, 88.7%. Additionally, gender classification is performed at 96.9%.

The research in [35] suggests to add additional information to features of audio signals in order to increase the accuracy of speaker classification. The information to be added is duration of speech units such as phonemes or Hidden Markov Model (HMM) states that make up a phoneme. The study forms vectors using Gaussian mixtures and uses universal background model to capture speaker related information present in the audio signals. The findings of the study are that the accuracy has increased when the model uses the durational information about phonemes. The authors found that even more accurate results can be achieved if lexical features are also added to vectors. Moreover, the last finding of the study is that with longer test samples, accuracy again improves. This study is very similar to the current study, but with an older methodology.

The next study conducted by R. Karadaghi is named as "Open-set speaker identification" writing of which is motivated by criminal investigations [36]. Since the number of speakers possible speaking a voice in the audio is not known, this model needs to be open-set, meaning given audio may contain voice of unknown person or, in other words, the person out of our training dataset. The author explains the challenges of this task such as environment and background noise because usually audio speech used in criminal investigation are not recorded in a supervised manner, hence, the noises of street and voices of other people talking near are expected which makes the task more difficult. In addition, the duration of audio recordings also varies due to the same reason. By conducting this study, the author introduced a novel term called "vowel boosting". The results of the study showed that the vowels within the voice of a person have more speaker information that can be useful, thus, “vowel boosting” was introduced. It claims to increase reliability in speaker identification task where length of audio recordings varies.

Another study with the similar topic from the same author collectively working with H. Hertlein and A. Ariyaeeinia focuses on short audios that have various duration values [37]. Two architectures are compared on the same dataset which are GMM-UBM and i-vectors. The dataset used in this study is NIST speaker recognition evaluation corpus of 2008. The experiments showed quite interesting results that also align with previous papers discussed earlier. When the data for a model to learn is sufficient, i-vectors together with within-class covariance normalization (WCCN) used for variability compensation for the same speaker performed better than GMM-UBM, however, when the data is short and the duration of audio recordings varies, the performance and accuracy of i-vectors are not different from GMM-UBM. These experiments conclude that depending on the nature of data, its length and average duration, one should choose a speaker identification methodology carefully.

1.5 Assumptions and Limitations

This study of speaker identification assumes that tested speakers will be previously known ones from our database. It is a closed set task, meaning if other out-of-database speakers are given to the system, they will not be correctly identified as unknown, but will be mapped to any similar speaker. This is not secure for applications in any sort. Only case of application is that we are hundred percent sure that tested speakers will certainly be from our known set, which rarely happens in the real world. However, for the case of experiments and studies, this feature of the task can be tolerated.

Furthermore, not only speaker identification, but also any field working with human voices has to consider the fact that a person's speech is subject to change depending on age, health and emotional state. In addition, background noises are an inevitable part of recordings which affects the accuracy of systems working with human voices [38]. Background noise and the quality of the input device (the microphone) can create additional challenges for voice recognition systems. These factors make the systems not robust, since depending on the time of the day, background noise, health of a person and a gradual change in a person’s voice over years, the systems achieve lower accuracy values. This is an ongoing problem in the speech field and researchers try to partially solve it with augmentation. This study also utilizes the augmentation with various noise and reverberations in order to form a robust model.

Speaker recognition systems are also vulnerable to malicious attacks. If a recording of an authorized person would be acquired and replayed to the system, the system will unintentionally give a permission to an unauthorized person. This is called a replay attack. The way researchers and developers cope with this is storing each session uniquely and comparing the given recording with the database records [39]. The idea is that a person can not utter the words in the same exact manner and in the same exact background, thus, if the given recording exactly matches a record in the database, it means someone acquired a recording illegally and replays it back to the system. Replay attacks further included text-to-speech technologies to create a voice sample of the authorization phrase from speech samples of an authorized person. However, this approach needs a lot of speech data from an authorized person, and researchers have already developed methods to differentiate synthesized speech from a real one. To sum up, speaker identification and verification has its own challenges when it comes to security, and each side of the battle, either attackers or defenders, tries to come up with new methods to outsmart the other side. Although security is a huge concern for nowadays systems, this study does not focus on this aspect and does not perform experiments on the robustness of models on this matter.

2 RESEARCH APPROACH OR METHODOLOGY

The next subsections give insight about how humans generate voice and how technology captures it and acquires necessary information.

2.1 Anatomy of Speech

As human-beings, we are capable of understanding speech and differentiating voices of other human-beings. How are these voices generated? There are three processes going on sequentially when we speak. The first part is producing basic sound that will eventually be a spoken word. This basic sound is generated by vocal fold vibration and often called "buzz" sound. Organs actively taking part in this process are lungs and vocal ford. The second step is resonating an incoming sound. The organs participating in the resonance process are throat, oral cavity and nasal cavity. In this step, "buzz" sound is modified and amplified, producing a unique sound. Next, this amplified sound is modified by the organs that are last in this chain of speech production which are mouth, tongue, lips and teeth. These organs modify the incoming sound generating recognizable words. In summary, vibration, resonance and articulation form the process of speaking which is, in general, air in our lungs transforming to the spoken words that we understand. The second stage in this process produces a unique sound, hence, speaker recognition focuses heavily on this part. The last part in the process is where speech recognition mainly focuses on.

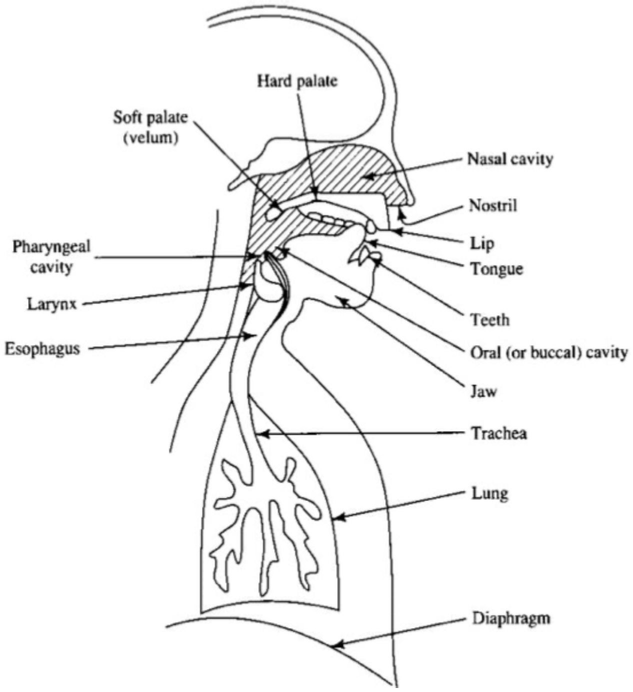

Figure 1: Voice production organs and anatomy of speech production. This picture is acquired from [40].

Due to the fact that our vocal tract shapes, glottis sizes and other voice producing organs are different, no two humans sound the same and each voice is unique. In addition, our voices are also unique because of different speaking and pronunciation styles and different choice of vocabulary. Our voices are so various that even the same person cannot utter the same phrase or sentence in the exact same manner. These factors, however, should not contribute to the fact that voice can be used as a biometric authentication which in fact, is not true. Even though two human-beings can be differentiated using the information about their voices due to the fact that they are certainly different, biometrically identifying a human-being with his/her voice cannot be implemented with a high accuracy. This is because voice is not as stable as fingerprint or eye retina, it changes not only in long-term as years due to aging, but also in short-term as days due to mood, health and time of the day. Because of these reasons, voice is not used by researchers and developers as a biometric authentication method, but it is still a very useful identifier and verifier method.

We understood how our voices are anatomically different, but how do computers understand the difference? First of all, the speech produced by us - humans needs to be converted and transformed to the format which computers would understand to further work on them. The sound is converted to binary format by transforming it to the digital sound from the original analogue version. This is done by dividing an audio recording into audio frames, which are a basic temporal fragments of the recording. Their length varies from 10 milliseconds to 50 milliseconds. Such digital coding attempts to represent the analogue signal accurately so that later, the reconstruction of the same signal would not lose its values. However, researchers are not concerned with reconstructing the audio signal, but representing it with useful parameters and features. Therefore, the frames that make up an audio signal are used to generate features that will represent the audio signal in the most informative way possible. The researchers and programmers have found two ways of representing sound, which are temporal based and spectral based [41].

Time based representation is soundwaves we usually see in music players. The sound is represented as frequencies over a time interval and depicted as waves in music players or any other programs working on sounds. The computer acquires numbers which are frequencies over a time interval and processes them. Below is a picture of the author saying a random sentence “I like this class, kind of” represented in a time-based manner (Figure 2).

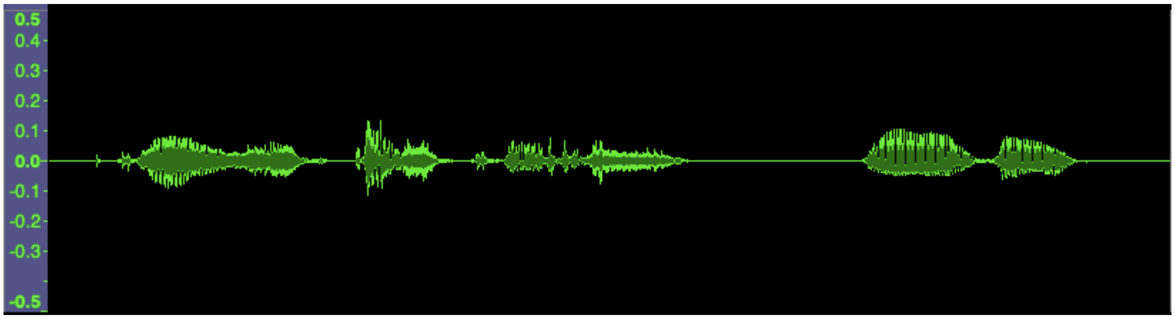

Figure 2: A visual representation of time-based sound representation of a sentence “I like this class, kind of” by the author.

On the other hand, spectral representations transform audio signal from time-space to frequency-space that has more information and easier to work with. They are distribution of amplitude over frequencies, and also called frequency spectrum. When we visually represent the spectrum of frequencies, we obtain an image which is called a spectrogram in speech field. Spectrograms display not a single number of a frequency over a time interval, but a spectrum of frequencies. Figure 3 displays the same sound in Figure 2 as a spectrogram.

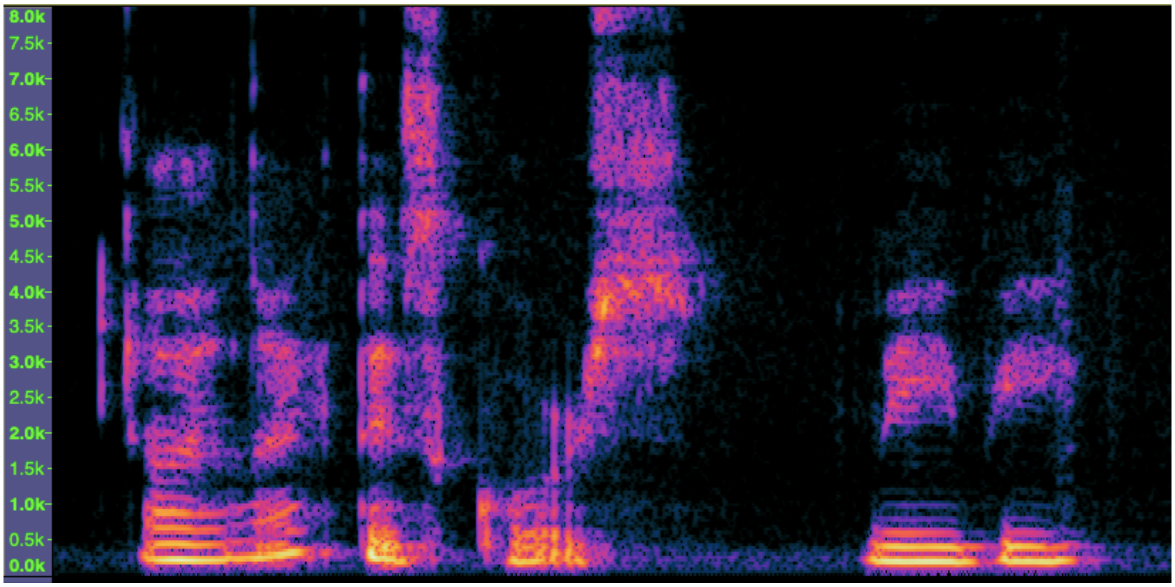

Figure 3: Spectrogram of the same random sentence “I like this class, kind of” as in Figure 2.

Looking at the spectrogram, one can observe that colors are changing in the image depending on the segment of audio signal. Each segment has a dominant frequency called formant or energy. The dominant frequencies are easily distinguishable by their color. The lighter the color, the dominant is the frequency, thus, one can observe the formants in the relatively bottom part of the image.

Inspecting the form of these formants over the frequency domain, and collecting them gives us the tone of the speaker. This information is very crucial for the technologies or methods to identify speakers and differentiate between them since tones existing in the audio signals uniquely represent the speaker.

Many algorithms work on frequency-based representations and transfer the frequencies into meaningful information. Examples for such algorithms can be Perceptual Linear Predictive (PLP) [42], Linear Prediction Coefficients (LPC) [43], Linear Predictive Coding Cepstrum (LPCC) [44], Real Cepstral Coefficients (RCC) [45], Mel Frequency Cepstral Coefficients (MFCC) [46], etc. The Table 1 lists 7 features by popularity and shows their representation, whether it is temporal of spectral. Algorithms working with sounds on spectrograms are Convolutional Neural Networks (CNN)[47] which take input as images of the sounds. This paper will discuss only the first type of methodologies, where the input is numbers in matrices, not pixels of images.

Table 1: Features and their representations. The information in the table is derived from [41].

Feature Name	Representation (Temporal, Spectral)
Short time energy or root mean square (RMS) or spectrum power or volume or loudness	Temporal
Zero crossing rate (ZCR)	Temporal
Mel-frequency cepstral coefficient (MFCC)	Spectral
Spectral centroid or brightness or frequency centroid	Spectral
Short time fundamental frequency or pitch or harmonic frequency	Spectral
LPC-derived cepstral coefficients (LPCC)	Spectral
Linear predictive coding (LPC)	Spectral

In general, the algorithms extracting a unique voice of a speaker creates segments of sound that consist of many dominant frequencies. These dominant frequencies are called formants and can be seen in Figure 3 by lighter colors in the bottom of the image. Together they form a tone of a voice and it can be saved in a digital format, similar to fingerprints or any other biometric data, which will serve a purpose of identifying a speaker.

2.2 Speaker Identification versus Speaker Verification

After we understand how voices are produced and how unique they can be, we can focus on how algorithms work to differentiate voices of speakers. To discuss technologies used to accomplish the task of speaker identification, one should understand its difference from speaker verification. Speaker identification is the process of identifying a speaker based on his/her voice. This task can be closed set or open set. A closed set speaker identification is identifying a speaker in the set of known speakers, and if any unknown speaker voice is given, the system will try to match it with the most similar speaker voice in the dataset. Such systems can be useful, if we are certainly sure that no other speakers or imposters would use the system. On the other hand, open set speaker identification takes into account the possibility that a given voice for identification may be out of the known speakers of the system.

While speaker identification answers the question "Who is speaking?", speaker verification answers the question "Is this person who he/she claims to be?". The input is not only voice, but also the claimed identity. The Figure 4 depicts the differences of these two fields of speaker recognition in a clear way.

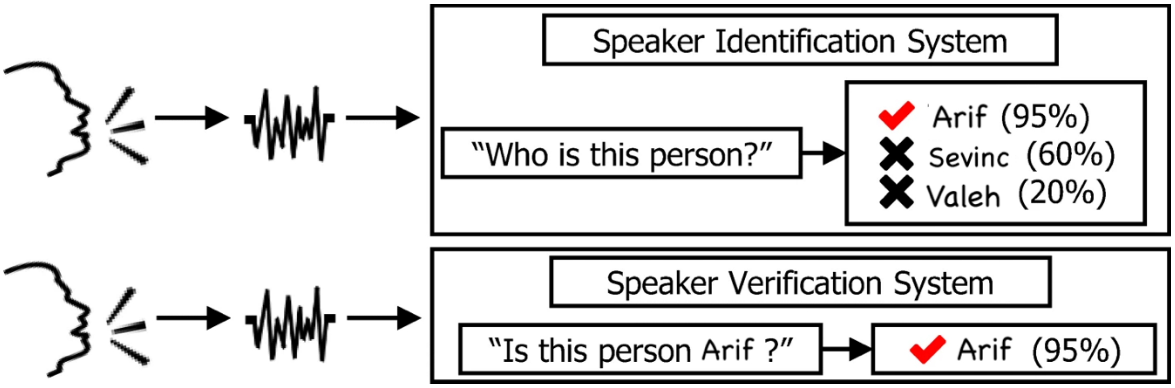

Figure 4: The main difference between speaker identification and speaker verification. This image is created and modified using the similar image from [48].

Figure 5 and Figure 6 depict the process flow of speaker verification and speaker identification, respectively. The speaker verification system takes the voice, and compares it with the speaker model of the claimed identity. Based on the similarity, the verification model gives a score to be inputted to the threshold. If the model of the claimed identity and a given voice match, the system accepts the speaker, if not, declines him/her based on the predefined threshold. On the other hand, speaker identification does not require the information about the claimed identity. The model tries to find the identity itself based on the maximum value of similarity scores with each speaker. It is necessary to mention that speaker identification can be open-set or closed-set problem. When the task is a closed-set, it means the model assumes that every given audio input will definitely belong to one of the speakers whom it already saw and learnt about. Therefore, the identification number of the found speaker is returned. However, if speaker identification is open-set, it is not certain that the voice belongs to the known speakers. In such case, additional check is required to implement, where the confidence is checked with a defined threshold. Another difference between these systems is that, speaker verification is a binary task, either yes or no, whereas, identification is a multiclass task. Furthermore, the increase in the number of speakers does not affect verification, but hugely affects identification since the system performs a search on known speakers to find the correct one. These two systems, both speaker identification and verification, together form a speaker recognition task. The algorithms and methodologies discussed in this section are for the use in speaker recognition, thus, applicable to both verification and identification tasks. The only difference in the approach is the end stage of evaluation where identification does a search based on an utterance and verification compares a speaker model with an utterance.

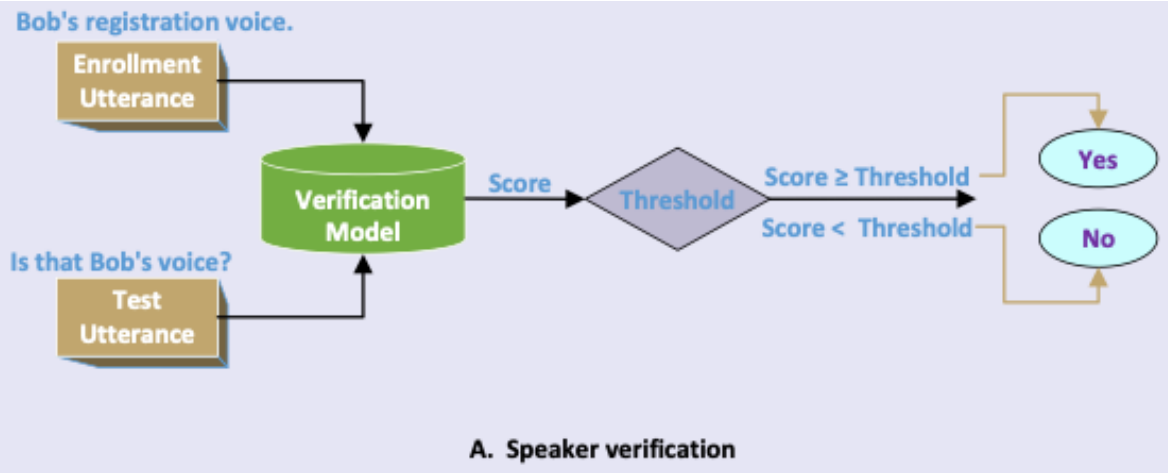

Figure 5: The flowchart of speaker verification. This image is acquired from [49].

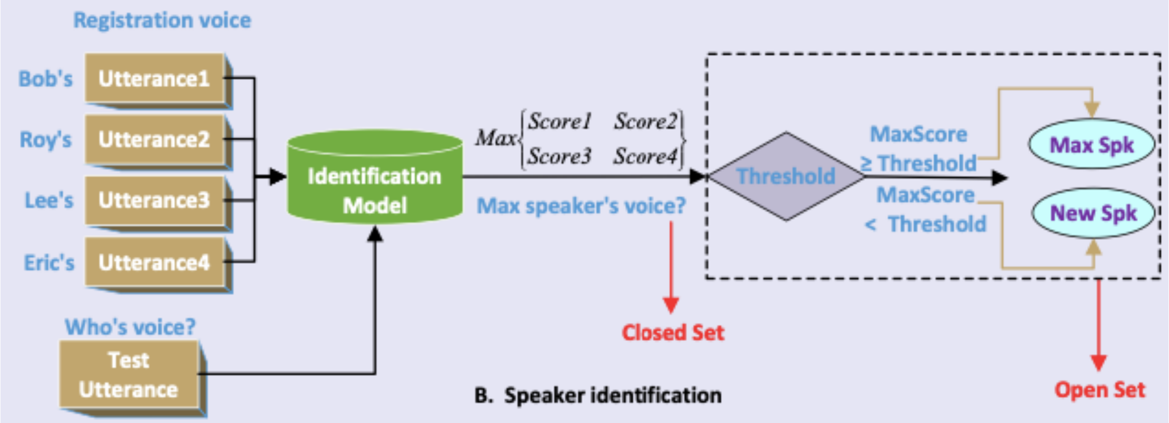

Figure 6: The flowchart of speaker identification. This image is acquired from [49].

2.3 Technology

As mentioned earlier, speaker recognition technologies started to develop as early as from 1960 [4]. At that time, algorithms were a simple template matching using frequency spectrum. Later in the 1970s and 1980s, features changed from frequency spectrum to LPC, LPCC, and at the end, MFCC. The algorithms were Dynamic Time Warping [50], which still matched sequences, and Hidden Markov Model (HMM) [51] which transferred the algorithms of speech and speaker recognition from template-based to statistical-based models. Starting from the 1990s, with the introduction of the Gaussian Mixture Models [52], the main speaker recognition approach has changed to use this flexible, robust and efficient method for that time. Later in 2000, Gaussian mixture model-Universal background model (GMM-UBM) [53] was developed by the same author, D. Reynolds, and acquired huge popularity since it enabled researchers to train practical applicable models in the real world. This method created one background model called Universal Background Model (UBM) that captured all information and voice patterns of all speakers and later adapted it to create Gaussian Mixture Model (GMM) for each speaker. Some adaptations were also developed which were GMM-HMM and GMM-SVM that combines the model with Hidden Markov Models and Support Vector Machine. The Figure 7 depicts the architecture of GMM-UBM for the task of speaker recognition. The pipelines for both speaker identification and verification can be easily observed.

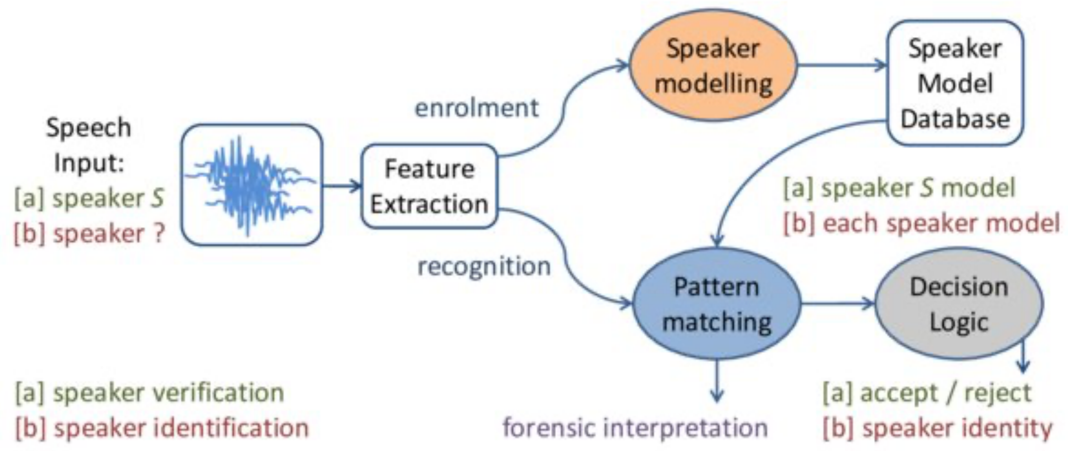

Figure 7: Architecture of GMM-UBM for speaker verification and identification. This picture is acquired from [54].

In the 21st century, two new approaches were created which made use of GMM-UBM methodology and its logic. These methods were Joint Factor Analysis (JFA) [55] and i-vectors [56], which represent each utterance with vectors. Additional to these main methods, back-end techniques have also been developed in the 21st century which include but are not limited to within-class covariance normalization (WCCN) [57] and probabilistic linear discriminant analysis (PLDA) [58]. These techniques contribute to discriminating speakers even more in evaluation phase of new utterances given to the system.

The i-vector approach, introduced in 2000, is very similar to JFA. JFA represents the GMM mean supervector by four components [59]. The approach also stores the variability in channels and speakers similar to GMM-UBM and uses all these information as variables in a mathematical formula of

$$M = m + Ux + Vy + Dz \quad (1)$$

where m is a GMM supervector which is independent of speaker and channel variances, U is a subspace of sessions in a matrix form and V is a subspace of speakers in matrix and D is a diagonal matrix that together form M, which is a GMM supervector with session and speaker information. The variables x, y and z are random vectors with a standard normal prior. The formula for i-vectors is similar, except all variabilities, be it channel or speaker, are combined into one variability matrix [59]. The formula for this approach is

$$M = m + Tw \; (2)$$

where we need to find w that represents i-vectors when M is GMM mean supervector of an utterance, m is UBM mean supervector and T is total variability matrix which holds speaker and channel variability. I-vectors were very popular for representing utterances and applying them in speaker recognition. However, the disadvantage of i-vectors and other previous generative models is that these architectures were utilizing unsupervised learning. Therefore, they intrinsically aimed to generate a distribution of acoustic signals, not differentiate speakers. Ten years later in 2010, as neural networks took a rise among the machine learning field, a new method was developed similar to i-vectors. The idea stayed the same because representing utterances as vectors makes it easier to differentiate and find similarities between recordings with simple methods. The difference was in the approach. While i-vectors have a clear mathematical formula for finding vectors, this new method took second to the last layer of neural networks and used it to represent utterances. These vectors were named as x-vectors and they performed better than i-vectors [3]. D. Synder and other authors of this new system have tested i-vectors and x-vectors with publicly available datasets, which are SITW Core and SRE16 Cantonese in [3]. The results show that x-vectors were at least 1% and at most 4% better than i-vectors in the task of speaker recognition. This study has been conducted using the latest technology in speaker recognition which is x-vectors.

Below is a figure that represents the summary of speaker recognition history (Figure 8). It depicts how feature extraction and technologies have evolved starting from 1930 to our days. It also shows when practical use of speaker recognition technologies have changed from small scale to large scale.

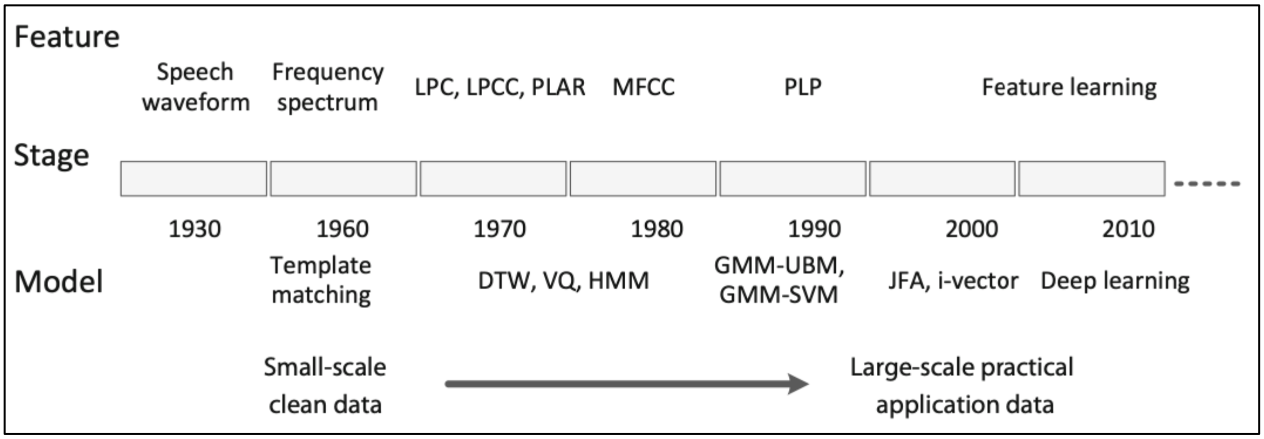

Figure 8: History of speaker recognition technologies and features. This picture is acquired from [4].

After the creation of new models such as i-vectors and x-vectors, speaker recognition architecture took a clear form (Figure 9). The architecture of speaker recognition with the latest technologies consists of 3 distinctive parts: the first part is training where lots of recordings of different speakers are given to the model to extract features. Each recording gets a preprocessing where Voice Activation Detection (VAD) [60] is applied to get exact frames of voice and disregard silent frames. The discussed embedding vectors have different way of acquisition in this phase. While i-vectors make use of universal background model and adaptation of it with speaker and channel variations, x-vectors make use of artificial neural networks and represent the audio signals with a layer in neural networks. This difference is depicted in Figure 10.

As mentioned above, the extracted features are vectors that each represent a different utterance and are unique in nature. These vectors are fed into the second phase, which is named as enrollment phase, where speaker models are acquired from them. The features of a speaker can be averaged to get a model of that speaker. So far described two phases can be observed in Figure 11.

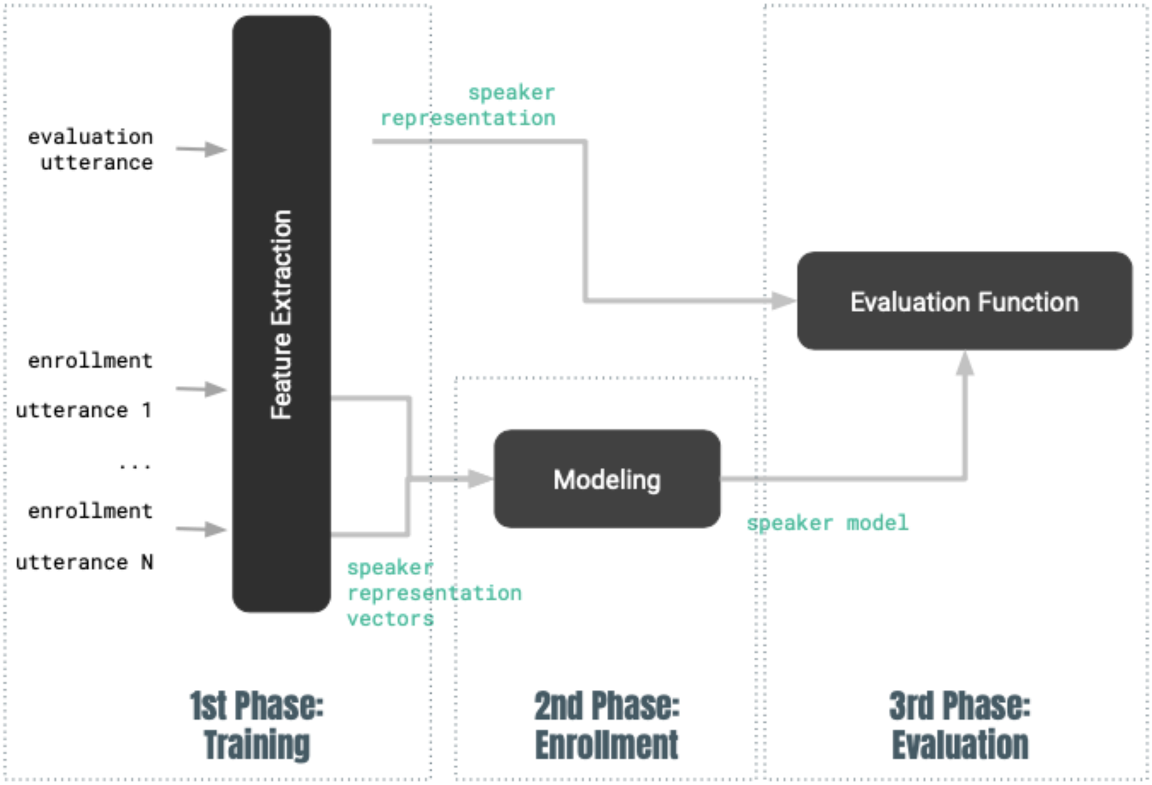
speaker representation
evaluation utterance
Feature Extraction
Evaluation Function
enrollment utterance 1
...
enrollment utterance N
Modeling
speaker model
speaker representation vectors
1st Phase: Training
2nd Phase: Enrollment
3rd Phase: Evaluation

Figure 9: The architecture of a speaker recognition system.

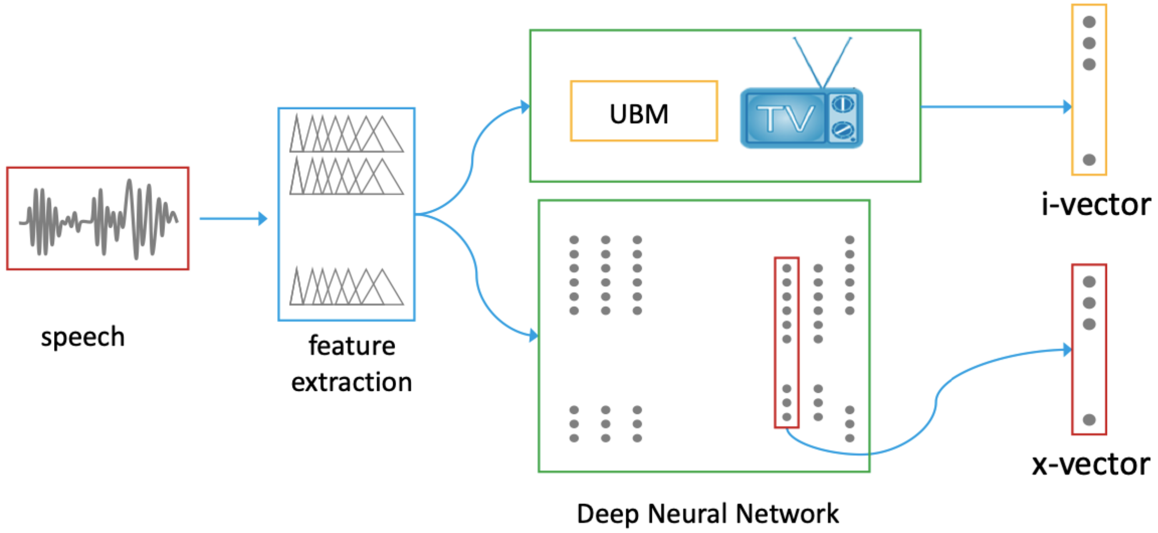

Figure 10: Comparison of pipelines of speaker recognition with i-vector and x-vector architectures. This image is acquired from [61].

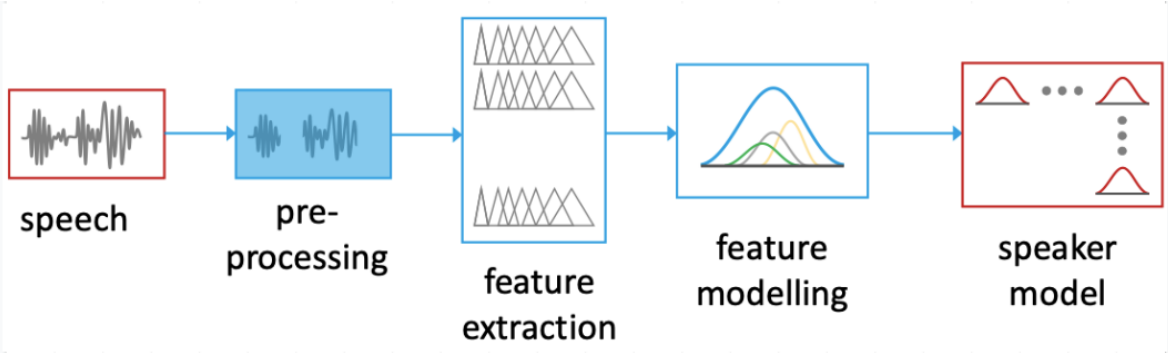

Figure 11: Pipeline of speaker recognition. This image is acquired from [61].

The third phase is evaluation, and it differs for speaker identification and verification. In speaker identification, the evaluation phase is calculating probabilities of a recording belonging to speakers. An incoming recording passes through feature extraction and then its features are compared to those of other speakers. At the end, we get probabilities of a recording belonging to speakers in our known set.

After an overall architecture is explained, more detailed and technical features of the architecture used in this study can be discussed. The first part of the architecture - feature extraction makes use of Time Delay Neural Network (TDNN) [62]. This network structure takes into account the current frame and frames before and also after the current frame. In general, one can say that TDNN is taking context into account when learning. A general idea and workflow of this network is depicted in Figure 12. The output of this process is speaker embeddings which are called x-vectors. X-vectors, as discussed earlier, are an efficient and compact way of representing audio recordings and are, in fact, a layer within a network.

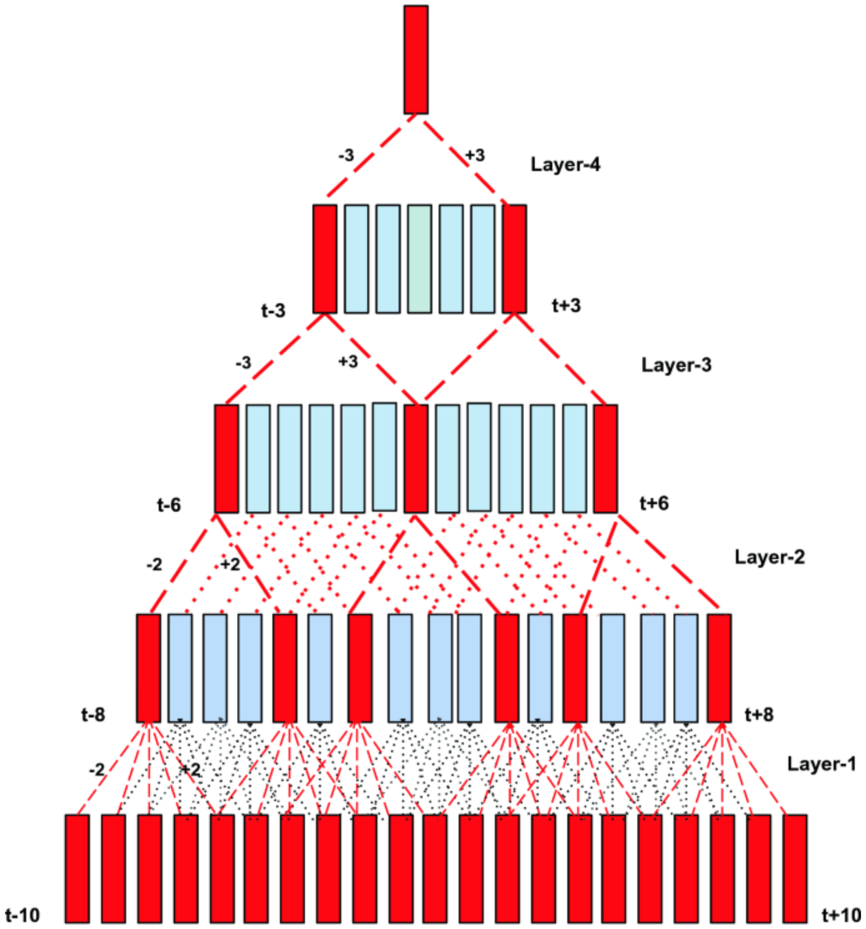

Figure 12: Time Delay Neural Network. This image is acquired from [63].

Figure 13 describes this process in more detail. As generally shown in the image, the overall architecture consists of 5 hidden layers using rectified linear unit (ReLU) [64] activation and batch normalization, a statistics pooling layer, 2 hidden layers to reduce dimension and a softmax layer. Speech feature frames are inputs to the first frame-level layers. These layers capture the temporal information taking into account adjacent frames and, by that, the context. The pooling layer, which is coming next, aggregates the information outputted from the last frame-level layer. It calculates mean and standard deviation across frames of a utterance and converts information in the process-flow from frame-level to recording-level. This is why the next two layers after pooling are named as recording-level layers, and they, in fact, represent utterances. Even though these layers can both represent the utterances, the first and the second layers were compared conducting experiments and it was found by an empirical way that the first layer gives more accurate representation [3]. Therefore, the first layer after the pooling one is where we get our embeddings. The last layer is a softmax layer that gives each class a probability of uttering the given recording input. Based on the result and the ground truth, backpropagation updates weights and the model to be more accurate.

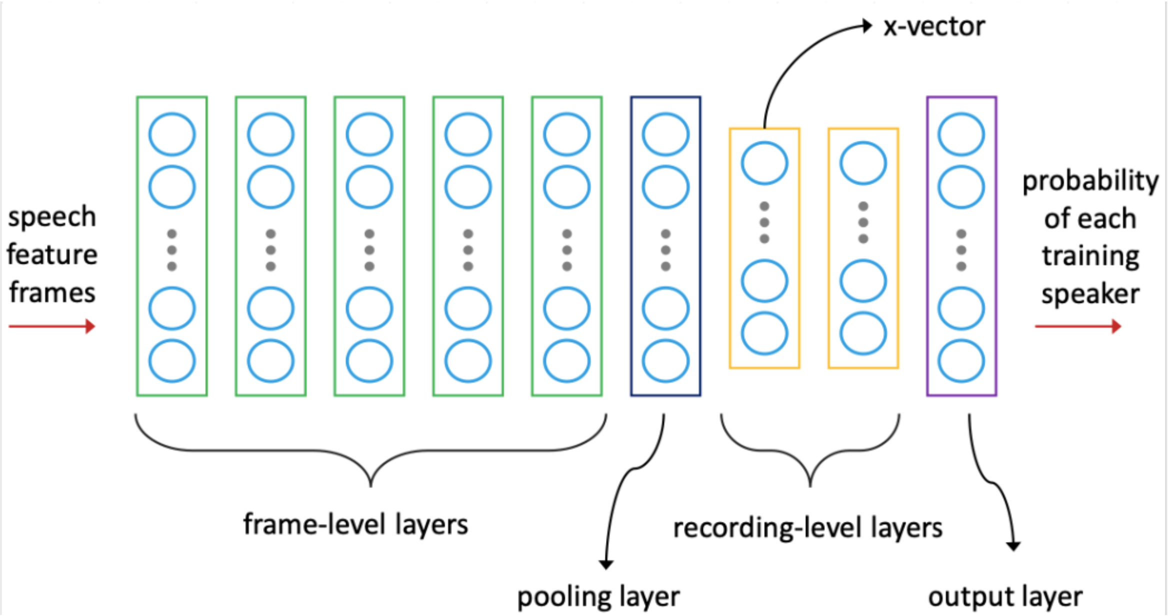

Figure 13: TDNN architecture of speaker recognition. This picture is acquired from [61].

Table 2 gives additional information about the layers within TDNN. For each nine layers, layer context, the number of total context and the size of input and output of layers are given. With the given input having T frames in total, t is our current time frame. The first 5 layers, which we named earlier as frame-level layers, add more context and by that, increase the size of the total context. Hence, starting with 5, the last layer has in total 15 time frames. The statistics pooling layer aggregates the output from frame 5 across T and finds the mean and standard deviation over T and stores it in the vector of size 1500. Since the statistics layer aggregates the outputs, consequent layers work on context {0} and the size of T which means an entire recording. The next layer is our embedding layer x-vectors and till this point and including it, there are 4.2 million parameters.

Table 2: Architecture of TDNN layer by layer.

Layer	Layer Context	Total Context	Input x Output
Frame 1	{t-2, t+2}	5	120x512
Frame 2	{t-2, t, t+2}	9	1536x512
Frame 3	{t-3, t, t+3}	15	1536x512
Frame 4	{t}	15	512x512
Frame 5	{t}	15	512x1500
Statistical Pooling	[0, T)	T	1500Tx3000
Segment 6	{0}	T	3000x512
Segment 7	{0}	T	512x512
Softmax	{0}	T	512xN

The architecture of TDNN and x-vectors for speaker identification can be found in the toolkit named SpeechBrain [65] and it is also used in this study for conducting experiments. SpeechBrain is a public, accessible, easy-to-use tool that is aimed to deliver speech functionalities in a holistic way as our brain does [66]. The configuration of training is an easy process where we need only modify the human-readable train.yaml file to obtain the training configuration we want. Moreover, the availability of clear documentation with examples makes it easier to start working with the toolkit, understand the code and modify, if necessary.

The dimension of x-vectors used in the experiments conducted in this study is 512. The features are filterbank features of input signals derived from 23 Mel filters which are used to average the spectrogram banks. The activation function used in the training of neural networks is different from the original architecture and is LeakyReLU - Leaky Rectified Linear Unit that adds a small slope for negative values compared to ReLU that has a flat slope [67]. The optimizer is Adam optimizer instead of classical stochastic gradient descent as it handles sparse gradients on noisy problems. There are 5 TDNN blocks and their channel sizes vary as 512, 512, 512, 512, 1500. The last value is changed during experiments for better fitting.

3 RESEARCH RESULTS AND ANALYSIS OF RESULTS

This section introduces the different datasets used in this study and describes experiments conducted with these datasets at the same time analyzing the results.

3.1 Datasets

3.1.1 English AudioMNIST Dataset.

The first dataset used in this research is the English dataset named AudioMNIST digits [68]. It is an audio version of a popular MNIST (Modified National Institute of Standards and Technology) database which consists of images of handwritten digits. The AudioMNIST dataset has 60 speakers and each of the speakers has uttered digits from 0 to 9 fifty times, thus, the total number of recordings are 3000. The audio recordings are at 48 kHz. Its total duration is 5 hours and 21 minutes.

Different subsets of the dataset have been used in the experiments. In addition to using all of the data, various subsets have been generated and used for training and testing in the experiments. The training datasets are usually 5 digits concatenating together various number of times. These training subsets are the following: only 5 digits of the dataset from 0 to 4 uttered 50 times, and from 0 to 4 uttered only 5 times. Test datasets are a subset of the all data and also a subset of the dataset from 3 to 7 uttered 50 times for the first training set. For the second and third training sets, test sets are from 0 to 4 and 5 to 9 uttered 50 and 5 times, respectively. Lastly, the third training tests an additional subset from 3 to 6 uttered 5 times. The sample rate of these recordings is 48 kHz. The ratio of female to male speakers is 20:80.

3.1.2 Azerbaijani AudioMNIST Dataset.

The Azerbaijani dataset is similar to the previous English dataset. There, 59 speakers have uttered digits from 0 to 9, but various times. Some speakers uttered the digits 10 times, some less and some more. On average, each speaker uttered digits 5 times. This variety complicates the problem. Nevertheless, the same test and training datasets have also been created for the Azerbaijani dataset. The sample rate of these recordings is 16 kHz. The ratio of female to male speakers is 32:68. Its total duration is 1 hour and 17 minutes. This dataset was collected during the research with the help of ADA University students and is shared to the web for the use of the public [69].

The division of this dataset to train and test sets is similar to the English dataset. The training datasets are concatenation of 5 digits from 0 to 4. The first training set concatenates all repetitions of digits, which makes the number of samples various for each speaker since each of them uttered digits different number of times. The second training set concatenates only 4 repetitions as it is the largest number of samples each speaker has in common. The testing sets are also similar to the previous dataset. The concatenation of all digits and digits from 3 to 7 were created for the evaluation of the first training, and concatenation of digits from 0 to 4, 5 to 9 and 3 to 6 repeated 4 times are created for the second training.

Additional to the sequential order of digits, the study also experimented with subsets of the digits which were phonetically divided so that both train and test sets will have the same vowels within digits. The more detailed explanation on how the digits are grouped can be found in the "Experiments" section and "Experiments on the AudioMNIST Datasets" subsection.

3.1.3 Azerbaijani DAC Dataset.

This dataset not only has digits from 1 to 9 like previous ones, but also has commands that speakers uttered, hence, named as Digits And Commands (DAC) dataset. The dataset was also collected during the research with the help of ADA students like the previous Azerbaijani dataset, however, it is not publicly shared and is accessible only for ADA students. The commands in the DAC dataset are popular voice commands which are usually asked from mobile phones, like "What is the weather today?", "Call my mom", "Open contacts" and etc. There are in total 29 speakers and 86 texts, consisting of 9 digits and 77 commands. The sample rate of these recordings is 48 kHz. The ratio of female to male is 38:62. In total, this dataset is 17 hours and 42 minutes.

An updated dataset is also extracted from the current dataset, named DACW. The DACW dataset consists of chosen 20 speakers from the speakers set who uttered all commands and digits. Their female to male ratio is 45:55 which is a more balanced set of speakers out of all data sets. This dataset has 9 digits and 71 commands uttered. Some commands were removed which had less than 3 times repetitions making 72 commands out of 77 and one more command which had 5 words in it was not used, making in total 71. Each utterance is on average 10.3 seconds and has on average 293.5 files. Each speaker uttered each text about 14.7 times. The least amount of average repetition made by a speaker is 6.8 times and the most is 27.6 times. In total, it is 15 hours.

The DACW dataset was created to test the relationship of speaker identification to the duration of recordings. Therefore, DACW has 3 parts in it. The first part of the dataset - DACW-2 has only two word commands in it. There are 48 commands that have two words and 45 of them are used. The second part, DACW-3 has 3 word commands. In the original dataset, there are 27 commands and 25 of them are used. The additional 20 commands were generated using two word commands and one word digits making in total 45 utterances. The last part of the dataset is called DACW-4 consists of 4 word commands. The number of original utterances is 1, thus, 44 new commands were generated. These newly generated utterances are either combinations of original three-word commands and one-word digits, or combinations of two two-word commands. At the end, all datasets - DACW-2, DACW-3 and DACW-4 have 20 speakers and 45 utterances. The sample rates are kept as in the original dataset. Because each recording in DACW-4 is longer than DACW-3, and each recording is DACW-3 is longer than DACW-2, the total duration of these datasets differ by around an hour so that DACW-2 is 2 and half hours with 4404 files, DACW-3 is 3 and half hours with 4171 files and DACW-4 is 5 hours with 4112 files. The number of files are similar. The content of these datasets consists of the same commands and combinations of them, but no two utterances are identical in these sets. Repetitions of each command were exclusively divided into three parts to be used in generating these datasets.

3.1.4 RIRS Dataset.

RIRS dataset is an augmentation dataset used to enrich the datasets for training [70]. It is named as Room Impulse Response and Noise Database, and as the name suggests it is a database of real and simulated room impulse responses and noises. Room impulse response is a transfer function between the original sound source and the microphone. The noises in this dataset have two types: isotropic and point-source. Isotropic noise is a kind of sound that emits equal power in all directions [71] and point-source noise emits spherical spreading in all directions [72]. The RIRS dataset combines real RIRs from many available datasets and generated ones by the authors. The point-source noises are taken from a very popular database named MUSAN [73]. There are 843 point-source noise files and 325 real RIRs files. The simulated RIRs are divided into three sets which are small, medium and large rooms. Each room set has 200 folders with room numbers on it and each such room folder has 100 simulated RIR sounds. In total, small, medium and large rooms together have 60 thousand files. All the files in the RIRS dataset are in 16kHz.

3.2 Experiments

The dataset in all experiments have been splitted into 80% train, 10% test and 10% validation. Train set is used to train our model and learn how to differentiate speakers. With the validation set, we tune the hyperparameters of the model to learn better by observing how better it becomes on identifying the validation set. After training is done, accuracy is calculated using the test set and observing how our trained model performs on unseen data of learnt speakers.

3.2.1 Experiments on the AudioMNIST Datasets.

The first set of experiments have been conducted on the English AUDIO MNIST dataset. 10 digits uttered by 60 speakers were merged together and grouped by the number of repetition that is uttered. Hence, if a speaker utters each digit 50 times, making 500 files, in total, concatenating these digits together results in 50 longer files. On average, each speaker spoke 5.24 minutes. With 80% of this dataset, the training gave very high results. When tested with a test set that is unseen to the model, the model gave 99.4% accuracy. Testing the model performance on all dataset, including train set and also validation, gave 99.8%. When tested with a subset of these known digits consisting of 3 to 7, the accuracy stays the same, which means subsets of the train set are also recognizable.

Further experiments are conducted to check the performance on a subset that is out of the train set. For that, a new subset is created consisting of digits from 0 to 4. This set has an average of 4.47 minutes for each speaker. The experiment also resulted in high accuracy with 99.9% accuracy. Another subset of digits consisting of a sequence from 5 to 9 is used to test this model and understand if the model has any relationship with the spoken text. It turns out that the relationship exists since the model scored 34.1% for unseen digits.

The experiment was conducted again but with less number of repetitions, particularly 5. With a 90% decrease of files, the average time per speaker decreased to 27 seconds, on average. The test accuracy is 96.7% for digits 0-4. When tested with an unknown set of digits - from 5 to 9, the accuracy dropped 15.3%. This could indicate that some sense of text dependency exists. Another set of digits - from 3 to 6 is also created to test the model. This set consists of two numbers, 3 and 4, from known, and the other two numbers, 5 and 6, from unknown digits sets. The accuracy of the test set was 29.3%, which is an almost twice better result than the previous set where all digits were unknown. The experiments are depicted in the table below (Table 3).

Table 3: Results of experiments on the English AudioMNIST dataset.

Combination	The number of speakers	Utterances	Average Time	Accuracy over all dataset (%)	**Accuracy over test dataset (%)**
60x10x50	60	0-9	5.24 minutes	99.79	99.4
	60	3-7			99.36
60x5x50	60	0-4	4.47 minutes	99.97	99.90
	60	5-9			34.11
60x5x5	60	0-4	27 seconds	97	96.67
	60	5-9			15.33
	60	3-6			29.33

The experiments conducted on the english digits dataset were repeated with the Azerbaijani dataset. The difference and also a challenge of this dataset is that the dataset in the Azerbaijani language varies in terms of the number of repetitions each speaker made while uttering a digit. Therefore, there is a question mark in the first cell of Table 4 because the number of repetitions is unique for each speaker. In addition to the repetitions being inconsistent and less than the english dataset, the cumulative average time is also less. It is 2 minutes for each speaker, twice less than the english dataset. Using 80% of all utterances for training, we acquire 85.3% for the test set. It is 14% less than the results obtained with the english dataset, but this was expected due to the factors mentioned above. The subset of the trained digits also resulted in the same accuracy.

In the next experiments the repetition number is fixed to 4. Excluding the speakers who uttered digits less than 4 times, the resulting set has 36 speakers. The average time of utterances is 48 seconds. The average duration being more than the english dataset is due to the reason that digits in the Azerbaijani dataset are sometimes uttered longer. This set gave 86.7% accuracy on identification of speakers. Similar to the previous experiments, when tested with an unknown set of digits(from 5 to 9) the accuracy dropped to 53.3%. Combining two known and two unknown digits to the testing set, the accuracy increased 13%, resulting in 66.7%. The results are similar to the ones with the english dataset. The information about these experiments can be observed in Table 4.

Table 4: Results of experiments on the Azerbaijani AudioMNIST dataset.

Combination	The number of speakers	Utterances	Average Time	Accuracy over all dataset (%)	Accuracy over test dataset (%)
59x10x?	59	0-9	2.02 minutes	96.71	85.3
	59	3-7			85.29
36x5x4	36	0-4	48 seconds	97.92	86.67
	36	5-9			53.33
	36	3-6			66.67

Considering the first two experiments in Table 3, we can see that the model identifies the subset of the trained digits. Even though the training set is from 0 to 9, any subset (from 3 to 7 in this case) can be used to identify speakers. The next 2 experiments show that when the training set is from 0 to 4, the test set of 5-9 gives really low accuracy. The next experiment was conducted to see the accuracy rate when speakers will only utter the digits 5 times, which gives pretty decent results. Again, when trained with 0-4, 5-6 gives low accuracy when tested. Combining 2 digits from the train dataset and 2 out of train dataset shows higher accuracy than the previous test, which concludes that the model depends on the digits and not the voice of the speaker.

In Table 4, we can see similar experiments that conclude the same idea. Because the number of utterances change for speakers, we have collected those speakers that have uttered each digit at least 4 times. This gives 36 speakers for the last 3 experiments. However, the test results are low due to the smaller dataset which brings overfitting. This gap between training and test has been mitigated from test being 70% to 80%, but still the issue exists. When more data will be collected as a future work, this problem will be resolved.

Another issue with the experiments is that only one way of text-dependency is tested which is training digits from 0 to 4 and testing with other digits. It could be the case that the model trained with sounds of letters in digits from 0 to 4, never sees and understands sounds existing from 5 to 9. In that case the observed phenomenon would be not a text-dependent characteristic of the architecture but adapting to and learning the training set. Therefore, new experiments have been conducted where digits are sorted in both ways: sequential, and also phonetical. The sequential order of digits is defined as from 0 to 4 and from 5 to 9. Both used as training sets and test sets. In order to create phonetically correct datasets, the sounds forming the digits must be observed in detail. There are in total 9 vowels in the Azerbaijani language and all of them are used in digits. Table 5 shows which vowels exist in which digits, and the following sounds in brackets approximate these sounds to English ones.

Table 5: Digits in Azerbaijani and the corresponding vowels sounding when uttering them. In brackets, it is written how the vowel approximately sounds in English.

Digit	Vowels
0: Sıfır	ı (eu)
1: Bir	i (ee)
2: İki	i (ee)
3: Üç	ü (ew)
4: Dörd	ö (eo)
5: Beş	e (e)
6: Altı	a (ah), ı (eu)
7: Yeddi	e (e), i (ee)
8: Səkkiz	ə (a), i (ee)
9: Doqquz	o (oh), u (oo)

Using this information, the train and test datasets were divided into groups such that each vowel would appear both in train and test to eliminate the hypothesis that the model reacts to sounds existing in the training set. There are 3 groups formed:

1. a = {5,6,8,9}. These digits hold the unique vowels to be used in the training set. 2 vowels (u) and (o) from the digit 9, and 1 vowel (ə) from the digit 8 do not exist in any other digit.
2. b = {0,1,2,7}. The vowels in these digits are repetitions from the training set and will be used for testing. In this manner, the vowels seen in the training set will also be seen in the test set, so text-dependency will be texted again in a more challenging way.

3. c = {3,4}. These digits also hold unique vowels that do not exist in other digits. The experiments will sometimes include one for training, and another one for test, or all for training or all for test.

The experiment settings with these new digit groups will be as in Table 6.

Table 6: Phonetic experimental setting of Azerbaijani AudioMNIST dataset with digit groupings.

Train	Test
c1+a	c2+b
c2+a	c1+b
c1+c2+a	b
a	c1+c2+b

The experiments to check text-dependency of a model both in sequential and phonetic way were conducted using less number of speakers which is 33 and using all repetitions a speaker uttered for digits. Since the number of repetitions varies per digit, when concatenating digits in the group equal times there were utterances left. These utterances are also concatenated to increase data samples. It is clear that the accuracy numbers will be very low if we decrease the number of speakers and keep the number of repetitions equally at 4 as it is the largest number of repetitions common for speakers per digit. The architecture of x-vectors needs a lot of data to show good results, therefore, when decreasing the number of speakers, the amount of repetitions should increase so that the model will have enough data samples to learn and perform more than 60%. The results can be seen below in Table 7.

Table 7: Text-dependency experiments with sequential and phonetic sorting of digits.

Train set	Test Set	Accuracy (%)	Test set #2	Accuracy #2 (%)
0-4	0-4	81.82	5-9	59.09
5-9	5-9	82.61	0-4	56.52
3-5-9	3-5-9	69.57	0-4-7	52.17
4-5-9	4-5-9	60.87	0-3-7	56.52
3-4-9	3-4-9	73.91	0-2-7	56.52
5-6-9	5-6-9	82.61	3-4-7	43.48

In the table, the first two rows are sequential testing and the last four rows are phonetic testing, in the order given previously in Table 6. Observing the differences in the accuracy columns for each experiment, one can say that the model can differentiate between texts whether the digits are in sequential order or in phonetic grouping. Overall, the models trained with digits in sequential order performed better than the phonetic setting defined above. This could be due to the grouping or the dataset, in general. Even though the accuracies change for a model in each experiment, identifying speakers with the set of unknown words seems to be staying the same around 50%. This can conclude that x-vectors indeed have information about the given texts, and perform poorly on unknown words. Nevertheless, this dataset is very small to derive such conclusions and some experiments resulted in overfitting. The next experiments are conducted with a bigger dataset.

3.2.2 Experiments on the DAC Datasets.

The experiments in the previous section introduced us a little bit to the identification and formed an idea about its relationship with text. However, the toy datasets above are not enough to draw a conclusion. With such small datasets, it is easier for a model to overfit and perform unexpectedly on a different set. Hence, a bigger dataset is used to test the hypothesis and conduct more experiments to also research the existence of a relationship between speaker identification and duration of recordings.

The bigger dataset used for the purpose of this study is the DAC dataset. The DAC dataset is divided into two parts for text dependency experiments. The first part consists of only digit utterances, and the second part is only command utterances. In the digit dataset each digit utterance on average is 1.74, and its total duration is almost 1 hour and half. The commands are on average 2.35 seconds long, so the duration of these two sets does not differ much. On the contrary, the commands are a lot more than digits, meaning there are only 9 digits, however, the commands are 77. Therefore, even though the duration of each utterance of these sets do not differ, at the end, the duration of the command dataset is almost 16 hours.

The results are shown in Table 8. The accuracy of the model trained on the digits dataset is 85.58% and the model trained on commands gives 92.16%. This is predictable since commands have more data than digits. The cross-testing of these datasets gave similar results with each other. Digits model scored 62% on the commands dataset, and commands model scored 58% on the digits dataset. Even though we have switched the dataset, the same type of experiments gave again around 50%-60% accuracy. Low accuracies show that the models have text dependencies since the drop in accuracy is huge. The text-dependency experiments with both Azerbaijani AudioMNIST and DAC datasets performed the same and revealed the relationship between the architecture and text used in the training.

Table 8: Accuracy results of experiments with DAC digits versus commands.

	DAC Digits	DAC Commands
DAC Digits	85.58%	62.42%
DAC Commands	57.74%	92.16%

The next experiment has been conducted on the subset of the DAC dataset which has all the commands and digits, but less number of speakers. 20 speakers out of 29 were chosen since these speakers have uttered all 86 texts (77 commands and 9 digits), thus can be grouped together. As mentioned earlier, the speakers uttered commands and digits various times. It on average ranges from at least 6.8 times to the most is 27.6 times. An average of all speaker average values gives the value of 14.7. This means that if we pick any utterance of any speaker, the number of times it was repeated by that speaker will be 14.7. Without limiting the repetition times and giving all the dataset to train, we get an accuracy of 97.95%. If we are to limit it somehow, so the range of repetitions will not be drastically huge, we can set the limit as at most 10 repetitions. Most of the speakers, to be precise 75% of the speakers, have uttered a text more than 10 times, on average. The other 25% on average repeated below that number. The logic behind this creation of this subset is, if there are more than 10 repetitions for an utterance for any speaker, only 10 are kept, if it is low, all are kept. With this setting, the accuracy decreased to 87.03. These experiments form an idea that if a system needs to be built for text-dependent closed-set speaker identification in the Azerbaijani language, each text should be repeated more than 10 times for an accuracy above 90%. The results are depicted in Table 9.

Table 9: Accuracy results of experiments with DACW various repetition times.

Repetition Times	Test Accuracy (%)
All repetitions	97.95
At most 10 repetitions	87.03

Another experience has been conducted to test the relationship of models with the duration of recordings. For these three datasets have been used which are DACW-2, DACW-3 and DACW-4. Three models have been trained using these datasets and cross-tested with them. The results are shown in Table 10.

We can observe the performance of each model closely. For DACW-2, the most accurate result is with its own dataset, which can be predicted. When tested on DACW-3 and DACW-4, the result was 4% lower. For DACW-3, the most accurate result is with DACW-4 and DACW-3, and it is lower for DACW-2 by 3%. DACW-4 has the highest accuracy for its own dataset, then it is lower by 4% for DACW-3, and by 5% for DACW-2.

Table 10: Accuracy results of experiments with DACW-2, DACW-3 and DACW-4.

	DACW-2	DACW-3	DACW-4
DACW-2	97.96%	93.07%	93.26%
DACW-3	96.19%	99.04%	99.47%
DACW-4	95.28%	96.24%	99.76%

By viewing the table again, we can observe a pattern in diagonals, where each model scored relatively good in identifying its own dataset. In addition, we can conclude that if we are trying to achieve higher accuracy for a dataset, the longer the recordings are the better. The highest accuracy has been shown by DACW-4 with 99.76% accuracy. For cross-tests, however, DACW-3 seems to do better than DACW-4. It is 1% more accurate in testing DACW-2 than DACW-4. Also, DACW-3 performed better on testing DACW-4 than vice-versa. This can conclude that if we are going to test our model with different combinations of the same utterances, the best dataset to train our model will be the middle or average number of words of our datasets. For example, if we know that our model will be trained for two-word commands, as well as, three and four-word commands, we would train our model with a three-word dataset since it is near to other two datasets. With the pattern exposed in the tests, we can clearly see that the more the dataset goes in one direction of the number of commands, the farthest one loses accuracy. As an example, DACW-4 performs relatively poorly on DACW-2, but better for DACW-3. Overall, models trained with the longer durations perform better in identifying speakers.

4 SUMMARY AND CONCLUSIONS

In summary, speaker identification is a technology that has been researched for a long time due to its useful applications, in investigations and customer service. This field is evolving starting from 1960, and the latest architecture was introduced in 2018 which utilizes the idea of vectors and embeddings. The embedding method called x-vectors is derived from the deep neural network and claimed to be very accurate. This technology was used in experiments in this paper. The aim of this study was to find out if the architecture of x-vectors has information about the text given in the dataset. Since most research tested the architecture in a text-independent setting, creating a text-dependent model is a new approach in which x-vectors can be used. Thus, its relationship with text, how accurately it performs when confronted with unknown text, is to be tested. Additionally, the relationship between x-vectors and duration of recordings given as input to the neural network is observed.

For these purposes, three datasets were used to train the models. These are digits datasets such as AudioMNIST in English and Azerbaijani, and digits and commands (DAC) dataset in Azerbaijani. The latter one is the largest dataset among the three, being in total almost 18 hours. All of these datasets were used to test text-dependency, whereas, the DAC dataset was used to test duration dependency. Moreover, there is a scarcity in studies of speaker identification in the languages belonging to a Turkic group, and even less research papers on this task with the Azerbaijani language. Therefore, this research will hopefully benefit the local community and contribute to new studies.

The experiments in this study showed quite interesting results. First of all, all experiments with data sets regarding text-dependency showed that it exists. When training a model to identify speakers with a defined set of digits and testing with digits unseen to model be it in English or in Azerbaijani AudioMNIST datasets, the model performed very poorly. Even though speakers were identified with an accuracy higher than 80% with digits defined in the training set, with unseen digits it is around 50%. Conducting the experiment with the larger DAC dataset gave the same results. When the model is trained with commands, with unseen and the same text commands, it gave 92.16% accuracy, however, with unseen and different text digits it resulted in 57.74%. The same pattern was observed when conducting the experiment vise-versa. When the model is trained with digits, with unseen and the same text digits, the result was 85.58% accuracy, however, with unseen and different text commands the outcome was 62.42%. In summary, x-vectors architecture showed a text-dependency characteristic in experiments with the English and Azerbaijani dataset varying from 1 hour to 18 hours, even though the texts of the inputted data were not defined.

For testing duration-dependence of x-vectors, commands of various lengths, such as 2, 3 and 4, were utilized, which are generated subsets of the DAC dataset. The outcome of the experiment shows that the longer the duration, the better the model learns and identifies a speaker. The subset with longer duration showed better results also in cross-testing, where the model trained with the utterances consisting of 4 words tested with utterances consisting of 2 or 3 words.

An additional experiment with DAC dataset concluded that for achieving an accuracy in text-dependent speaker identification higher than 95%, each text should be repeated at least 10 times and more than that for higher results.

To conclude, x-vectors architecture for speaker identification is a text-dependent architecture which shows better results with the known texts and poor results with unseen text when recognizing speakers. The difference in accuracy values between test cases of known and unknown texts can be from 20% to 40%. It also performs slightly better with longer utterances than with shorter ones. Furthermore, since deep learning requires a lot more data than any other architecture, this study defined that the number of repetitions for texts in text-dependent speaker identification using x-vectors should be 10 and more to achieve accuracies higher than 95%.

The future work of this study will be experimenting with various length of utterances starting from 1 command to 6 commands to further ensure the observed pattern. Moreover, the architectures of x-vectors and i-vectors can both be tested on a small dataset and a larger dataset with a potential gradual increase in sizes, to see the performance and determine which architecture performs better on which size of data. The next studies will also be focused on open-set speaker identification, where unknown speakers who are not in the database of the training set will be correctly identified as unknown.

REFERENCES

[1] Erin Digitale. 2016. Mom's voice activates many different regions in children's brains. Stanford Medicione.

[2] K.von Kriegstein ,D.R. Smith , R.D. Patterson, S.J. Kiebel, T.D. Griffiths. 2010. How the Human Brain Recognizes Speech in the Context of Changing Speakers. Journal of Neuroscience. 10.1523/jneurosci.2742-09.2010

[3] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, Sanjeev Khudanpur. 2018. X-vectors: Robust DNN Embeddings for Speaker Recognition. ICASSP 2018 - 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 10.1109/ICASSP.2018.8461375

[4] Thomas Fang Zheng, Lantian Li. 2017. Speaker-related robustness issues. SpringerBriefs in Electrical and Computer Engineering. 10.1007/978-981-10-3238-7_3

[5] Y. Imamverdiyev and L. Suxostat. AZ-SRDAT – РЕЧЕВАЯ БАЗА ДАННЫХ ДЛЯ АЗЕРБАЙДЖАНСКОГО ЯЗЫКА. Challenges of Information Technology, 2013, 1(7), 67-73.

[6] Havva Celiktas, Cemal Hanilci. 2017. A study on Turkish text — dependent speaker recognition. 2017 25th Signal Processing and Communications Applications Conference (SIU). 10.1109/siu.2017.7960304

[7] Arsha Nagrani, Joon Son Chung, Andrew Zisserman. 2017. Voxceleb: A large-scale speaker identification dataset. Interspeech 2017. 20.01.2017. 10.21437/interspeech.2017-950

[8] Seyed Omid Sadjadi, Sriram Ganapathy, Jason Pelecanos. 2016. The IBM 2016 speaker Recognition System. The Speaker and Language Recognition Workshop (Odyssey 2016). 10.21437/odyssey.2016-25

[9] Noor Salwani Ibrahim, Dzati Athiar Ramli. 2018. I-vector extraction for speaker recognition based on dimensionality reduction. Procedia Computer Science. 10.1016/j.procs.2018.08.126

[10] Desh Raj, David Snyder, Daniel Povey, Sanjeev Khudanpur. 2019. Probing the information encoded in X-Vectors. 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). 10.1109/asru46091.2019.9003979

[11] Linda Gerlach, Finnian Kelly, and Anil Alexander. 2019. More than just identity: speaker recognition and speaker profiling using the GBR-ENG database. IAFPA conference 2019, 14-17 July 2019, Istanbul, Turkey.

[12] J. Bonastre, and et al. 2003. Person Authentication by Voice: A Need for Caution. EUROSPEECH 2003 – GENEVA.

[13] Vijayaditya Peddinti, Daniel Povey, Sanjeev Khudanpur. 2015. A Time delay neural network architecture for efficient modeling of long temporal contexts. Interspeech 2015. 10.21437/interspeech.2015-647

[14] Brecht Desplanques, Jenthe Thienpondt, Kris Demuynck. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification. Interspeech 2020. 10.21437/interspeech.2020-2650

[15] Varun Sharma and P. K. Bansal. 2013. A Review On Speaker Recognition Approaches And Challenges. International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 Vol. 2 Issue 5, May - 2013

[16] Hansen, John H.L., and Taufiq Hasan. Speaker Recognition by Machines and Humans: A Tutorial Review. IEEE Signal Processing Magazine 32, no. 6 (2015): 74–99. https://doi.org/10.1109/msp.2015.2462851.

[17] Yiyin Zhou. 2008. Speaker Identification. Project Report. Columbia University.

[18] Charan, Rishi, A. Manisha, R. Karthik, and M Rajesh Kumar. A Text-Independent Speaker Verification Model: A Comparative Analysis. 2017 International Conference on Intelligent Computing and Control (I2C2), 2017. https://doi.org/10.1109/i2c2.2017.8321794.

[19] Kunzel, Hermann. Automatic Speaker Recognition of Identical Twins. International Journal of Speech Language and the Law 17, no. 2 (2011). https://doi.org/10.1558/ijsll.v17i2.251.

[20] Antony, Anett, and R. Gopikakumari. Speaker Identification Based on Combination of MFCC and UMRT Based Features. Procedia Computer Science 143 (2018): 250–57. https://doi.org/10.1016/j.procs.2018.10.393.

[21] Ozgur Orman and Levent Arslan. 2001. Frequency Analysis of Speaker Identification. A Speaker Odyssey. The Speaker Recognition Workshop Crete, Greece, June 18-22, 2001

[22] Furui, Sadaoki. Speech and Speaker Recognition Evaluation. Text, Speech and Language Technology, n.d., 1–27. https://doi.org/10.1007/978-1-4020-5817-2_1.

[23] S. Subha and P. 2015. Kannan. Speaker Identification Techniques – A Survey. International Journal of Advanced and Innovative Research (2278-7844), 190, Volume 4 Issue 10

[24] Van, Khoa N., Tri P. Minh, Thang N. Son, Minh H. Ly, Tin T. Dang, and Anh Dinh. Text-Dependent Speaker Recognition System Based on Speaking Frequency Characteristics. Future Data and Security Engineering, 2018, 214–27. https://doi.org/10.1007/978-3-030-03192-3_16.

[25] E.O., Aliyu, Adewale O. S., and Adetunmbi A. O. Development of a Text-Dependent Speaker Recognition System. International Journal of Computer Applications 69, no. 16 (2013): 1–7. https://doi.org/10.5120/12043-7021.

[26] Ghahabi, Omid. Deep Learning for I-Vector Speaker and Language Recognition: A Ph.D.. Thesis Overview. IberSPEECH 2018, 2018. https://doi.org/10.21437/iberspeech.2018-37.

[27] Senoussaoui, Mohammed, Patrick Cardinal, Najim Dehak, and Alessandro L. Koerich. Native Language Detection Using the I-Vector Framework. Interspeech 2016, 2016. https://doi.org/10.21437/interspeech.2016-1473.

[28] Farnaz Ganjeizadeh, Howard Lei, Andrew Maganito and Gopikrishnan Pallipatta. 2014. Reducing the Computational Complexity of the GMM-UBM Speaker Recognition Approach. International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 Vol. 3 Issue 3, March - 2014

[29] Boulkenafet, Z., M. Bengherabi, F. Harizi, Omar Nouali, and Cheriet Mohamed. 2013. Forensic Evidence Reporting Using GMM-UBM, JFA and I-Vector Methods: Application to Algerian Arabic Dialect. 2013 8th International Symposium on Image and Signal Processing and Analysis (ISPA*)*, 2013. https://doi.org/10.1109/ispa.2013.6703775.

[30] Kinnunen, T., E. Karpov, and P. Franti. 2006. Real-Time Speaker Identification and Verification. IEEE Transactions on Audio, Speech and Language Processing 14, no. 1 : 277–88. https://doi.org/10.1109/tsa.2005.853206.

[31] Poddar, Arnab, Md Sahidullah, and Goutam Saha. 2015. Performance Comparison of Speaker Recognition Systems in Presence of Duration Variability. 2015 Annual IEEE India Conference (INDICON), 2015. https://doi.org/10.1109/indicon.2015.7443464.

[32] Biagetti, Giorgio, Paolo Crippa, Alessandro Curzi, Simone Orcioni, and Claudio Turchetti. 2015. Speaker Identification with Short Sequences of Speech Frames. Proceedings of the International Conference on Pattern Recognition Applications and Methods, 2015. https://doi.org/10.5220/0005191701780185.

[33] Schmidt, Ludwig, Matthew Sharifi, and Ignacio Lopez Moreno. 2014. Large-Scale Speaker Identification. 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014. https://doi.org/10.1109/icassp.2014.6853878.

[34] Jahangir, Rashid, Ying Wah Teh, Uzair Ishtiaq, Ghulam Mujtaba, and Henry Friday Nweke. Automatic Speaker Identification through Robust Time Domain Features and Hierarchical Classification Approach. Proceedings of the International Conference on Data Processing and Applications - ICDPA 2018, 2018. https://doi.org/10.1145/3224207.3224213.

[35] L. Ferrer, H. Bratt, V. R. Gadde, S. Kajarekar, Elizabeth Shriberg, M. Sonmez, A. Stolcke, A. Venkataraman. 2003. Modeling duration patterns for speaker recognition. INTERSPEECH

[36] Rawande Karadaghi. 2017. Open-set Speaker Identification. PhD Thesis. University of Hertfordshire.

[37] Rawande Karadaghi, Heinz Hertlein, and Aladdin Ariyaeeinia. "Open-Set Speaker Identification with Diverse-Duration Speech Data." *SPIE Proceedings*, 2015. https://doi.org/10.1117/12.2176335.

[38] Zhao, Xiaojia, Yuxuan Wang, and DeLiang Wang. Robust Speaker Identification in Noisy and Reverberant Conditions. 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014. https://doi.org/10.1109/icassp.2014.6854352.

[39] Wu, Zhizheng, Sheng Gao, Eng Siong Cling, and Haizhou Li. A Study on Replay Attack and Anti-Spoofing for Text-Dependent Speaker Verification. Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific, 2014. https://doi.org/10.1109/apsipa.2014.7041636.

[40] M. Honda. 2003. Human Speech Production Mechanisms in NTTTechnical Review.

[41] Lu, Lie, and Alan Hanjalic. 2009. Audio Representation. Encyclopedia of Database Systems, 160–67. https://doi.org/10.1007/978-0-387-39940-9_1442.

[42] Hermansky H. 1990. Perceptual Linear Predictive (PLP) analysis of speech. J Acoust Soc Am 87(4):1738–1752

[43] K.R. Aida–Zade, C. Ardil and S.S. Rustamov. 2006. Investigation of Combined use of MFCC and LPC Features in Speech Recognition Systems. World Academy of Science, Engineering and Technology 19.

[44] Jain, Shivani, and Brij Kishore. Comparative Study of Voice Print Based Acoustic Features: MFCC and LPCC. International Journal of Advanced engineering, Management and Science 3, no. 4 (2017): 313–15. https://doi.org/10.24001/ijaems.3.4.5.

[45] Biswas, Sangeeta, Shamim Ahmad, and Khademul Islam Molla. 2007. Speaker Identification Using CEPSTRAL Based Features and Discrete Hidden Markov Model. International Conference on Information and Communication Technology, 2007. https://doi.org/10.1109/icict.2007.375398.

[46] Vergin R. 1999. O'shaughnessy D,Farhat A. Generalized mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition. IEEE Trans Speech Audio Process 7(5):525–532

[47] Milosevic, Nemanja. 2020. Introduction to Convolutional Neural Networks. https://doi.org/10.1007/978-1-4842-5648-0.

[48] Tan, Choon Beng, Mohd Hanafi Hijazi, Norazlina Khamis, Puteri Nor Nohuddin, Zuraini Zainol, Frans Coenen, and Abdullah Gani. A Survey on Presentation Attack Detection for Automatic Speaker Verification Systems: State-of-the-Art, Taxonomy, Issues and Future Direction. Multimedia Tools and Applications 80, no. 21-23 (2021): 32725–62. https://doi.org/10.1007/s11042-021-11235-x.

[49] Bai, Zhongxin, and Xiao-Lei Zhang. Speaker Recognition Based on Deep Learning: An Overview. Neural Networks. 140 (2021): 65–99. https://doi.org/10.1016/j.neunet.2021.03.004.

[50] Sakoe H, Chiba S. 1978. Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans Acoust Speech Signal Process 26(1):43–49

[51] Rabiner L, Juang B. 1986. An introduction to hidden Markov models. IEEE ASSP Magazine 3(1):4–16

[52] Reynolds D. 2015. Gaussian mixture models. Encyclopedia of biometrics, pp 827–832

[53] Reynolds DA, Quatieri TF, Dunn RB. 2000. Speaker verification using adapted Gaussian mixture models. Digit Signal Proc 10(1–3):19–41.

[54] Finnian Kelly. 2014. Automatic Recognition of Ageing Speakers. PhD Thesis. Trinity College.

[55] Dehak N, Dumouchel P, Kenny P. 2007. Modeling prosodic features with joint factor analysis for speaker verification. IEEE Trans Audio Speech Lang Process 15(7):2095–2103.

[56] Dehak N, Kenny P, Dehak R et al. 2011. Front-end factor analysis for speaker verification. IEEE Trans Audio Speech

Lang Process 19(4):788–798.

[57] Hatch AO, Kajarekar SS, Stolcke A. 2006. Within-class covariance normalization for SVM-based speaker recognition. INTERSPEECH.

[58] Ioffe S. Probabilistic linear discriminant analysis. 2006. European conference on computer vision. Springer, Berlin, pp 531–542.

[59] Jorgen J. Antonsen. 2017. Open Set Speaker Identification. Master Theses, Department of Electronic Systems, Norwegian University of Science and Technology

[60] Javier Ramírez, Juan Gorriz, and Jose Segura. 2007. Voice Activity Detection. Fundamentals and Speech Recognition System Robustness. Robust Speech Recognition and Understanding. https://doi.org/10.5772/4740.

[61] Finnian Kelly, Anil Alexander, Oscar Forth, and David van der Vloed. 2019. From i-vectors to x-vectors – a generational change in speaker recognition illustrated on the NFI-FRIDA database**.** 15th July 2019, IAFPA conference, Istanbul

[62] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang, Phoneme recognition using time-delay neural networks, IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 37, no. 3, pp. 328–339, Mar. 1989.

[63] Kumar, Ankit, and Rajesh Kumar Aggarwal. Hindi Speech Recognition Using Time Delay Neural Network Acoustic Modeling with I-Vector Adaptation. International Journal of Speech Technology 25, no. 1 (2020): 67–78. https://doi.org/10.1007/s10772-020-09757-0.

[64] Abien F. Agarap. 2018. Deep Learning using Rectified Linear Units (ReLU).

[65] Mirco Ravanelli, et al. 2021. SpeechBrain: A General-Purpose Speech Toolkit. https://speechbrain.github.io/

[66] Mirco Ravanelli, et al. 2021. SpeechBrain: A General-Purpose Speech Toolkit.

[67] Bing Xu, Naiyan Wang, Tianqi Chen and Mu Li. 2015. Empirical Evaluation of Rectified Activations in Convolutional Network.

[68] Soren Becker, Marcel Ackermann, Sebastian Lapuschkin, and Klaus-Robert Muller, and Wojciech Samek. Interpreting and Explaining Deep Neural Networks for Classification of Audio Signals. CoRR abs/1807.03418, 2018

[69] Samir Rustamov and Natavan Akhundova. 2020. Azerbaijani AudioMNIST dataset of digits. https://github.com/Natavan-A/AudioMNIST_AZ

[70] Ko, Tom, Vijayaditya Peddinti, Daniel Povey, Michael L. Seltzer, and Sanjeev Khudanpur. A Study on Data Augmentation of Reverberant Speech for Robust Speech Recognition. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017. https://doi.org/10.1109/icassp.2017.7953152.

[71] Habets, Emanuël A., and Sharon Gannot. Generating Sensor Signals in Isotropic Noise Fields. The Journal of the Acoustical Society of America 122, no. 6 (2007): 3464–70. https://doi.org/10.1121/1.2799929.

[72] Attenborough, Keith. Sound Propagation in the Atmosphere. Springer Handbook of Acoustics. 2007, 113–47. https://doi.org/10.1007/978-0-387-30425-0_4.

[73] Snyder, David, Guoguo Chen, and Daniel Povey. MUSAN: A Music, Speech, and Noise Corpus. 2015. https://doi.org/https://doi.org/10.48550/arXiv.1510.08484.