TEXT-INDEPENDENT, AUTOMATIC SPEAKER RECOGNITION SYSTEM EVALUATION WITH MALES SPEAKING BOTH ARABIC AND ENGLISH

by

SAFI S. ALAMRI

B.S., King Saud University, 2006

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado

in partial fulfillment of

the requirements for the degree of

Master of Science

Recording Arts

2015

This thesis for the Master of Science degree by

Safi S. Alamri

has been approved for the

Recordings Arts Program

By

Catalin Grigoras, Chair

Jeff M. Smith

Leslie Gaston

November 19st, 2015

Alamri, Safi, Saad (M.S., Recording Arts)

Text-independent, Automatic Speaker Recognition System Evaluation with Males Speaking Both Arabic and English

Thesis directed by Professor Catalin Grigoras

ABSTRACT

Automatic speaker recognition is an important key to speaker identification in media forensics and with the increase of cultures mixing, there's an increase in bilingual speakers all around the world. The purpose of this thesis is to compare text-independent samples of one person using two different languages, Arabic and English, against a single language reference population. The hope is that a design can be started that may be useful in further developing software that can complete accurate text-independent ASR for bilingual speakers speaking either language against a single language reference population. This thesis took an Arabic model sample and compared it against samples that were both Arabic and English using and an Arabic reference population, all collected from videos downloaded from the Internet. All of the samples were text-independent and enhanced to optimal performance. The data was run through a biometric software called BATVOX 4.1, which utilizes the MFCCs and GMM methods of speaker recognition and identification. The result of testing through BATVOX 4.1 was likelihood ratios for each sample that were evaluated for similarities and differences, trends, and problems that had occurred.

The form and content of this abstract are approved. I recommend its publication.

Approved: Catalin Grigoras.

DEDICATION

I dedicate this to my wife and three beautiful children for being the sunlight in my life. Also to my parents for making me the person I am today. Lastly, to my friends who helped me achieve my goals and push me to always go further.

ACKNOWLEDGEMENTS

I would like to thank my wonderful advisor Catalin Grigoras for all of his support and help with this thesis and everything I've learned. He's helped me overcome my difficulties every step of the way. Also to Jeff Smith for all of the knowledge and support he's given me throughout my time here. And to Leah Haloin for being the backbone for our program and making sure I'm always on track. I hope this thesis is a testament for all the time and knowledge you all have invested in me.

I'd like to also thank Nathanial Lynch for helping me with everything for my international status and helping me stay here to study and achieve my academic goals.

Lastly, I'd like to say thank you to the Saudi Arabian Cultural Mission to the U.S. and my Kingdom as a whole for helping me to get to and get through my education in the U.S. I am blessed to have such an opportunity to learn and further develop my skills in such an outstanding program that will certainly reflect on the practical side of my work in the criminal forensics field.

TABLE OF CONTENTS

CHAPTER

LIST OF ABBREVIATIONS

ASR	Automatic Speech Recognition
RBFs	Radial Basis Functions
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
FASR	Forensic Automatic Speaker Recognition
FBI	Federal Bureau of Investigations
FFT	Fast Fourier Transform
FSC	Forensic Speaker Comparison
GMM	Gaussian Mixture Model
GMM-UMB	Gaussian Mixture Model-Universal Background Model
HASR	Human Assisted Speaker Recognition
HMM	hidden Markov model
LL	Log Likelihood
LLR	Log Likelihood Ratio
LR	Likelihood Ratio
MFCC	Mel-Frequency Cepstral Coefficient
NIST	National Institute of Standards and Technology
NNs	Neural Networks
PCM	Pulse Code Modulation
SNR	Signal to Noise Ratio
UBM	Universal Background Model
VOIP	Voice over Internet Protocol
VQ	Vector Quantization
WAV	WAVeform audio format

LIST OF FIGURES

Figure

LIST OF TABLES

Table

CHAPTER I

INTRODUCTION

This chapter presents the Automatic Speaker Recognition problems and actual solutions. This includes the systems, history, and concepts. This chapter also goes on to explain Forensic Automatic Speaker Recognition. The objectives of this paper are also explained. The last pieces of this chapter are the motivation behind this research, the benefit, and uses of the the technologies and tools.

Automatic Speaker Recognition Background and History

Automatic Speaker Recognition (ASR) can be classified into two fundamental tasks: Speaker Identification and Speaker Verification. (1) Speaker Identification is determining who the speaker is in the provided sample. The speaker usually has no identity, so it is generally assumed the unknown speaker must come from a set of known speakers fixed from the system. (2) Speaker Verification is determining whether or not the speaker is the claimed person based on the results of Speaker Identification. In this thesis, the emphasis is on Speaker Identification. The process can be classified as either text-dependent or text-independent. This all depends on the cooperation of the involved parties and the available information. The text-dependent application requires the speaker to speak a pre-determined text in order to identify. The text-independent application is designed to identify the speaker through the recognition system regardless of what the speaker says. (2) These apply to the various systems used in ASR.

ASR systems have several applications, both commercial and forensic. Some of the commercial applications include telephone banking, voicemail, prison call monitoring,

voice dialing, and biometric authentication. (3) The focus here is on the forensic application. This includes systems like Batvox, Matlab, and Vocalise. It can be used for investigative and evidential purposes. These systems have two prominent processes: feature extraction and classification. Feature Extraction takes small portions of samples that will be stored and used later on for identification while discarding the useless information, like background noise. The most common technique for feature extraction is the Mel-Frequency Cepstral Coefficients (MFCCs). (4) Classification is a two-phase process that starts with speaker modeling, which is the features of a new speaker and then uses speaker matching, which includes the features saved in the database. (3)

The first attempts at ASR were made by Pruzansky from Bell Labs in the 1960s. (5) He used digital filter banks and spectrograms to correlate sample producing a similarity measure. Taking this method, Pruzansky and Matthews worked together to further develop and add linear discriminators. It was Doddington at Texas Instruments who later took out filter banks and put in formant analysis. (5) It was during the 1970s that text-independent and text-dependent methods were developed. During the 1980s, the hidden Markov model (HMM) was applied directly to text-dependent models and applied to text-independent models along with Vector Quantization. The central theme of the 1990s was to increase robustness of systems. There was also a shift in normalizing the likelihood values of intra-speaker variation, which was further developed upon in the 2000s. In the 2000s, there was also a shift in creating systems that could be used for commercial use. (5)

Models of Automatic Speaker Recognition

For text-independent application, there must be a speaker model in place. The speaker model is a recognition system that has trained speaker samples stored in a database

that uses acoustic feature vectors extracted from each trained sample as comparison to any given sample. This is what allows the text-independent application to have no restrictions on the words the speaker can use, but also makes it a more challenging method of ASR because of the different linguistic content and potential phonetic mismatch. (4) There are many models, such as the Hidden Markov Model (HMM), Mel-Frequency Cepstral Coefficients (MFCCs), Vector Quantization (VQ), Gaussian Mixture Model (GMM), Neural Networks (NNs), and Radial Basis Functions (RBFs). (4)

The two models used in this paper are the Mel-Frequency Cepstral Coefficients (MFCCs) and the Gaussian Mixture Model-Universal Background Model (GMM-UMBs). MMFCs are the most successful features in ASR. The goal of the MFCC is to model the vocal tract's spectral envelope consisting of the formants and a smooth curve connecting them and using it as an identifier. This happens by taking the spectral envelope and d applying a filter based on human perception experiments, known as Mel-Frequency analysis, which applies filters to the spectral envelope and creates the spectrum known as the Mel-Spectrum. Cepstral transformation is then performed on the Mel-Spectrum and the outputs are the MFCCs and speech is then represented as a sequence of cepstral vectors. (6)

GMMs are the most successful speaker model in ASR. The first approach to speaker modeling was Vector Quantization (VQ) in the 1980s. (7) VQ maps the features by associating them with quantized, non-overlapping feature spaces and organizing it all into a 'code book'. (7) An extension of VQ is the GMM, which takes overlapping feature clusters with a non-zero probability. GMMs have become the dominant modelling approach in speaker recognition since its introduction is 1995 by Douglas Reynolds and

Richard Rose. (8) The GMM is a weighted cumulative of the features observed from a sample when compared to the trained model, the outcome being the Log Likelihood (LL). The higher the value of the LL, the higher probability that the model and evidence are the same speaker. The GMM is representation of the cumulative observed features from the speaker that were taken from the underlying model. Originally, GMM's were trained as estimates of maximum likelihood ratio from the data with a free-range parameter that was relevant to the number of Gaussians. Soon, a Universal Background Model (UBM) was represented by a GMM, helping to normalize a score and add substantial robustness to the duration variability. (9)

Forensic Automatic Speaker Recognition

Forensic Automatic Speaker Recognition (FASR) has been around since the 1960s, where it was created to make it easier and more accurate to do speaker recognition. This meant creating an algorithm that then makes quantitative analysis of the speech signals. (4) This has been further developed with the use of models to present the data and algorithms. These models, talked about above, are the baseline of FASR and help to get the most accurate results. This is by using them for hypothetical-deductive reasoning based on the Bayes theorem, which takes new data and combines it with background data to give posterior odds for an adequate outcome, called the likelihood ratio (LR). (4) There is commercial use available of ASR, but FASR systems are more advanced in the options available and the depth of the testing. FASR is used by forensic scientists and they are often highly trained in media forensics. Regardless of who is conducting FASR, there are still drawbacks.

Issues in Forensic Automatic Speaker Recognition Systems

One of the biggest problems forensic scientists face is non-optimal quality of the given recordings. It's hard to get proper results when a recording is full of background noise, very quiet, full of buzz, or poorly transmitted. These environmental factors make it hard for a forensic scientist to work with a voice sample without using some form of audio enhancement to clean up the sample. (10) The recording device also makes a difference. The recording will not be as good over phone lines with low bandwidth availability as well as on low quality devices. There is also a lot of question surrounding new technologies and transmission effects, such as the Voice over Internet Protocol (VOIP) transmissions. VOIP transmissions are programs like Skype and Facetime, where the communication is voice, but using a non-telephone based system. Recordings downloaded from the Internet are compressed to decrease file size resulting in decreased voice information and overall quality. There is also a general concern in regards to recording equipment, such as clip-on microphones and hidden recording devices. Having concealed recording equipment can give the recording different acoustic filtering effects, and decrease quality of the recorded sample. (10) Another part of that problem is voice changing and voice disguise. When someone talks in a different style, tone, volume, with something blocking the mouth, or with a voice disguiser, it makes it harder for the FASR systems to correctly identify. (11) At this point, there needs to be some human-based speaker recognition done before any ASR evaluation can be done. Another issue that forensic scientists face is an uncooperative speaker. Sometimes, whether a criminal case or private forensic analysis, there is an uncooperative speaker, so it's hard to obtain a sample to compare to the evidence. From there, text-independent FASR can be used and the evidence can be compared to a database

of speech samples. (10) These results can be used either for private customers or for the judicial system. Forensic scientists are commonly called upon to do FASR and present the results to the court.

The Likelihood Ratio Approach

Likelihood Ratio is the outcome of FASR testing and is a probability used to determine the level of confidence in the two hypotheses, the Null and Alternative Hypotheses. (10) An example would be a LR used to determine the closeness of a suspect's sample recording, the model, to the evidence recording that was provided. In this case, the null hypothesis would state that the suspect is not the speaker in the evidence, while the alternative hypothesis would state the suspect is the speaker in the evidence. This is accomplished by taking features extracted from the recording through MMFCs and front-end processing and comparing the extracted features to the features of a model representative of the claimed speaker using the GMM. The result of the GMM is the LR, which is comprised of the ratio of evidence and model match scores. (9) These scores are then compared to a standard that would either accept or reject the model as the speaker in the evidence. The standard value is the magnitude of the LR from the number 1. (10) For example, if the LR=10, then the evidence would be 10 times more likely to be the model, however if the LR is between 0 and 1, the probability of the model being the speaker in the evidence are less likely. If independent features are analyzed and there is a set of LRs, then the mean of those LRs can be used. In forensics, the LR is an important piece for creating a result to be used in court as evidence. However, the LR cannot be used to 100% determine someone as the evidence speaker. It can be used to set a degree of confidence in the forensic scientist and the people of the court, but cannot be used for complete certainty.

Objective of the Study

- To evaluate cross-language challenges of automatic speaker recognition.
- To determine the use of text independent speaker identification in multilingual environments.
- To evaluate the application of Likelihood Ratio (LR) in measurements of reliability and validity of automatic speaker recognition and forensic voice comparison.
- To find out the benefits provided by the text independent speaker identification.

Motivation

In recent times, there has been an increase in the interactions between Arabic and English people. Globalization has encouraged the movement of people within Arab and English countries. Moreover, recent increase in terrorist activity in the Middle East poses a greater need to identify the people who are involved. This includes the videos seen that are produced by these terrorist organizations. However, there doesn't exist any software or data to help in identifying the English second language/Arabic first language speakers. By taking two samples of the bilingual speaker speaking in English and a sample of the same person in Arabic, there is hope in creating a likelihood ratio (LR) that uses both languages. That LR can then be used to identify the person using either an Arabic or English sample and comparing it to the available database.

Another motivation is the lack of a likelihood ratio or accompanying system that does what is trying to be accomplished. There are likelihood ratios used in software today that can only identify and compare against samples of the same language. Being able to set it up to compare one sample of a bilingual speaker against two languages and being able

to identify the speaker when comparing one of two languages to the single language databases would make it a lot more successful in the media forensics field.

Benefit

This research will be of use to a wide range of people. For law enforcement agencies, it will enable them to recognize speakers who have heavy accents or speak another language. Moreover, the research sheds the new light on the issue of benefits and innovations in the sphere of automatic speaker recognition programming. The core idea of implementing this technology stems from the need to specify, designate and identify the speech patterns to make more accurate inferences on speaker identity, which can be widely used in forensic analysis, criminal authentication and detection, as well as in ensuring legal access to the computer accounts. Ultimately, this will boost their efforts in combating crime. It will also facilitate in dealing with terrorism, especially if the terrorists communicate in Arabic, since it is the language focused on in this paper. For general purposes, it can be used in media forensics, voice identification, and improving biometric access and speaker recognition for bilingual speakers. An example of this would be biometric speaker recognition technology, which now is not efficient for bilingual speakers with heavier accents. Improving this technology would include making it easier for the recognition system to identify the bilingual speaker using a more accurate likelihood ratio and features specific to the speaker. This will also increase the efficiency of judicial systems because speaker recognition would be more accurate for people who speak either of the languages.

Tools and Technologies Used

BATVOX

Agnitio (2015) defines BATVOX as a biometric tool that uses advanced technology and it is designed to compile expert reports for evidence purposes and performing speaker verification, by forensic experts. (11) Batvox Basic is a specialist 1:1 tone biometric device designed for investigation experts as well as scientific police who conduct voice recognition duties, while BATVOX Pro can be used by large organizations with multiple users (Agnitio, 2014). (12) It operates by entering audio formats which are run against samples to find a match. This means utilizing the hybrid approach, which increases the strength of conclusions made by the user and producing evidence for court hearings that is precise and reliable (Agnitio, 2015). (11) This technology can be used in compiling professional reports which can be used as evidence in court of law. Batvox has various features, which include: case management and speaker recognition duties. In case management, the consumer is capable of sorting out audios along with voice samples by cases. In order to achieve this single or various calculations are involved. This will facilitate the investigation making it possible for the consumer to identify of the voice. In terms of speaker recognition duties, BATVOX facilitates either recognition of unidentified voices alongside voices imminent from identified speakers or LR computations making one on one comparison. This research focuses on the 4G of Agnitio technology and the improvements that it has seen over the course of development.

Basically Agnitio technology can be classified into generations. This is based on their time of development. These generations are broken down as 1G, 2G, 3G and most recent 4G. 1G was the first technology of Agnitio and it came in from University of Madrid.

The 1G incorporated superior GMM technology through various normalization methods that ranked top 5 within the NIST 2004 assessment. 2G came in 2007, but this time it added channel compensation methods. Agnitio created 2G with the assistance of Universidad Autonoma de Madrid. Development of 3G based on Joint Factor Analysis (JFA) happened in 2009. JFA is a creation of Kenny from CRIM in 2008. JFA works on the assumption that nearly all of the inconsistency is accounted by channel factors as well as a speaker. 3G has great accuracy with enhanced rate of operation. With 3G, it was possible to divide speakers into independent audio parts. 3G has the capabilities for gender identification and streaming analysis. It was now achievable to sense the number of speakers in an audio. Development of 3G was under the leadership of Niko Brummer who implemented expertise from NIST, speaker identification evaluation, year 2000 – 2008.

In 2012, there was the release of 4G, which is because of i-vector, an improvement of JFA model. 4G has speed of up to 10 times faster than 3G with memory required performing voice detection. This advancement could operate on smart phones as well as tablets and seek out over 1 million tone of voice print catalog with a sole server in only some few minutes.

Text independent expertise applies where there is natural as well as spoken speech. In most scenarios, natural, conversational speech is normally unavailable for authentication. Therefore, short and set passphrase conversations are used. Text-dependent methods are less flexible but more accurate than text-independent technologies because a lot of information is used. In the event that the case has two speakers in the audio, automatic segmentation applies preceding gender recognition for each active speaker, then text independent recognition it applied. 4G technology has further developed text-independent

technology, which has continued to prevail over the major weak points and stabilize its use in media forensics.

Audio Enhancement

Audio enhancement refers to the process of eliminating noise from audio files that are of poor quality. As a result, the quality of the file is improved for analysis. Audio enhancement can be done through the time domain level detection or the frequency domain filtration. (12) In the time level domain detection, the amplitude envelope of the audio signal is treated. The general level of the audio file is leveled to determine the audio exists in the background when the desired signal is absent. In the frequency domain filtration, the spectral subtraction is used to reconstruct the signal. This technique ensures that the most appropriate audio file is obtained, and therefore, it can complement other approaches. Improving the quality of the audio will require the use of software and hardware. These would include Fast Fourier Transform (FFT) Analyzers, computers, speakers, printers, digital filters, audio equipment, and proprietary software playback systems, among others. (13)

Summary

In this chapter, we explored the history, background, and details of Automatic Speaker Recognition. We also talked about methods and models used in ASR, such as text-independent. Another topic we discussed was Forensic Automatic Speaker Recognition, with a brief history, conditions, and applications in the field of forensics. Then we discussed the likelihood ratio and its relevance in FASR. We listed out the objectives of this article as well, while also talking about the motivations and benefits of this research.

The last subject we touched on was the tools and technologies we used, which were audio enhancement and BATVOX.

CHAPTER II

LITERATURE REVIEW

Cross-language Challenges of Automatic Speaker Recognition

Text-independent speaker recognition in English versus Arabic environments involves the use of two languages. In military and intelligence operations, automatic speaker recognition systems have been experiencing cross-language challenges for several years, even though communication channels are being monitored real time and on a large scale. Nevertheless, the impact of cross-language on these systems are not disclosed because of various reasons. (14) Furthermore, these systems have not been adequately researched despite many countries being multi-ethnic or multi-lingual. A portion of the research that has already been conducted on automatic speaker recognition systems, however, has investigated various elements of the cross-language problems. Research on cross-language training did not yield results on the individual effects of the variables because they were combined. In a test that was based on a Mixer Corpus, which contains English and Arabic languages among others, results indicated that same languages had a higher performance than unmatched languages. Nonetheless, Künzel (2013) states that the cross-language problem may be attributed to the capacity of the normalization procedure.

Automatic forensic speaker recognition systems are affected by the cross-language problem in a quantitative way because advanced automatic systems use low-level acoustic features instead of the highly specific language features. These low-level features depict the characteristics of the general resonance behavior of the vocal tract of the speaker. Consequently, adequate amount and quality of speech material will ensure that language

usage will be minimized. Various language mismatch types are probable with the extreme ones implying speaker models of different languages, reference populations, and test samples. There is also a possibility of matching a test sample for language and reference population, but the system response will be affected in the opposite way. On the other hand, increasing the similarity between test samples and the reference population will reduce the number of false rejections, whereas the number of false acceptances will increase. (14)

Text Independent Speaker Identification in Multilingual Environments

Most forensic labs of the US government use automated and manual voice analysis tools to determine the possibility of a match between the suspect's voice and the speaker in the evidence. The Federal Bureau of Investigations (FBI) has installed a Forensic Automatic Speaker Recognition (FASR) system that is characterized by both text and channel independence. Notwithstanding, the US had been experiencing terror attacks. This developed the need to incorporate new capabilities into the FASR system so that it could deal with terrorists who used other languages, apart from English. (16) Multilingual data can facilitate the development of systems that are capable of recognizing multilingual speakers despite using any language. However, previous assumptions that language differences in acoustic approaches does not affect performance have not yet been proved. (15)

On the other hand, Gold and French (2011) argue that Forensic Speaker Comparison (FSC) can be analyzed by Automatic Speaker Recognition System (ASR) and the ASR with human analysis, known as Human Assisted Speaker Recognition (HASR). In ASR, a specialist software is used to determine the level of similarities between speech samples and it is based on statistical models of features that have been automatically

extracted from recordings. In HASR, an automatic system is used alongside the analysis of the acoustic-phonetic and/or auditory kind. (16)

In recent times, speaker recognition systems that are based in multilingual environments have received special attention from researchers. As a result, speaker models have been trained and tested in different languages. (17) Previous research shows that in conditions where there is a language mismatch, the level of accuracy tends to decline. However, they fail to show how that problem can be alleviated. On the contrast, research by Luengo et al. (2008) seeks to maintain recognition accuracy in language-mismatched conditions by finding a robust parameterization. Most speaker recognition systems use the Gaussian mixture models (GMM) of short-term spectral features that characterizes the vocal tract filter during articulation. This system can capture the vocal tract characteristics of each phoneme and speaker. On the other hand, text-independent speaker recognition systems develop various problems in a multilingual environment because the model is trained in one language but tested in another.

The problem, however, is that the different phonetic content of both languages increases the system error rate. (17) This problem can be solved by training and testing the system using recordings from the two languages. As a result, the characteristics of phonemes will be learned. But, this challenge can also be solved by ensuring that each language uses a different speaker model. A language detector can be used to determine the appropriate model for each language. (17)

Likelihood-Ratio Systems (LR)

The LR can be applied in various evidence types in which the question is whether two samples have a similar origin. In an LR framework, forensic scientists are tasked with the probability of obtaining similar observations between a sample of unknown origin and that whose origin is known. This analysis is done under two hypotheses that are different and contrasting in nature. (18) The validity of LR systems is measured with a large number of test samples that are of known origins. These samples are run through the system, and the output is determined whether to be good or bad based on its desired value. Measuring validity depends on both the test set and the system. A relevant validity measurement requires the test samples to be closely matched with the case trial conditions. (18) If a system declares that two samples have the same origin whereas they have different origins, then an error occurs. Correct-classification rates and log-likelihood-ratio (LLR) cost can be used as validity metrics.

However, Morrison (2010) notes that correct-classification rates result from binary decisions that are formulated using posterior probabilities. The support strength of a likelihood ratio towards a certain hypothesis is determined by the size of the likelihood ratio. (18) Morrison (2010) also argues that the size of the likelihood ratio represents a numeric expression of evidence strength with respect to competing hypotheses.

On the other hand, the log-likelihood-ratio (LLR) cost is a metric of validity, which is appropriate for use in the likelihood-ratio framework. Furthermore, it can be applied in automatic speaker recognition and forensic voice comparison. Morrison (2010) defines forensic voice comparison as the process of comparing audio recordings of a known

speaker with that which the identity of the speaker is unknown with the aim of presenting expert testimony in court. (19)

Reliability and Validity Measurements

In developing forensic comparison systems, validity and reliability measurements are important. When reliability is a primary concern, longer voice recordings should be used instead of shorter voice recordings. Estimates of the extent of reliability and validity of forensic comparison systems should be calculated by forensic scientists who should closely match the test conditions with the trial conditions. (19) Morrison (2010) also indicates that the accuracy of a forensic comparison system can be determined by testing it using a test set of similar or different origins and then comparing the output with the knowledge input. However, comparing known and unknown samples is a task of determining evidence strength and not a binary decision. According to Morrison (2010), the task of a forensic scientist in a likelihood-ratio framework is to present a strength-of-evidence statement to the court. This statement would be responding to the question: 'what is the probability that observed differences between known and unknown samples will be more prominent in one hypothesis than the other contrasting hypothesis?' The first hypothesis states that the known and unknown sample have a similar origin, whereas the other one states that the two samples have a different origin. (19).

The numerator and denominator of the likelihood ratio are known as a similarity term and typicality term. When calculating the strength of evidence, forensic scientists must consider both the degree of typicality and the degree of similarity between the samples based on relevant population. (19) Forensic scientists are required to present the probability of evidence and not the probability of hypotheses because of logical and legal

reasons. The systems they use have to be up to the Daubert ruling, which says that an expert witness' testimony must be based on their scientific knowledge rather than a system that is user-friendly to anyone. This means that in order for a system to be used in court, it must be understood and explainable by the expert who uses it for court. For example, the Batvox system by Agnitio is used by law enforcement in many countries due to its complex, highly qualified system. This means that the expert needs to know the ins and outs of Batvox if he/she were to be presenting evidence ran through Batvox in court. (20) The trier of fact is tasked with determining the probability of guilty or not guilty beyond reasonable doubt. The trier of fact makes decisions based on all the evidence presented in court, and the forensic scientist is required to provide a statement of strength regarding a certain piece of evidence. (19)

Benefit of Automatic Speaker Recognition for Bilingual Speakers

The LR framework is used in various evidence types that include voice and DNA samples, among others. In evidence analysis, the LR framework is used to determine whether two samples with known and unknown identities have a similar origin. Legal and logical reasons require that forensic scientists provide the probability of evidence as opposed to the probability of the hypothesis. This means that the LLR cost is used in the LR for validation. It is also used in automatic speaker recognition and forensic voice comparison. This has played a key role in combating crime because forensic evidence is now admissible in a court of law. Due to automatic speaker recognition and forensic voice comparisons, there is a steady advance in finding models and producing software that would be able to identify speakers who may be speaking another language different from

the native language. This will make in increase in our methods of tackling crime around the world because of the always advancing systems.

Text-independent speaker recognition in Arabic and English environments would enhance speaker identification in both languages. As a result, the government relations between countries that speak English and Arabic would be improved, due to their ability to identify speakers with more advanced dual language ASR systems. This means bringing together those countries in combating crime from extremist groups who speak out in both English and Arabic. Working side by side would also help these countries to work on their foreign policies regarding each other. This can be a big push towards peace and alliance between countries that didn't have that opportunity before. Overall, an ASR system for bilingual purposes would help in combating crime against bilingual speakers who are harder to verify voice, being able to identify the members of extremist groups, and bring together powerful countries and increase their crime-stopping efforts.

CHAPTER III

ANALYSIS AND REQUIREMENTS

Introduction

The purpose of this thesis is to study the Arabic v. Arabic and Arabic v. English voiced samples from different media recordings, made in real life conditions, different that the lab controlled conditions, with no control over: the microphones, the recording equipment and their settings, the environment and background noises of other signals (hum, music, voices, reverberates), and the time delay between samples. The aim of this chapter is to explain the process used and data retrieved from testing the text-independent samples and analyze the results of the Arabic v. Arabic and Arabic v. English likelihood ratios (LRs). The overall process is five steps: preprocessing samples, file extraction, modeling, comparison of the features and models, and analysis of the results. (22)

Methodology

During an investigation there are various ways used to gather and analyze data. In this case, the methodology involves the steps and procedures of how to get the results needed from the input materials. The methodology aims at revealing the processes that takes place in the operations of BATVOX. It also further describes speaker recognition basics as well as how BATVOX organizes and stores memory. Lastly, this covers instructions of how to use BATVOX, as well as what the inputs are, to get a comprehensive result. The BATVOX technology involves an in-depth discussion on the capabilities of 4G technology. This includes an enrollment and testing workflow that work together to identify the gender and identity of the speaker.

Automatic Speaker Recognition with BATVOX

In the enrollment phase of BATVOX, the first step is inputting the speech signal for enrollment. The flow of the ASR process is seen in Figure 1. From there, detection of voice activity is done by processing the speech signals to find the presence of a human voice and finding any insufficient conditions. After the detection, the features are extracted from the speech signal. The goal of feature extraction is to emphasize the relevant data within a signal while removing irrelevant data through lossy compression of the sample. (25) This helps the patterns in the data stick out and become more noticeable in a feature vector when compared to the normal signal. Another look at the ASR system can be seen in Figure 2. The features presented here are the result of Mel Frequency Cepstral Coefficients (MFCC). MFCCs are one of the most common methods to feature extraction used in ASR because they offer a more compact representation of the signal, basically windowing a version of the signal with the necessary features highlighted and unnecessary features removed. This can be used in speech processing applications like language recognition, emotion recognition, and speaker recognition. The MFCC was first introduced in the 1980s by Steven Davis and Paul Mermelstein. (25)

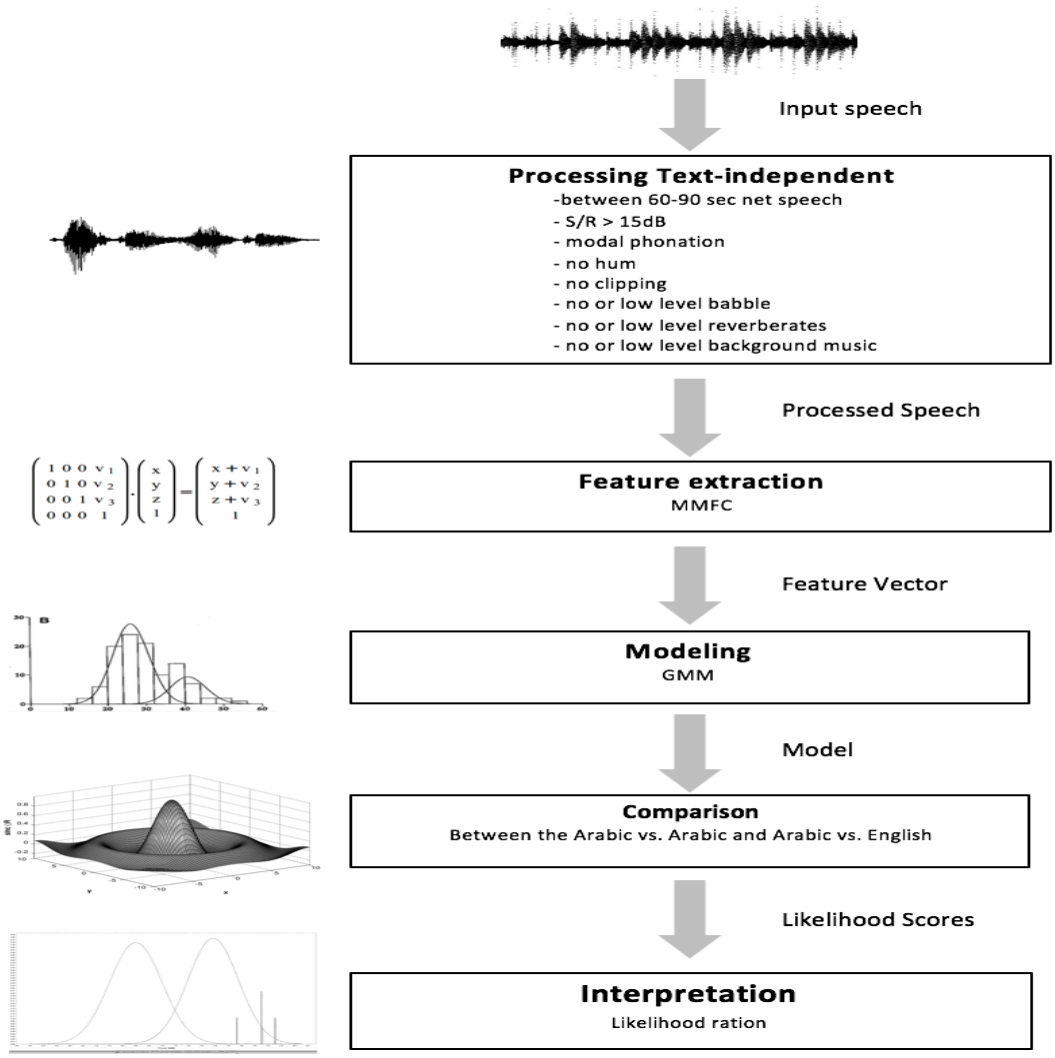
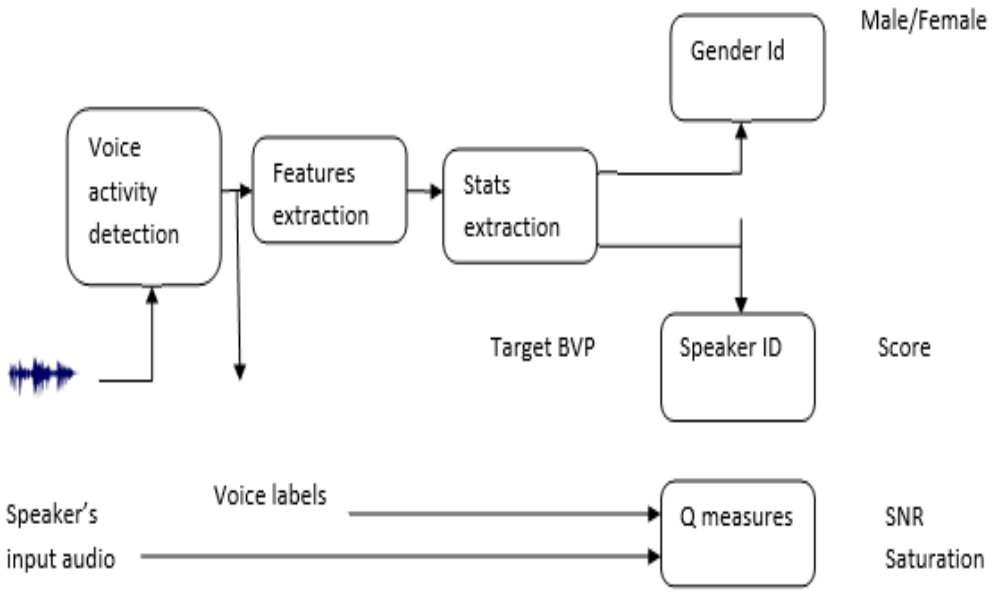


Figure 1: Overview of the Automatic Speaker Recognition Process

The first step to MFCCs is to take a frame of the speech signal from the sample. Then it is pulled apart into frequency components by the fast Fourier transform (FFT), which is a more efficient form of the discrete Fourier transform (DFT). The result of this is a spectral envelope of the signal that has the data and properties of the vocal tract related to the speaker. From there, the Mel frequency scale, which is a set of bandpass filters that give high resolution to lower frequencies, is applied to the spectral envelope. After the filters

are applied, there is logarithmic compression also applied. On top of those, the discrete cosine transform (DCT) rids the signal of all correlation. At the end of all this, an MFCC feature vector is taken from up to the first 20 coefficients and a final feature vector is



created from that.

Figure 2: BATVOX Testing Diagram

Speaker Modeling

Speaker modeling is creating a model that is trained from a set of feature vectors to be used as a basis for comparison against testing samples. In text-independent speaker recognition, there is no relationship between the speaker model and the recognition utterances. (24) This means the model needs to be general enough to fit the average features of a speaker, but different enough to distinguish between features of different speakers. BATVOX compares the GMM of the speech signal to the Universal Background Model

(UBM). UBM has 256 Gaussians full covariance GMM. The stats calculation module gets the zero as well. The i-vector mining module estimates the 400 breadth i-vector from the stats by means of the Total Variability (T) matrix. To judge against two speakers, use a Gender Dependent SPLDA with 120 Eigen voices to evaluate test along with enrolment i-vectors.

Preprocessing

In this section, we will discuss the process of obtaining and enhancing the samples used, which is called preprocessing. According to the National Institute of Standards and Technology (NIST), the conditions that can affect recognition performance are: sex, age, pitch, handset type, noise, number of calls made, dialect, region, and channel. (23) In this study, media networks with an open source were used. The objective was to identify males who could speak both Arabic and English. The recordings we used were taken from YouTube. Usually, quality recordings are made in a studio using lavaliere microphones. The recordings used were up to 4 years apart. The following collections of voiced samples were built:

a) 35 Arabic male speakers for the Arabic Reference Population;
b) 20 Arabic male speaker models, different than (a);
c) 20 Arabic male speaker test files, same speakers as (b);
d) 20 English male speaker model and test files, same speakers as (b, c).

The initial recordings were collected by downloading the files from Internet sources and the samples and their conditions are seen in Table 1. From each speaker, the audio samples were prepared by respecting the following conditions:

a) About 60-90 seconds of net speech length.

b) Signal to noise ratio greater than 15dB.

c) Normal speech / Modal phonation.

d) No major distortions, clipping, background noises or signals such as voices, music or babble. For some of the samples the hum was removed by using notch filters. The audio signals were extracted as WAV PCM 8 KHz sampling frequency, 16-bit, mono files.

These conditions were chosen because they are the minimum requirements for audio samples to be ran successfully through the BATVOX program.

Audio distortion refers to the alteration of the waveform of the signal, the discounting of noise and amplification of sound. Distortion can be linear or nonlinear. (23) In linear distortion, there are frequency and time dependent characteristics of the amplitude and phase response of the transfer function. It can be caused by system inhomogeneities and reflections in the propagation path. In nonlinear distortion, changes in the frequency content of the input result in the transfer of energy from one frequency at the input to several frequencies at the output. (23) Examples of distortion are clipping and clicks, among others. Noise refers to the unwanted signals that interfere with communication, and the processing or measurement of an information-bearing signal. It presents itself in various degrees in nearly all environments. (23) The types of noise include white noise, thermal noise, and acoustic background noise, hum, among others. Noise can result in transmission errors and, therefore, the disruption of the communication process. In communication, noise and distortion are the main limiting factors.

On the other hand, a click refers to a sharp sound like that of a switch being operated or when two hard objects come into contact. Clipping refers to the distortion that limits a signal after it exceeds the threshold. It occurs when a signal is digitized or when a sensor with limited data measurements capacities records a signal. It can also occur when a digital or analog signal is transformed. Reverberates refers to a series of quickly repeated sounds that bounce off a surface like an echo and when present, they affect the forensic analysis of voice samples. In the recordings, some distortions were present whereas others were not. The following distortions were not present: wind, clicks, reverberates, electronic voice disturbances, wow & flutter, mobile phone burst.

However, the following distortions in the form of clippings and lossy compression artifacts were detected and removed. In addition, noise and other signals were detected in the recordings. Periodic noise in the form of tones was detected and fixed whereas sirens were absent. For non-periodic noises, hisses and broadband noises of very low levels were present, and they were fixed. Nonetheless, other forms of non-periodic noises like coughs, page turns, acoustic impulses, and pedals, among others, were not detected. Furthermore, background voices and music were detected and removed.

A parameter refers to a variable that is kept constant during the study or a characteristic that distinguishes one sample from another. In this study, language was used as a parameter. The voice samples were restricted to using only English and Arabic languages. Moreover, the gender of the participants was used as a parameter. Consequently, the participants used were males only. Furthermore, the gender confidence of the voices was included in the parameters. This would ensure that the voice depicted some male

characteristics. As a result, some male voices had a low gender confidence, and they were therefore removed.

Collected Data

The collected data is split into two main categories: Arabic reference population and the Model/Test population. There is 35 people in the Arabic reference population. In the Model/Test population, there is 20 Arabic native speakers who each have 4 samples: Arabic model, Arabic test, English model, and English test. Altogether, there is 55 speakers to work with. The samples come from a variety of Arabic speaking countries, as displayed in Figures 3 through 6. They display the countries the speakers are from. The data was all collected from public media websites. All of the samples were at least 60 seconds long, the voices were strictly males from 20-65, the samples couldn't be longer than 4 years apart, and the speaker was the only voice present during his speaking.

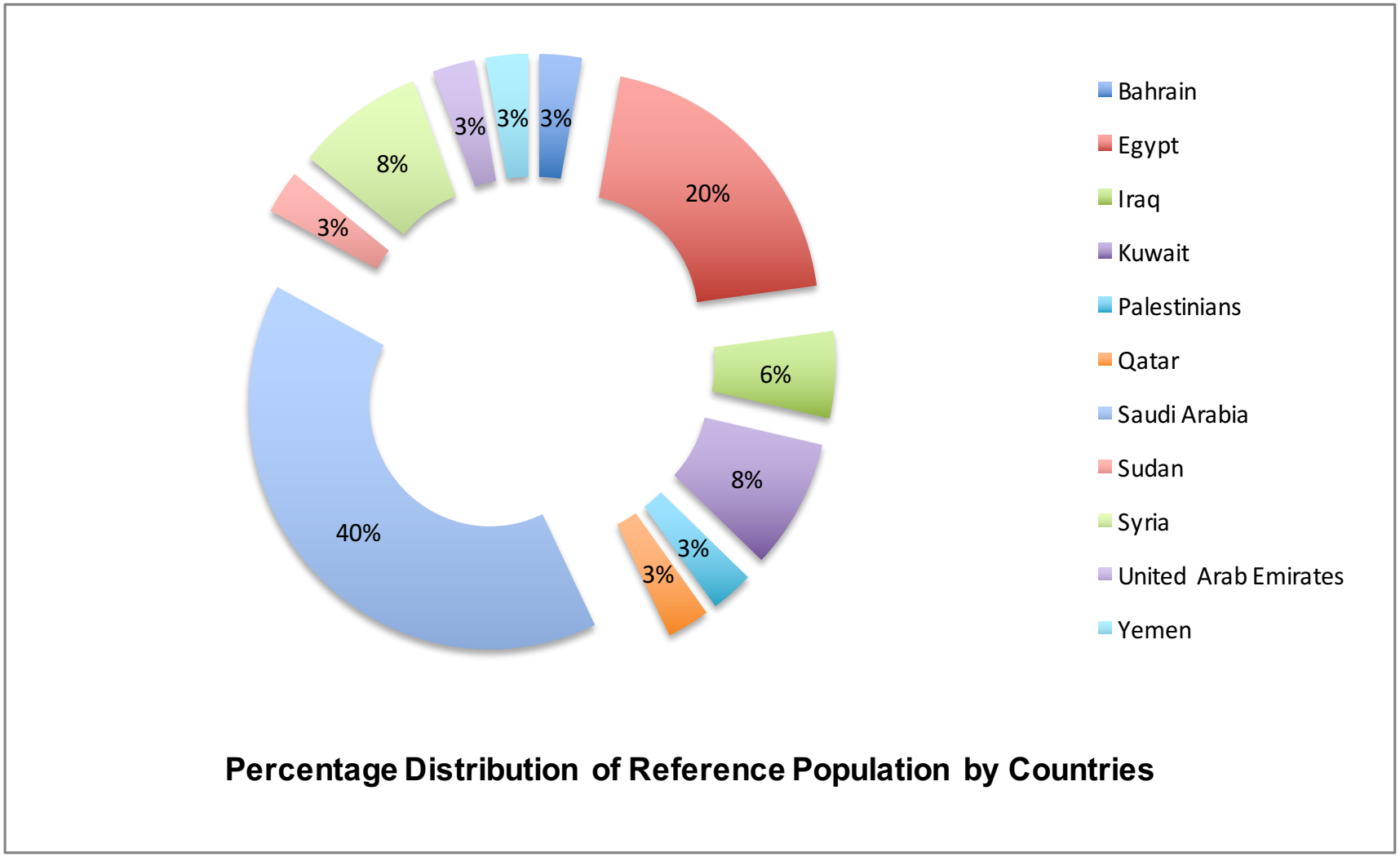


Figure 3: Percentage distribution of Reference Population by Countries

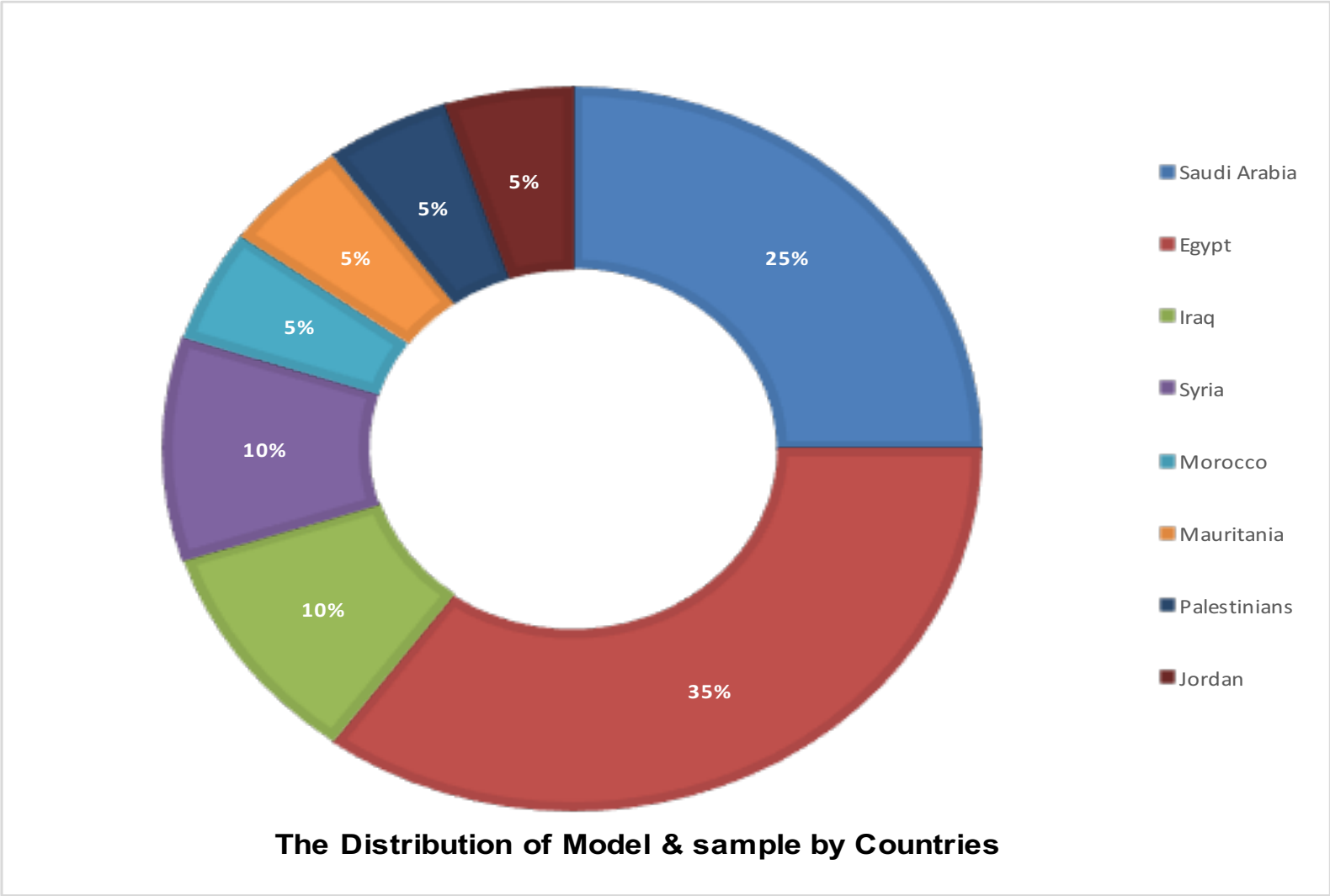


Figure 4: Percentage Distribution of Arabic Models & Samples by Speaker's Country

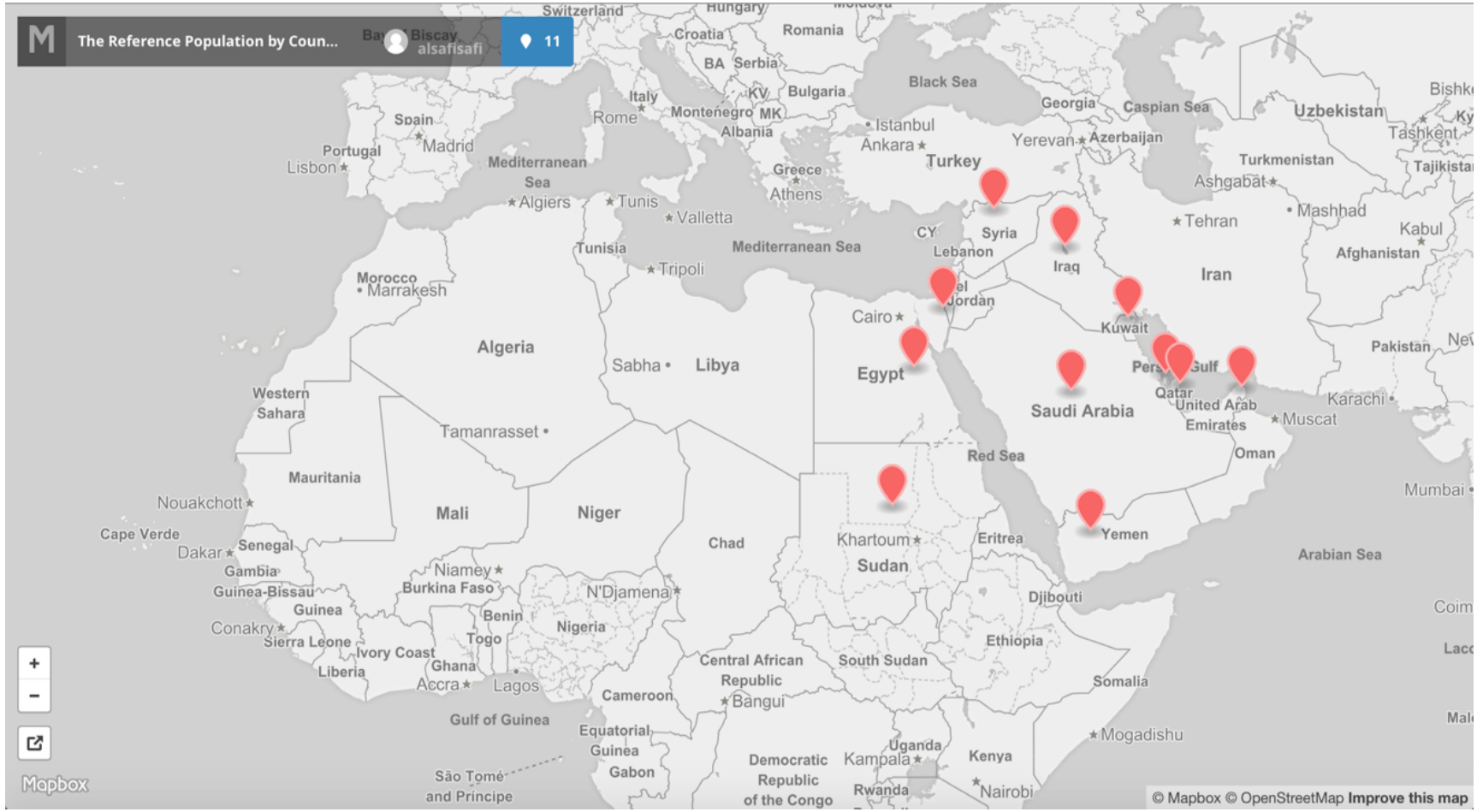


Figure 5: Map Distribution of Arabic Population by Speaker's Country

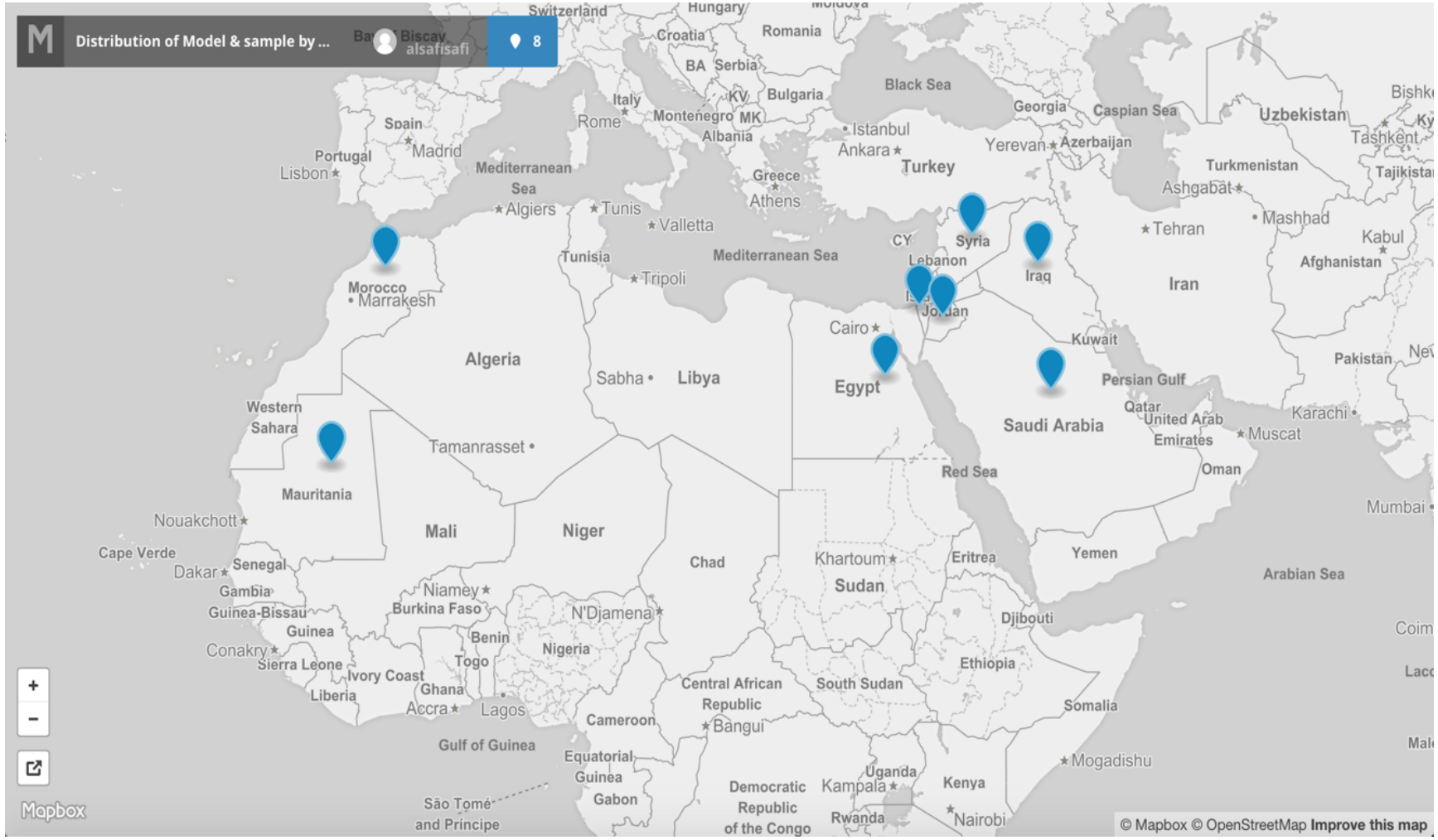


Figure 6: Map Distribution of Arabic Models & Samples by Speaker's Country

Diagram of Speakers

Batvox requires a reference population trained from data provided. This is where the 35 samples from the Arabic reference populations come into play. The two main processes used in Batvox are training and recognition. In training, the suspect is the person whose identity is known and is the voice we want to compare with the evidence. The suspect's sample is then trained into a model that represents the characteristics of that speaker's voice. (22) In recognition, the suspect's model is added to a larger reference population model and compared against the evidence to make a LR and identification. The diagram composed for the speakers is seen in Figure 7.

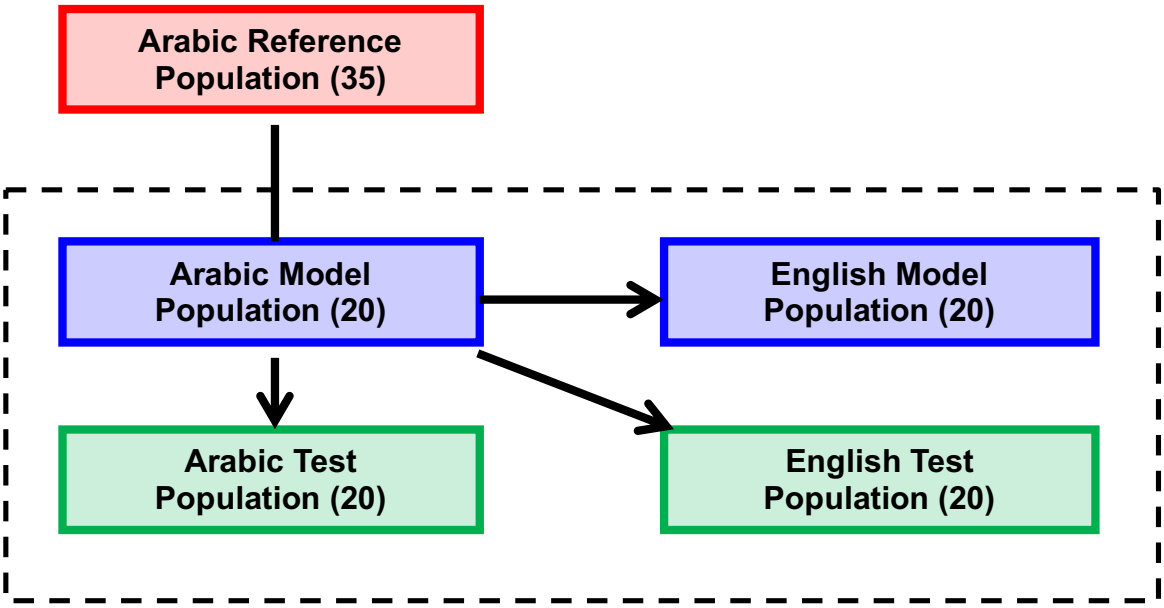


Figure 7: A Diagram Showing the Goals of a Number of Samples Hypothesis

The BATVOX Analysis

In this study, the samples were being tested for speaker recognition in males speaking Both Arabic and English languages. The Arabic models were compared to Arabic and English test samples in an attempt to determine the correct speaker, and the results were reported as likelihood ratios.

The samples were run through BATVOX for forensic analysis. BATVOX Basic was used to conduct the analysis because it only involved a single user. Three tests were done and they included Arabic model vs. Arabic test, Arabic model vs. English model, and Arabic model vs. English test voice samples. For both tests, the Arabic reference population was 35 whereas the test population was different in each test. Furthermore, the model populations were developed from the test population of each test.

Problems

During the first step, the following problems occurred:

- Not long enough net speech sample
- S/N < 15db
- harsh and falsetto phonation samples
- different transmission channels and muffling
- hum
- different degrees of clipping, including heavy
- high level of background music and/or babble.

For short speech, speaker #9 had less than 60 seconds of net speech on his English test. Speaker #2, on his Arabic test, had S/R less than 15db. Speaker #7's English test had

"harsh" and high pitch phonation. The following speakers had different channels: speaker #8 AM, speaker #9 EM and ET, speaker #13 ET, speaker #14 EM and ET, speaker #16 EM and ET, speaker #17 ET, speaker #18 EM, speaker #19 ET, and speaker #20 EM and ET. There was clipping on speaker #5's ET by .56%. Table 1 shows the analytic analysis of the audio samples. Since the voiced signal is affected by distortions and contaminated by other noises, then it is expected the voice feature extraction algorithms to reflect these phenomena. Results from a study by Herman Kunzel in 2014 showed that clipped samples up to 50% had no effect on the S/N when enhanced. This means that speaker #5's ET sample's S/N isn't impacted by clipping the original sample or enhancement on the clipped sample. (26)

Table 1: Speaker Samples with Conditions

Speaker #	Language	Sample	Net speech	S/R	Clipping	Babble	Reverberates	Different channels	Phonation
1	Arabic	Model	>60sec	>15dB	No	No	No	No	Modal
		Test	>60sec	>15dB	No	No	No	No	Modal
	English	Model	>60sec	>15dB	No	No	No	No	Modal
		Test	>60sec	>15dB	No	No	No	No	Modal
2	Arabic	Model	>60sec	>15dB	No	No	No	No	Modal
		Test	>60sec	≈14dB	No	No	No	No	Modal
	English	Model	>60sec	>15dB	No	No	No	No	Modal
		Test	>60sec	>15dB	No	No	No	No	Modal
3	Arabic	Model	>60sec	>15dB	No	No	No	No	Modal
		Test	>60sec	>15dB	No	No	No	No	Modal
	English	Model	>60sec	>15dB	No	No	No	No	Modal
		Test	>60sec	>15dB	No	No	No	No	Modal
4	Arabic	Model	>60sec	>15dB	No	No	No	No	Modal
		Test	>60sec	>15dB	No	No	No	No	Modal
	English	Model	>60sec	>15dB	No	No	No	No	Modal
		Test	>60sec	>15dB	No	No	No	No	Modal
5	Arabic	Model	>60sec	>15dB	No	No	No	No	Modal
		Test	>60sec	>15dB	No	No	No	No	Modal
	English	Model	>60sec	>15dB	No	No	No	No	Modal
		Test	>60sec	>15dB	Yes	No	No	No	Modal
6	Arabic	Model	>60sec	>15dB	No	No	No	No	Modal
		Test	>60sec	>15dB	No	No	No	No	Modal
	English	Model	>60sec	>15dB	No	No	No	No	Modal
		Test	>60sec	>15dB	No	No	No	Yes	Modal
7	Arabic	Model	>60sec	>15dB	No	No	No	No	Modal
		Test	>60sec	>15dB	No	No	No	No	Modal
	English	Model	>60sec	>15dB	No	No	No	No	Modal
		Test	>60sec	>15dB	No	No	No	No	Harsh

8	Arabic	Model	>60sec	>15dB	No	No	No	Yes	Modal
		Test	>60sec	>15dB	No	No	No	No	Modal
	English	Model	>60sec	>15dB	No	No	No	No	Modal
		Test	>60sec	>15dB	No	No	No	No	Modal
9	Arabic	Model	>60sec	>15dB	No	No	No	No	Modal
		Test	>60sec	>15dB	No	No	No	No	Modal
	English	Model	>60sec	>15dB	No	No	No	Yes	Modal
		Test	<60sec	>15dB	No	No	No	Yes	Modal
10	Arabic	Model	>60sec	>15dB	No	No	No	No	Modal
		Test	>60sec	>15dB	No	No	No	No	Modal
	English	Model	>60sec	>15dB	No	No	No	No	Modal
		Test	>60sec	>15dB	No	No	No	No	Modal
11	Arabic	Model	>60sec	>15dB	No	No	No	No	Modal
		Test	>60sec	>15dB	No	No	No	No	Modal
	English	Model	>60sec	>15dB	No	No	No	No	Modal
		Test	>60sec	>15dB	No	No	No	No	Modal
12	Arabic	Model	>60sec	>15dB	No	No	No	No	Modal
		Test	>60sec	>15dB	No	No	No	No	Modal
	English	Model	>60sec	>15dB	No	No	No	No	Modal
		Test	>60sec	>15dB	No	No	No	No	Modal
13	Arabic	Model	>60sec	>15dB	No	No	No	No	Modal
		Test	>60sec	>15dB	No	No	No	No	Modal
	English	Model	>60sec	>15dB	No	No	No	No	Modal
		Test	>60sec	>15dB	No	No	No	Yes	Modal
14	Arabic	Model	>60sec	>15dB	No	No	No	No	Modal
		Test	>60sec	>15dB	No	No	No	No	Modal
	English	Model	>60sec	>15dB	No	No	No	Yes	Modal
		Test	>60sec	>15dB	No	No	No	Yes	Modal
15	Arabic	Model	>60sec	>15dB	No	No	No	No	Modal
		Test	>60sec	>15dB	No	No	No	No	Modal
	English	Model	>60sec	>15dB	No	No	No	No	Modal
		Test	>60sec	>15dB	No	No	No	No	Modal
16	Arabic	Model	>60sec	>15dB	No	No	No	No	Modal
		Test	>60sec	>15dB	No	No	No	No	Modal
	English	Model	>60sec	>15dB	No	No	No	Yes	Modal
		Test	>60sec	>15dB	No	No	No	Yes	Modal
17	Arabic	Model	>60sec	>15dB	No	No	No	No	Modal
		Test	>60sec	>15dB	No	No	No	No	Modal
	English	Model	>60sec	>15dB	No	No	No	No	Modal
		Test	>60sec	>15dB	No	No	No	Yes	Modal
18	Arabic	Model	>60sec	>15dB	No	No	No	No	Modal
		Test	>60sec	>15dB	No	No	No	No	Modal
	English	Model	>60sec	>15dB	No	No	No	Yes	Modal
		Test	>60sec	>15dB	No	No	No	No	Modal
19	Arabic	Model	>60sec	>15dB	No	No	No	No	Modal
		Test	>60sec	>15dB	No	No	No	No	Modal
	English	Model	>60sec	>15dB	No	No	No	No	Modal
		Test	>60sec	>15dB	No	No	No	Yes	Modal
20	Arabic	Model	>60sec	>15dB	No	No	No	No	Modal
		Test	>60sec	>15dB	No	No	No	No	Modal
	English	Model	>60sec	>15dB	No	No	No	Yes	Modal
		Test	>60sec	>15dB	No	No	No	Yes	Modal
Speaker #	Language	Sample	Net speech	S/R	Clipping	Babble	Reverberates	Different channels	Phonation

Results

The final results consisting on likelihood ratios are presented in Table 2. It was observed that the lower likelihood ratios are correlated with the acoustic problems from Table 1.

- Speaker #1: The speaker sample here had no problems in any of the audio samples, making speaker #1 a prime sample to work with. The LRs were all very high, with his English test being the highest. These numbers go along with our hypothesis really well.
- Speaker #2: The speaker sample here had a low SNR on the Arabic test, causing the LR to be very low. However, the English model and test were problem free, causing those LRs to both be very high and also in line with our hypothesis.
- Speaker #5: The English test was clipped by .56%, causing the LR to drop to a lower amount than would be accepted. But, the Arabic test was the highest LR of all of the Arabic test LRs and the English model was high as well.
- Speaker #7: The speaker sample had a harsh and high pitched voice in the English test, causing the LR to be quite low, while the Arabic test and English model were fine and got high LRs.
- Speaker #16: The transmission channels of the English model and test were different from the Arabic model and reference population, causing the English model LR to be the lowest of all of the English model LRs and the English test LR to be very low, too. The Arabic test, however, was fine and the LR was high.
- Speaker #19: The English test was muffled, causing the LR to drop to a significantly low number in comparison to the higher LRs that came from the Arabic test and English model.

Table 2: The Final Results

Speaker #	Arabic		English		LRs Arabic Model vs			Comments
	Model	Test	Model	Test	Arabic Test	English Model	English Test	
1	OK	OK	OK	OK	15108	16586	126103	OK
2	OK	SNR≈14dB	OK	OK	286	9216	65244	Noisy Arabic Test, the new Arabic Test is clipped
3	OK	OK	OK	OK	17807	1784	8640	OK
4	OK	OK	OK	OK	3183	790	1388	OK
5	OK	OK	OK	clipped	8631330	4583	449	OK, the English Test is clipped 0.56%
6	OK	OK	OK	channels	10542	2055	438	OK
7	OK	OK	OK	harsh voice	11577	7188	174	OK, Different phonation, higher pitch, harsh
8	OK	OK	channels	OK	3272	338	1308	OK
9	OK	OK	channels	channels	1613	137	92	Channel problems, OK for Arabic M v. T
10	OK	OK	OK	OK	510	2383	5246	OK
11	OK	OK	OK	OK	70027	3689	1714	OK
12	OK	OK	OK	OK	1626	956	631	OK
13	OK	OK	OK	channels	581	1083	422	OK, Different channels
14	OK	OK	channels	channels	1080	107	287	Different channels
15	OK	OK	OK	OK	60817	31049	64582	OK
16	OK	OK	channels	channels	851	9	285	Different channels
17	OK	OK	OK	channels	1104	924	402	OK, different channels, I keep the English Model
18	OK	OK	channels	OK	115	47	4401	OK, Different channels, I keep the English Test
19	OK	OK	OK	muffled	2380	2681	223	OK, we keep the English Model
20	OK	OK	channels	muffled	12177	31	0.1	Different channels + Muffled

Of the 60 test samples collected, which were Arabic model vs. Arabic test, Arabic model vs. English model, and Arabic model vs. English test, 43 passed criteria deemed appropriate for the experiment.

Intra/Between and Inter-Variability LRs

Arabic Model v. Arabic Test

In the Arabic model vs. Arabic test, the probability of identifying similarities between the samples was quite high and the highest likelihood ratio being recorded at 8,631,330. The lowest likelihood ratio is 115, without any problems. These figures represent the number of times that similar observations could be identified between voice samples of known origin and voice samples whose origin was unknown. The inter-variability high is 63. This means there is a gap between the intra and inter variabilities, between 115 and 63, which signifies that no sample from the reference population is the person speaking in the evidence. These numbers are displayed in figure 8. All numbers that are red represent the samples that had problems, for the Arabic model v. Arabic test in particular, there is one sample whose SNR was below 15 dB, as seen in figure 9. The red numbers were not accepted in the final result.

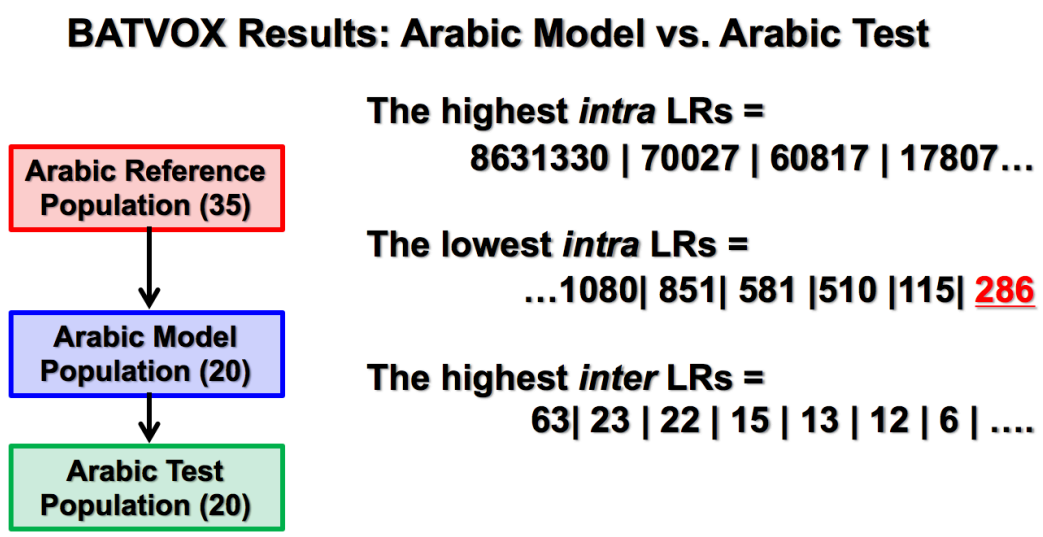


Figure 8: Arabic Model vs. Arabic Test Results

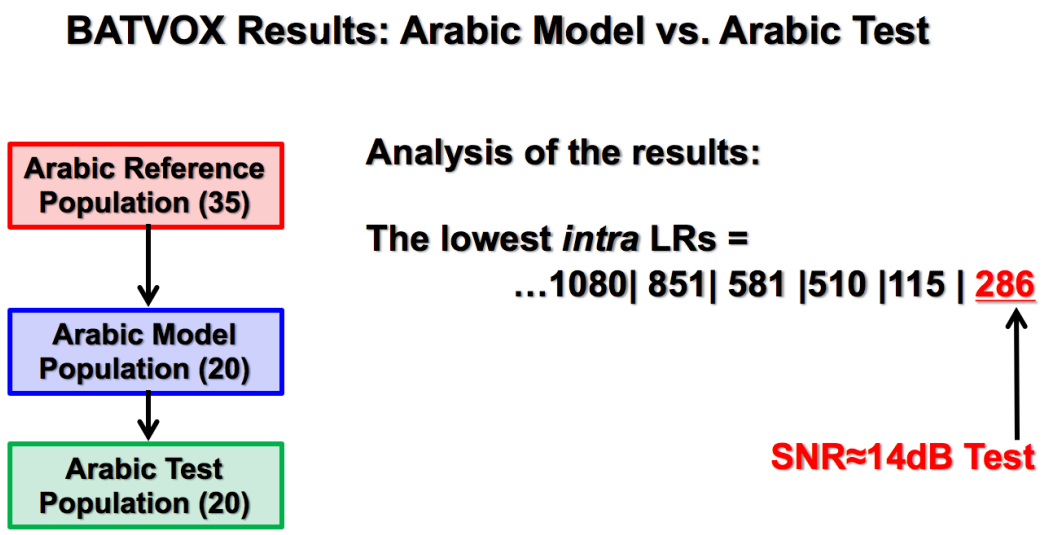


Figure 9: Arabic Model vs. Arabic Test Intravariability LR

Arabic Model vs. English Model

In the Arabic model vs. English model, the likelihood ratios were generally lower than Arabic model vs. Arabic test for each speaker. The highest ratio between Arabic model vs. English model being recorded at 31,049 and the lowest at 790. This indicates that the

probability of finding similarities between voice samples of known origin and those of unknown origin, from the two languages, had reduced. The highest inter-variability is 32, meaning that there is a gap between the intra and inter variabilities. These numbers are seen in figure 10 and 11.

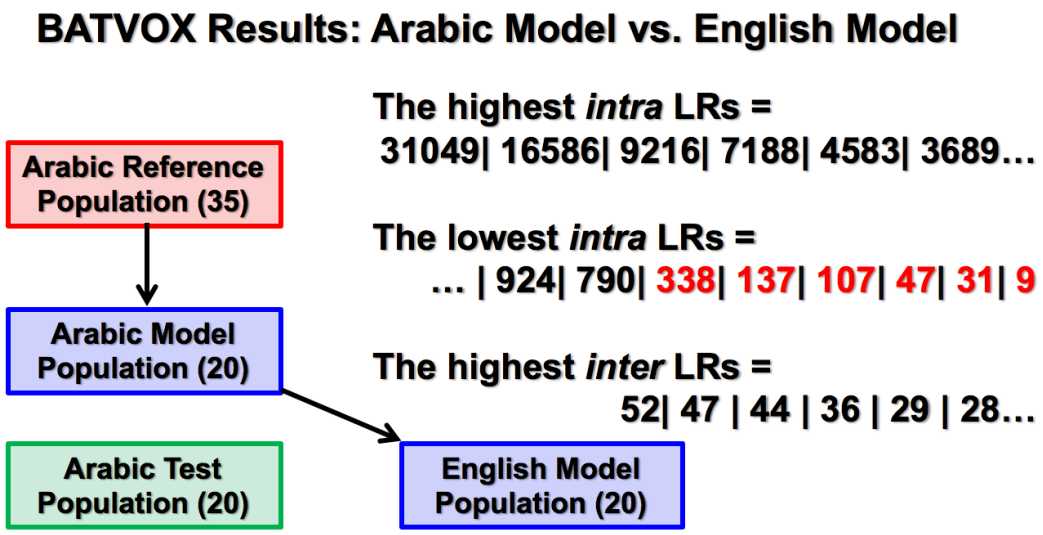


Figure 10: Arabic Model vs. English Model Result

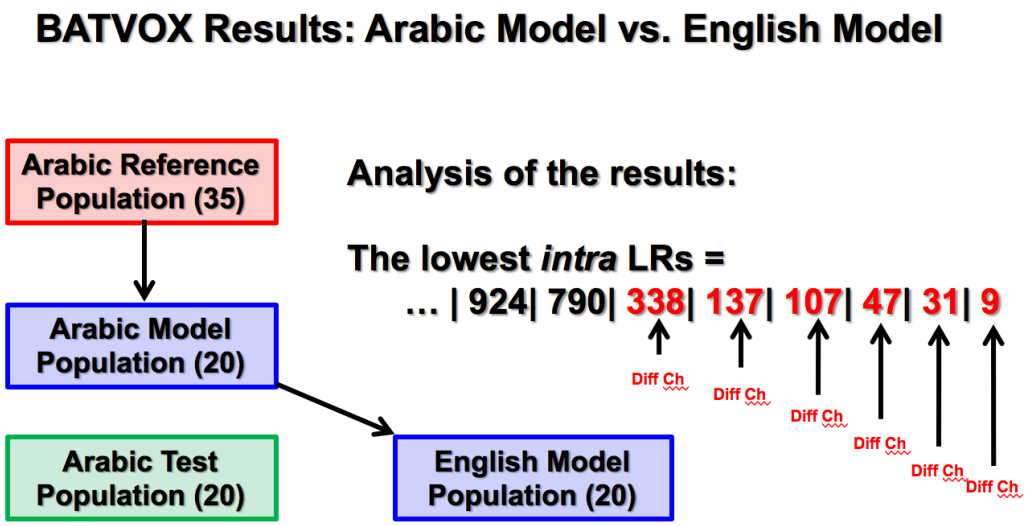


Figure 11: Arabic Model vs. English Model Intravariability LR

Arabic Model vs. English Test

As with English model samples, the Arabic model vs. English test likelihood ratios were generally lower than Arabic model vs. Arabic test for each speaker. The highest ratio being recorded at 126,103 and the lowest at 631. The highest inter variability is 98, meaning that there is a gap between the intra and inter variabilities. These numbers are all seen in figures 12 and 13.

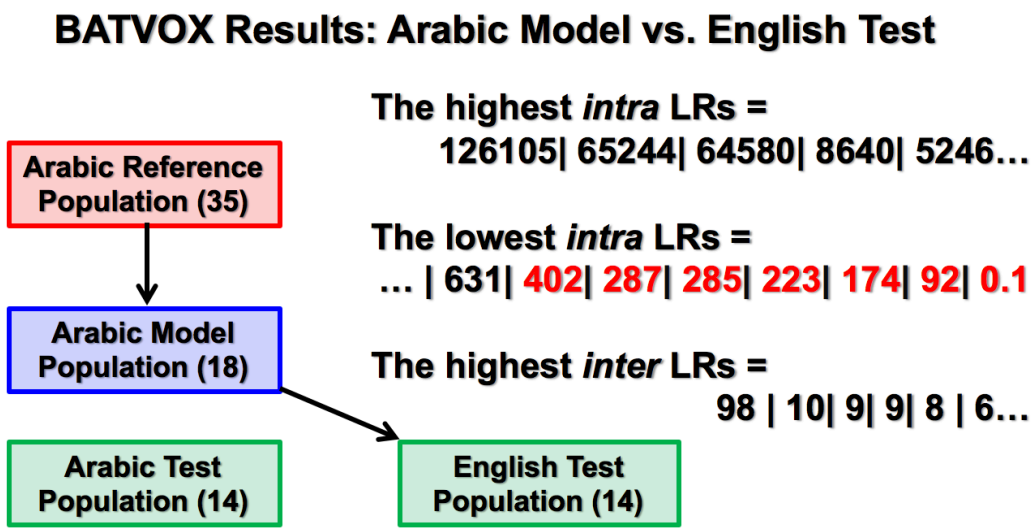


Figure 12: Arabic Model vs. English Test Results

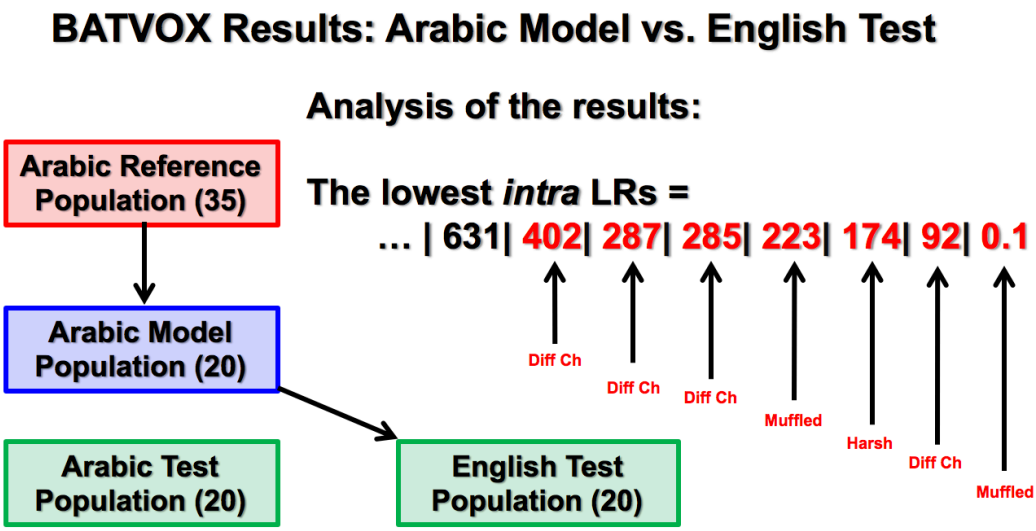


Figure 13: Arabic Model vs. English Test Intravariability LR

CHAPTER IV

CONCLUSION

Discussion Points

The quality and quantity of the voice samples is crucial. Low-quality samples contain plenty of distortions, noise and analyzing them is not easy. A quality sample, on the other hand, has few or no distortions, and this enhances the analyzing process. Quality samples result in more accurate and reliable LR results because the various voice characteristics are better identified by the tools conducting the analysis. The quantity of the voice samples is also important because reliable recognition processing requires an adequate length of each sample. Overall, the study would not be affected even if some samples had deficiencies because the researcher can clip out the unusable speech signals. This would work best if the samples all contained enough length to endure clipping. It is even harder to maintain accuracy for text-independent recognition processes in comparison to text-dependent when a sample has both low-quality and quantity due to no control over how the sample is made.

Noise is part of quality that stands out as a more common issue. It is present in almost all environments in different forms, and this makes it challenging to prevent it from occurring in samples, even controlled samples. It can be periodic or non-periodic depending on the nature of the environment where the recordings were made. For this thesis, hums, hisses, and broadband noises were detected and fixed as much as possible. However, some noises like harshness, muffled speaking and lossy compression artifacts are very difficult to fix, and remain a hit on the accuracy of text-independent speaker

recognition. Lossy compression artifacts refer to the distortion of media files like audio and images due to the application of lossy data compression. This compression involves discarding part of the data that is present in the media in an attempt to make it simplified enough for storage or transmission purposes. In audio files, it mainly works with the psychoacoustic models. Another common issue that arose in this experiment was the difference in transmission channels between samples of the same speaker. This happens when recordings are taken from networks like a public domain or from VOIP. The problem with different transmission channels is that the software is unable to properly pull the data from the waveforms when they emit different speech signals that don't correlate.

Evaluation of the Tested Hypothesis

The intra LRs were highest when the same language was being used, whereas they experienced a decline when comparisons were done across the two languages. On the other hand, the inter likelihood ratios were relatively low across the board, but the highest figures were recorded when comparisons were done between the two languages. As a result, the hypothesis being tested, whether or not Batvox can reliably perform comparisons with language independent samples, is accepted due to the percentage being in favor.

Practical Recommendations

The file recordings should be made in a different environment where the challenge of noise and other distortions will be limited. This would result in audio files that are of high quality and this would enhance the analysis process and the production of better results. Challenges that arise should also be detected and fixed. Others, such as the lossy compression artifacts, remain a difficult-to-solve problem and recording should work to eliminate the possibility of these challenges from appearing. Furthermore, the length of

samples should be long enough to ensure that any if contingencies were to arise, they would be dealt with without any impact on the sample and the sample would still be long enough to process. A large number of samples, as well, would enable the researcher to replace any samples that fail to meet the criteria for analysis and to increase the accuracy of the intra-variability LR. For instance, this thesis incurred a case where some male voices had a low gender confidence, and therefore, were eliminated from the analysis process because they did not meet the criteria. Nonetheless, a large number of samples collected allowed the option of replacing the inadequate samples with more fit samples. On top of that, a larger number of samples than originally planned for ensures that conclusions are more accurate and reliable because the samples will contain all the salient characteristics of the population and therefore, be a better representation of the population.

This thesis can be investigated more by changing the reference population to English and using Arabic second language speakers to see if the LRs are similar to the ones found here and if this creates a variable of similarity between the two languages. This topic can be continued by conducting further studies on how the problems of harshness, muffled speaking, different channels and lossy compression artifacts can be solved. The elimination of these challenges would result in higher-quality audio files that would provide more accurate and reliable results after being analyzed. Further studies can also be conducted using female voice samples to create more flexible systems that can be utilized by people from both genders. More studies can be conducted to compare the LRs of two different languages outside of English and Arabic using the same design as a way to see if the LRs of those studies compared to these results are trending or completely different. This would

be beneficial to further development of a system in which speaker recognition can be utilized using two different languages against a single language reference population.

BIBLIOGRAPHY

(1) Amino, Kanae, et al. "Historical and procedural overview of forensic speaker recognition as a science." *Forensic speaker recognition*. Springer New York, 2012. 3-20.

(2) Reynolds, Douglas. "An overview of automatic speaker recognition." *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)(S. 4072-4075)*. 2002.

(3) El-Samie, Fathi E. Abd. *Information security for automatic speaker identification*. Springer New York, 2011.

(4) Drygajlo, Andrzej. "Automatic Speaker Recognition for Forensic Case Assessment and Interpretation." *Forensic Speaker Recognition*. Springer New York, 2012. 21-39.

(5) Furui, Sadaoki. "50 years of progress in speech and speaker recognition." *SPECOM 2005, Patras* (2005): 1-9.

(6) Huang, Xuedong, et al. *Spoken language processing: A guide to theory, algorithm, and system development*. Prentice Hall PTR, 2001.

(7) Kelly, Finnian. "Automatic Recognition of Ageing Speakers."

(8) Temme, Steve. "Audio distortion measurements." *Application Note, Bruel & Kjar* (1992). http://www.bksv.com/doc/BO0385.pdf

(9) Gonzalez-Rodriguez, Joaquin. "Evaluating automatic speaker recognition systems: an overview of the nist speaker recognition evaluations (1996-2014)." *Loquens* 1.1 (2014): e007.

(10) Rhodes, Richard William."Assessing the strength of non contemporaneous forensic speech evidence." (2012).

(11) Agnitio. "*Product Data Sheet: BATVOX"*. Sept, 2015 from: http://www.agnitio-corp.com/sites/default/files/BATVOX_DS_51915_FINAL_Hires.pdf

(12) Agnito (2014). *BANOX in Keywords: 'Decision-support Software'*. Sept, 2015 from: http://pegasus.cl/descargas/BATVOX_Brochure_November_2014.pdf

(13) Koenig, Bruce E., Douglas S. Lacey, and Steven A. Killion. "Forensic enhancement of digital audio recordings." *Journal of the Audio Engineering Society* 55.5 (2007): 352-371.

(14) Künzel, Hermann J. "Automatic speaker recognition with crosslanguage speech material." *International Journal of Speech Language and the Law* 20.1 (2013): 21-44.

(15) Campbell, Joseph P., et al. *The MMSR bilingual and crosschannel corpora for speaker recognition research and evaluation*. MASSACHUSETTS INST OF TECH LEXINGTON LINCOLN LAB, 2004.

(16) Gold, Erica, and Peter French. "International practices in forensic speaker comparison." *International Journal of Speech Language and the Law* 18.2 (2011).

(17) Luengo, Iker, et al. "Text Independent Speaker Identification in Multilingual Environments." *LREC*. 2008.

(18) Morrison, Geoffrey Stewart. "Measuring the validity and reliability of forensic likelihood-ratio systems." *Science & Justice* 51.3 (2011): 91-98.

(19) Morrison, Geoffrey Stewart, and Hugh Selby. *Forensic voice comparison*. Thomson Reuters, 2010.

(20) Eriksson, Anders. "Aural/Acoustic vs. Automatic Methods in Forensic Phonetic Case Work." *Forensic Speaker Recognition*. Springer New York, 2012. 41-69.

(21) Decker, John F., and Joel Handler. "Voiceprint Identification Evidence--Out of the Frye Pan and Into Admissibility." *Am. UL Rev.* 26 (1976): 314.

(22) Alexander, Anil. *Forensic automatic speaker recognition using Bayesian interpretation and statistical compensation for mismatched conditions*. Diss. Institut de traitement des signaux SECTION DE GÉNIE ÉLECTRIQUE ET ÉLECTRONIQUE POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES PAR Bachelor of Technology in Computer Science and Engineering, Indian Institute of Technology, Madras, 2005.

(23) Doddington, George R., et al. "The NIST speaker recognition evaluation–overview, methodology, systems, results, perspective." *Speech Communication* 31.2 (2000): 225-254.

(24) Vaseghi, Saeed V. "Noise and distortion." *Advanced Digital Signal Processing and Noise Reduction, Second Edition. ISBNs: 0-471-62692-9 (Hardback): 0-4* (2000): 70-84162.

(25) Stadelmann, Thilo. *Voice Modeling Methods for Automatic Speaker Recognition*. Diss. Philipps-Universität Marburg, 2010.

(26) Künzel, Hermann J., and Paul Alexander. "Forensic Automatic Speaker Recognition with Degraded and Enhanced Speech." *Journal of the Audio Engineering Society* 62.4 (2014): 244-253