Minimizing Language Dependency in a Voice Authentication System

Literature Review

Ryan J. Mechery

The Massachusetts Academy of Math and Science

Worcester, MA

Author Note

This research was not supported by any funding or grants.

Correspondence regarding this paper should be addressed to rmechery@wpi.edu.

Table of Contents

Minimizing Language Dependency in a Voice Authentication System

Literature Review

Voice Authentication systems use voice biometrics to validate a user’s identity. These systems are simple and cost-effective to implement. However, voice authentication is not a reliable nor global method for validation because of language dependency.

Authentication

Methods of Authentication

Authentication is a method for the validation of one’s identity. There are three types of authentications: knowledge, ownership, and inherent based. Knowledge-based authentications involve traditional passphrases that are easy to remember but are prone to brute force[1] and dictionary attacks[2]. Ownership-based authentications involve a physical device such as smart cards or RFID chips. This method is more secure but impractical as an end-user would be locked out of a system if they did not have access to the integrated circuit (IC) chip, unlike a passphrase. Inherent-based authentication is the most secure method compared to former methods as it involves biometrics, which are measurements of unique characteristics of the human body (Barkadehi et al, 2018).

Types of Biometrics

Although biometrics are prone to forgery and replay attacks, they are secure because they cannot be forgotten or stolen from a user. There are two types of biometrics: physical and behavioral. Physical biometrics involve direct measurements that can be scanned from the body.

[1] Brute force attacks involve entering generated text to hack into a system.

[2] Dictionary attacks also involve entering generated text to hack into a system, but they are generated with words under the assumption that a majority of users will use words in their passwords to make them easier to remember.

Examples include facial, iris, and fingerprint biometrics. Behavioral biometrics involves measurements of a user's unique habits and examples include gait, signature, and namely, voice biometrics (Marinov & Skövde, 2003). Although each type of biometric is very different in both architecture and use case, they all have general advantages and drawbacks. The pros of physical biometrics are that they are widely used and implemented, are stored as direct measurements of the body, and do not need to be rescanned. However, physical biometrics are sensitive to injuries that disfigure the body, the process requires complex sensors that drive up costs, and most importantly, once a physical biometric is copied by an attacker, it is stolen forever and cannot be reset. Although behavioral biometrics are inconsistent as a user's behavior may change slightly over time, they are more difficult to copy compared to physical biometrics. (Jirik, 2021).

Voice Authentication

Need for Voice Authentication

In today's technologically advanced world, we rely, evermore, on third-party companies to keep our personal data safe. Although we may feel that this data is secure, it can be easily stolen by faulty and weak passwords. According to the 2016 Verizon Data Breach Investigations Report (DBIR), over 80% of data branches were due to weak and stolen passwords. Although most people worldwide have a smartphone model with some type of biometric for authentication, this fact does not apply to other devices such as laptops, computers, and tablets. Voice Authentication can be already implemented on many of these devices without the need for an upgrade[3].

[3] This may not be practical for *all* devices as voice authentication is dependent upon the quality of microphone. Microphones with poor frequency response, high levels of feedback, or plosive noise, can degrade the accuracy of a VA system (Antlion Audio, 2019).

Types of Voice Authentication

Speaker[4] Recognition (SR), or voice biometrics, is a vast field of study that aims to identify a person from their voice. This field has two main branches: Speaker Identification (SI), and Speaker Verification (SV). Speaker Identification is a one-to-many comparison that aims to identify a speaker, or person, from a list of registered speakers (Beigi, 2011). Example implementations include smart home assistants such as Google Home which can recognize different users from a spoken keyword. Speaker Verification, or voice authentication, is a one-to-one comparison that aims to verify the identity of a user. The key difference between these two branches is that speaker verification systems require an additional identifier, such as a password, to complete the validation process (Marinov & Skövde, 2003). This makes SV systems more ideal for authentication applications as they are less processing-intensive and more secure.

Architecture of Voice Authentication

Voice Authentication can be built for two types of systems: text-dependent and text-independent. Text-dependent systems are reliant on spoken text. Text-independent systems, on the other hand, can also use utterances as input, but these systems identify users based on *how* they speak. Although security and complexity are key differences, in terms of implementation, both systems follow the same general design, but text-independent systems require a training phase while this phase isn’t necessary for text-dependent systems.

Enrollment

The enrollment phase is the first step in verification and it is a one-time process that stores a user’s voiceprint or template (Sharma, 2020). There are three steps in the enrollment phase: audio capture, audio preprocessing, and feature extraction.

[4] The term *Speaker* will be used throughout this paper to identify the subject or user of a Voice Authentication system. This is not to be confused with a loudspeaker.

Audio Capture

Audio capture is important in the design of a VA system as the only information needed is an utterance while any other extraneous features are not needed. Two aspects of microphones are important, direction and sensitivity. Since microphones can either target sound at one or a variety of angles, they can either be unidirectional or omnidirectional in design. Although omnidirectional microphones are becoming increasingly popular in devices for picking up audio at any angle, they are prone to pick up extraneous noise (Beigi. 2011). Sensitivity is defined as the signal magnitude a microphone can pick up. Since the aim is to minimize background noise, low-sensitivity microphones are preferred as they can pick up relatively loud sounds without picking up too much background noise (Fox, 2021). If a certain type of microphone will only be used, then the design of the VA system can be altered toward that constraint. If a range of microphones will be used, which is typical for a biometric system, then that variation must be accounted for in the next step, audio preprocessing.

Audio Preprocessing

As described in the previous section, hardware features cause alterations in utterances, or speech, which can impact a computer's ability to compare two audio samples. There are many ways to process audio to preserve its necessary components but an important step is audio normalization. There are many types of normalization but in audio analysis, peak normalization is typically used. This process aims to scale the numbers in an audio dataset to ensure that the loudest sound does not go above exceed a range, typically -1 to 1 scale. This step ensures that the system can compare two audio samples even if a speaker speaks at a different volume, or different distances away from the microphone (Campbell, 2018).

Feature Extraction

Feature extraction is the process by which unique characteristics of an audio signal are isolated for further analysis. This is an important step in audio signal processing as even subtle variations make it a meaningless task for computers to compare two audio samples directly. There are many types of audio features, but they all follow a general hierarchy. Low-level features involve statistical features that only computers can distinguish and calculate, such as amplitude envelope, spectral centroid, spectral flux, zero-crossing rate, etc. Mid-level features involve features humans can make out such as pitch, beat, and Mel-frequency Cepstral Coefficients (MFCCs). High-level features involve musical elements such as harmony, rhythm, key, etc. (Mahanta & Padmanabhan, 2021). An audio analysis system can extract any one of these features, however, mid-level features are most typically used for speaker verification systems because they aren't abstract, high-level ideas, which are difficult to define, and they also give information about entire segments of audio data, unlike low-level features.

Mel-Frequency Cepstral Coefficients. MFCCs are the most common feature used in audio and music analysis. They measure the timbre, or vocal quality, of the human voice, and its coefficients are derived from the Mel-scale which is logarithmic in rate (Velardo, 2020). The function for converting from a frequency to the Mel scale is defined below:

$$Mel(f) = 2595 * \log_{10}\left(1 + \frac{f}{100}\right)$$

To better understand the importance of the logarithmic scale, although a 2000hz tone is four times greater in magnitude compared to a 500hz tone, humans only *perceive* the note's pitch to be around 1.7 times greater in frequency. MFCC derivation is quite complex, but at its core, in a process known as Fourier Transformation, a signal is decomposed into collection of sine waves to convert a time-domain signal into a frequency domain signal which is a function that measures the magnitude of different frequencies at a certain time (Tiwari, 2010). A Mel-frequency spectrogram, which is a three-dimensional graph[5], *can be constructed by combining both of these functions.*

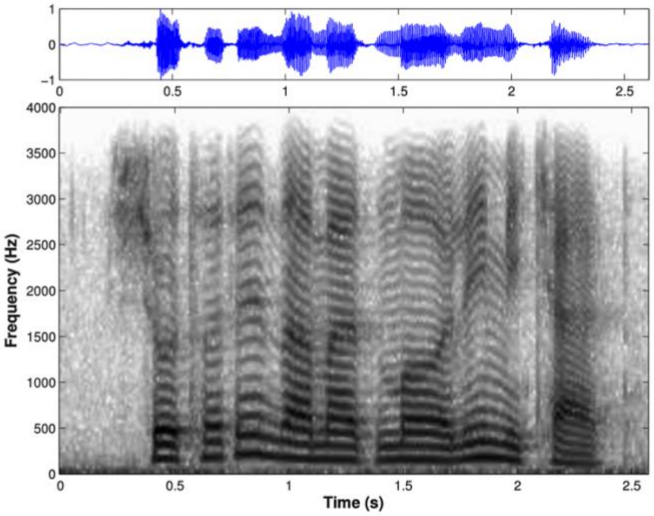


Figure 1. An example of a waveform (top) and a Mel spectrogram (bottom). (Beigi, 2011).

A spectrogram, which is shown at the bottom of Figure 1, is important as it visually conveys implicit features of audio that cannot be seen with a simple waveform. Furthermore, it allows humans to understand the patterns and trends a computer is aiming to recognize. Although most systems don’t actually store these graphs, the MFCCs that make up Mel spectrograms are stored in a two-dimensional feature vector or array. This step serves two main functions, it compresses audio data, but it also allows for computers to analyze this data in a model and store it for future reference.

Modeling

Modeling is the process by which a feature vector is analyzed or stored into a model/voiceprint that can be directly compared for similarity. Although two feature vectors can

[5] Although spectrograms have three dimensions, time, frequency, and magnitude, they are most often represented in two dimensional graphs with colors maps used to flatten the image. The colors maps visually convey the magnitude, or signal strength of each frequency.

be directly compared, this step saves on processing time as a model is simply a mathematical function[6]. To understand the significance of this compression, if 13 MFCC features are extracted from a one-minute audio sample framed at 20ms, the system would create a feature vector with just under 40,000 coefficients. Although this comparison of values would be quick[7], it would not be accurate as a mean or median is not a good indicator of similarity between two audio samples to determine if they come from the same speaker. Rather, comparing the distribution of MFCC coefficients is more important than comparing the individual values itself.

There are two main methods of modeling: statistical and machine learning. Statistical models are usually based on Probability Distribution Functions (PDF) which are less processing intensive to calculate but generally have a lower rate of authentication. Machine Learning based modeling is very processing intensive in the enrollment stage but improves the accuracy of authentication because it can run many epochs.

Gaussian Mixture Modeling. Gaussian or bell curves are a type of PDF that show the normal distribution of dataset. Speaker Verification systems usually create one model per user, but in applications outside of authentication or for storage reasons, some developers may opt to combine data from hundreds of speakers into one Universal Background Model (UBM). Gaussian Mixture Modeling (GMM) is perfect for this scenario as it combines different mixtures into one general distribution (Stadelmann, 2010). In an iterative process called Expectation Maximization (EM), a GMM system run different weights to find the best fit (Reynolds, 2009).

[6] Most systems do not actually store the model as simply a mathematical function. Rather, models are stored in a vector of numbers that make up the mathematical function.

[7] Attached in Appendix B is a python program written to compare two arrays of equal sizes, similar to the hypothetical description previously stated. Results of the program show that comparison of values, on an M1 Macbook Air, only takes 17.42 ms.

Validation

In the second phase, a system will use stored models to validate the identity of a user. At validation, the system will capture audio, and create a model of the claimed user to be compared with a stored model.

Similarity Measure

In statistical modeling-based systems, Likelihood Ratio Tests (LRTs) are used to determine the similarity of two audio samples. GMM or Hidden Markov Model (HMM) based systems, use a Log-Likelihood Ratio test which is calculated as follows,

$$\Lambda(X) = \log_{10}\left(X \middle| \lambda_{hyp}\right) - \log_{10}(X|\lambda_{\overline{hyp}})$$

This function simply calculates the probability that a claimed model is the same as the stored model and the number returned by it should be as close to zero as possible[8]. Due to variations in utterances, the log-likelihood may not exactly be zero so a developer has to determine a threshold value that would allow a user to be authenticated (Reynolds, 2000).

Language Dependency

Although text independent speaker verification systems are not restricted by what type of utterance a speaker provides, they are still language dependent to some degree. Language dependency is defined as a system’s need to be enrolled and validated in the same language.

Phonology

Phonology is the study of sounds in different languages. The basic unit of this field is called a phoneme which are semantically significant sounds that can change the meaning of a word (Ratner & Gleason, 2004). A study conducted by Maddieson in 1984 with 317 world languages found that although 70% of languages have between 20 and 37 phonemes, languages

[8] The Log LRT should be as close to zero as possible, only if the claimed speaker is a registered user.

can have as low as 11 and as high as 141 phonemes as seen in Native Hawaiian and Khoisan languages respectively. Although some languages may have the same number of phonemes, this is not an indication of similarity because they could have entirely different phones. Phones are defined as the different ways to pronounce a phoneme and their different pronunciations may not change the meaning of a word. In addition to phonemic differences, languages can vary in suprasegmental features of speech. Suprasegmental features are variations in pronunciation that occur beyond phonemes. Examples include prosodic features or intonations which include stressing syllables in the pronunciation of words in a phrase to convey some semantic significance. Pitch variation utilizations varies upon the language with Indo-European languages using it sparsely compared to East Asian languages like Mandarin.

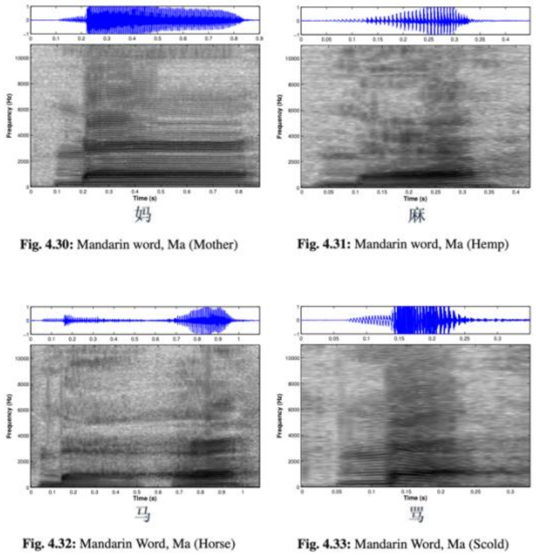


Fig. 4.30: Mandarin word, Ma (Mother)

Fig. 4.31: Mandarin word, Ma (Hemp)

Fig. 4.32: Mandarin Word, Ma (Horse)

Fig. 4.33: Mandarin Word, Ma (Scold)

Figure 2. Spectrograms (right) of a Chinese Speaker intonating the word “Ma” to convey different meanings. (Beigi, 2011).

As can be seen in Figure 2, although the pronunciation of a word may sound similar and contain the same types of phonemes, their pronunciations provide entirely different sound patterns.

Language Dependency in Speaker Verification Systems

Due to the phonetic variations of different languages, a speaker verification system, by nature, will provide different accuracies of authentication. A possible reason for this phenomenon is that more phonetically complex languages need a lot more enrollment audio for model creation and training than is practical for an authentication system. Some designs that account for this problem have been written about in academia, however this topic has not been fully researched.

Auckenthaler et. Al created a system for voice authentication in 2001 and tested it with twelve world languages from various families: Indo-European (6), Semitic (1), Sino-Tibetan (1), Altaic (2), Dravidian (1), and Austroasiatic (1). Overall, the researchers found that models trained with Chinese and Vietnamese obtained a 15% and 20% Equal Error Rate (EER), respectively, while models trained with English performed at around a 12% EER. While these results support the idea that more phonetically complex languages will have lower rates of authentication in SV systems, the experiment was not fully investigated. Although the selection of languages seems diverse, exactly half belong to the same language family. Second, the researchers admitted that they did not have a proper audio database of utterances at the time of testing. This means that the low accuracy rates could be due to other factors such as poor microphone quality and sampling errors.

Conclusion

Text Dependent SV systems have many different components and aspects that all impact accuracy of authentication. Many issues with SV systems have been identified and can be fixed with hardware and software revisions. However, many developers have failed to mention nor fix language dependency. Some researchers have developed experimental setups that try to account for this, but constraints of the time led to an insufficient database of recordings to train the system with, and a failure to choose an optimal architecture. Even if the experimental systems are shown to work with language dependency, they do not account for real world variables such as poor microphone quality, background noise, and processing power which are all important aspects of a universal biometric authentication system.

References

Antlion Audio. (2019). *How do you know if your mic is bad?* Retrieved from https://antlionaudio.com/blogs/news/how-do-you-know-if-your-mic-is-bad.

Auckenthaler, R., Carey, M. J., & Mason, J. S. (2001, May). Language dependency in text-independent speaker verification. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings* (Cat. No. 01CH37221) (Vol. 1, pp. 441-444). IEEE.

Fox, A. (2021). *What is microphone sensitivity? an in-depth description*. My New Microphone. Retrieved from https://mynewmicrophone.com/microphone-sensitivity/.

Barkadehi, M. H., Nilashi, M., Ibrahim, O., Zakeri Fardi, A., & Samad, S. (2018). Authentication systems: A literature review and classification. https://doi.org/10.1016/j.tele.2018.03.018

Beigi H. (2011) Speaker Recognition. In: Fundamentals of Speaker Recognition. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-77592-0_17

Campbell, T. (2018). *Audio Normalization: What Is It and Why Do We Do It?* Medium. https://medium.com/tannerhelps/audio-normalization-what-is-it-and-why-do-we-do-it-37b63176b914

Jirik, P. (2021). *5 Popular Types of Biometric Authentication: Pros and Cons*. PHONEXIA Speech Technologies. https://www.phonexia.com/en/blog/5-popular-types-of-biometric-authentication-pros-and-cons/

Maddieson, I. (1984). *Patterns of sounds*. Cambridge University Press.

Mahanta, S. K., & Padmanabhan, A. (2021). Audio Feature Extraction. Devopedia. Retrieved November 15, 2021, from https://devopedia.org/audio-feature-extraction.

Marinov, S., & Skövde, H. I. (2003). *Text Dependent and Text Independent Speaker Verification Systems. Technology and Applications*.

Ratner, N. B., & Gleason, J. B. (2004). Psycholinguistics. In L. R. Squire (Ed.), *Encyclopedia of Neuroscience* (pp. 1199–1204). Academic Press. https://doi.org/10.1016/B978-008045046-9.01893-3

Reynolds, D. A. (2009). Gaussian mixture models. *Encyclopedia of biometrics*, 741, 659-663.

Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using Adapted Gaussian mixture models. *Digital Signal Processing*, 2000.

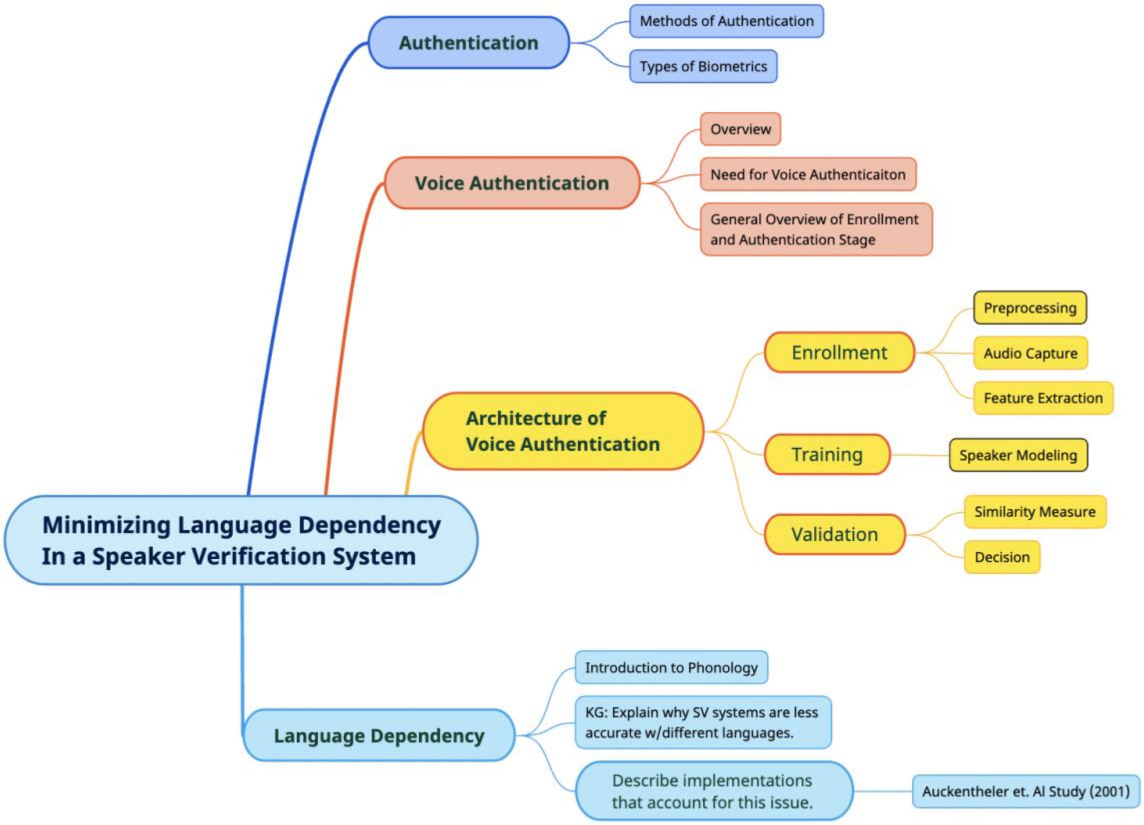
Stadelmann, T. (2010). *Voice Modeling Methods for Automatic Speaker Recognition*.

Tiwari, V. (2009). MFCC and its applications in speaker recognition. *International Journal on Emerging Technologies*.

Velardo, V (2020). *Types of Audio Features for Machine Learning* [Video]. YouTube. https://youtu.be/ZZ9u1vUtcIA

Appendix A

Literature Review Mindomo


Minimizing Language Dependency In a Speaker Verification System
Authentication
Methods of Authentication
Types of Biometrics
Voice Authentication
Overview
Need for Voice Authenticaiton
General Overview of Enrollment and Authentication Stage
Architecture of Voice Authentication
Enrollment
Preprocessing
Audio Capture
Feature Extraction
Training
Speaker Modeling
Validation
Similarity Measure
Decision
Language Dependency
Introduction to Phonology
KG: Explain why SV systems are less accurate w/different languages.
Describe implementations that account for this issue.
Auckentheler et. Al Study (2001)

Appendix B

Python Program that Compares Hypothetical MFCC Feature Vectors

```
import time #used to track processing time
import numpy as np
start_time = time.time() # tracks time

array1 = np.random.randn(13,3230) #array1 with 40,000 values
array2 = np.random.randn(13,3230) #array2 two has the same amount of
values as array1

differences = [] #this will store the differences between both arrays
for each value

#the nested for loop is used to iterate through the 2d arrays
for r in range(len(array1)):
    for c in range(len(array1[0])):
        subtract = array1[r][c] - array2[r][c]
        differences.append(subtract)

print("Execution Time:--- %s seconds ---\n" % (time.time() -
start_time))
```