Project Proposal

Project Title: Minimizing Language Dependency in a Speaker Verification System

Author: Ryan Mechery

Date: November 1st, 2021

Project Definition:

Text Independent Speaker Verification (TISV) systems are the most accurate implementation of voice authentication for English-based languages. However, these systems offer variable accuracies when tested with other world languages. The objective of this project is to create a system for Text Independent Speaker Verification and train it with different subjects who speak languages from linguistic families including Indo-European, Dravidian, and Sino-Tibetan. The overall aim of this project is to improve the design of the verification system in order to improve authentication accuracy. I expect the results to show that, after multiple rounds of testing, the system will return an equal, if not greater, authentication accuracy compared to the control model trained only with English speaking participants. This project will only train the system with audio samples of utterances, or spoken language, and unintelligible speech will not be used to train the system.

Background:

Phrase 1: Text Independent Speaker Verification (TISV) systems offer variable accuracies compared to English-trained models when tested with other world languages.
Phrase 2: The goal of this project is to engineer an SV system that offers comparable accuracy to English-trained systems.

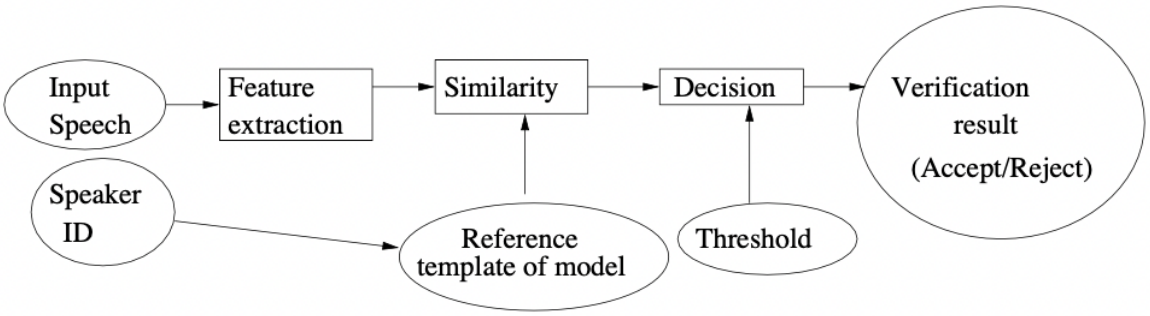
Voice Biometrics

Authentication is a method for the validation of one's identity. There are three main methods for authentication and they are confidential phrases that one can remember such as passwords, physical objects that only one can have such as smart cards, and the last one is biometrics which are just measurements of the unique characteristics of the human body (Marinov & Skövde, 2003). There are two types of biometrics: physical and behavioral. Physical biometrics are direct measurements of an individual's physical features. In contrast, behavioral biometrics are measurements of an individual's habits. Examples of unique behaviors that can be measured include signature, gait, and namely, voice biometrics. A key difference between these two biometrics is that physical biometrics can be directly measured while behavioral biometrics require pattern recognition. For some context, direct measurements of the minutiae or ridges on a finger can be compared for validation (Biometrics Institute). However, in the case of voice biometrics, two audio samples cannot simply be compared for validation. This is because natural and external factors cause a person's voice to change throughout the day. In order to compare two audio samples together, it has to find and extract implicit features within the audio which make it unique.

Speaker Verification (SV)

Voice Authentication is a type of behavioral biometric authentication that validates a user with their voice. Speaker Verification (SV) is another term for voice authentication and it is defined as a one-to-one comparison that aims to validate a user (Marinov & Skövde, 2003). In contrast to Speaker Identification (SI) which can identify a speaker from a list of registered users only using their voice, most Speaker Verification systems require an additional passphrase or ID number before it can complete the validation process (Beigi, 2011). Although SV systems are more processing intensive as they need to analyze voice data, separately, for each individual user, SV systems are more secure as they use two-factors for authentication: a passphrase and biometrics.

Speaker Verification systems are implemented in one of two categories: Text-Dependent and Text-Independent. Text-Dependent Speaker Verification (TDSV) systems require utterances, which are audio samples of spoken language, that are the same in both enrollment and validation. Although they are simple to implement, TDSV systems have an inherent security flaw in that all a hacker needs to be granted access by the system is an audio recording of the registered user. TISV systems, in contrast, focus on *how* someone is speaking. These systems extract unique vocal features that remain constant in a registered user's voice. As such, TISV systems aren't limited by predetermined utterances. However, they will encounter the same issue as TDSV systems if a hacker can spoof the system. So, TISV systems usually require the user to say a random prompt at validation and will use Speech to Text conversion to ensure that the user is saying the correct prompt. To hack a system like this, a hacker would need to create a generation system to recreate a registered user's voice (Marinov & Skövde, 2003).



This figure is a flowchart that illustrates the general architecture of a Speaker Verification system (Marinov & Skövde, 2003).

Language Dependency

Although Speaker Verification systems are language-independent in their design –meaning that the language of the speaker shouldn't affect the accuracy of the system – results have shown that the language of the speaker impacts the accuracy of the system. The reason for this lies in the structure of language itself. Since letters can be pronounced in different ways in each language, there are generally more unique sounds than letters of the alphabet. Unique sounds that can change the meaning of a word are called phonemes and their count varies between different languages. For some perspective, English has 40 phonemes but languages can have anywhere from 13 phonemes, as seen in the Hawaiian language, all the way up to 141 phonemes as seen in !Kung languages (Beigi, 2011). In a study conducted by Auckenthaler et al. which observed the accuracies of an SV system trained with various world

languages, results showed that languages with higher phoneme counts had a lower authentication accuracy compared to languages with lower phoneme counts. The reasoning behind this phenomenon is that Speaker Verification systems can only capture so much audio data from a user. If a user speaks in languages with a higher phoneme count, it is highly probable that when it comes to be authenticated into the system, the user will present the system with more phonemes than what it was trained with.

Experimental Design/Research Plan Goals:

Independent Variable: Language used to Train Model

Dependent Variable: Accuracy of Authentication for each User

Standardized Variables:

- The architecture used to train the models.
- I plan to use Gaussian Mixture Modelling (GMM) for now but it could change.
- The enrollment and validation utterance length.
- The microphone used will remain the same throughout all trials.
- Quiet Environment – I will not test in locations with background noise.

Control Groups:

- A model trained with English-speaking speaker(s).

Materials List:

- Computer
- Jupyter Notebooks
- Python
    - Librosa - Audio and Music analysis package.
- Audacity
- External Microphone - FIFINE Metal Condenser Recording Microphone K669B
- Google Translate

Procedure:

I will create a Speaker Verification system using Gaussian Mixture Modeling in Python. I will make the program so that I can record voice data and save models using either a command line interface (CLI) or graphical user interface (GUI). I will make a separate program that can read all the stored models in my enrollment directory and when given validation audio files, can return the validation result, acceptance and rejection, and other statistics (see Data Analysis). I will create a testing program that can test the second program, with a large number of registered speakers, at random, and it will return the accuracy of each model. I will add audio files of non-registered speakers, or imposters, from at least four people in the system to be used for training purposes.

In trial one of testing, I will record 40 seconds of enrollment audio data from at least two different speakers, preferably male and female, in these different language/families: English, Indo-European, Dravidian, and Sino-Tibetan. I will narrow down specific languages to use depending upon availability, but I will keep these four languages/ families in mind. I will use a

random sentence generator, such as the one found at https://randomwordgenerator.com/sentence.php, along with Google Translate, to provide 40 seconds of randomly generated validation utterances (40 s / 10s validation = 4 test cases) in the target language. I will take the enrollment data and create a model for each speaker which should be a total of 8 models and I will save that in the enrollment directory. I will use the testing program to complete 100 tests (could change based on processing power) with random users.

In second and subsequent trials of testing, I will decrease the enrollment and testing data and observe the accuracies of the system. Depending on the results, I will make adjustments in either the design of the system as a whole, or specific subsets of the design that can be changed to improve accuracy.

Risk/Safety Concerns:

The only concern associated with this project is that it involves human participants. Before recording, I will ask participants for consent to record their voices. To ensure the privacy of their voice recordings, I will not include the participants' names in any audio file named or any code publicly shared.

Data Analysis:

Authentication systems can only give a Boolean result of acceptance or rejection. However, the authentication system can be wrong so there are four outcomes of the SV system. The accuracy rate of each outcome is as follows: True Acceptance Rate (TAR), True Reject Rate (TRR), False Acceptance Rate (FAR), and False Reject Rate (FRR). I will graph the FAR and FRR in a Detection Error Tradeoff (DET) plot for each model and use that to compare different models. In addition to a DET plot, I will list the TAR, TRE, FAR, and FRE in a 4x4 confusion matrix for each model. I will record waveforms and spectrograms for all audio recordings and if needed, highlight some for clarification or further analysis in my theses.

Potential Roadblocks: (with action steps identified of how you might solve these):

1) **Design Limitations**: For this project I plan to use Gaussian Mixture Modelling in designing my Speaker Verification system. I have created an initial prototype which I trained with a small dataset and it can be found at this GitHub repository. This is a fairly simplistic approach that works with around 93% accuracy but multiple rounds of testing has shown false predictions. If preliminary results in experimentation with different languages shows that this type of approach cannot be reliably used for authentication, I will change the architecture of the SV system to use i-vector based approaches or Neural Networks.
2) **Dataset Size:** I plan to ask volunteers to help aid in providing audio data for my project. Since I will be testing the accuracy of different languages, I need multiple people that can speak each language. If I cannot gather enough participants, I will use a speech corpus, or audio dataset, online, such as the one provided by Mozilla. Since each speaker in this dataset wasn't recorded with the same microphone, adjustments will have to be made to the system to account for this.

References:

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

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

Marinov, S. (2003). Text dependent and text independent speaker verification system: Technology and application. *Overview article*.

Sharma, H. (2020, January 17). *Biometric System Architecture*. GeeksforGeeks. Retrieved November 1, 2021, from https://www.geeksforgeeks.org/biometric-system-architecture/.

Types of biometrics. Biometrics Institute. (2018, December 14). Retrieved November 1, 2021, from https://www.biometricsinstitute.org/what-is-biometrics/types-of-biometrics/.

Timeline: STEM Timeline