SPEAKER RECOGNITION SYSTEM

Caio Cruvinel - 1001079735

Daniel Seerig - 1001880734

Felipe Chicarelli - 1001006118

Pedro Soares - 1001080028

Rodrigo Abreu - 1001887915

University of Toronto

ECE446 - Sensory Communications

Professor Willy Wong

1 Abstract

The purpose of the project was creating a speaker recognition system capable to identify and verify an user using the voice, which could be used to control the access to some places or documents. The first phase of the project was to build a database with some users. The second phase consisted in develop a system that could find who in that database was closest to the signal input, and what was the probability that signal really belong to that person. Finally we have to test the project using for input some voice that belong and some that not belong to the database.

2 Introduction

2.1 Background

Nowadays we have access of all our bank accounts and personal information in our computer; in consequence, we need to look for a better way to protect this kind of information than the usual password which has been demonstrated to not be safe. Researchers started to look for a new secure protection method that was possible to rely on, and one way found was to use voice characteristics in order to identify an user using a voice recognition system.

The speaker recognition process is consisted of three major steps. First, a way to compare voices is needed, and it is done extracting features from an input signal. There are numerous algorithms that extract these features, and the most well know is the one that retrieve Mel-Frequency Cepstral Coefficients (MFCC). With these coefficients in hand, the next step is train models that represent speakers. It was usually done using statistical models, but nowadays Artificial Neural Networks are being used to create more robust and reliable models. The last step is composed of two tasks: identification and verification.

Both tasks use an input signal and compare it with the models that were already trained; however, they are two different processes with distinct objectives. The identification algorithm aims to find what is the most similar model in a database

compared with the input signal. On the other hand, a good verification algorithm will evaluate if the input speech really is the claimed speaker of the database or is someone else.

2.2 Literature

There is a numerous quantity of papers about Speaker Recognition. Usually, each paper focus on a single step of the whole process; however, papers giving an overview of the situation exist and are a good start point of study in this field. In their paper [1], Kinnunnen and Li restate the three steps mentioned before, and give options for each of these steps.

Regarding to the feature extraction, many algorithms were compared [2] and it was concluded that MFCCs is best used with a text-independent context (i.e. the features does not depend on the text that is being said) and a noise free environment. It is also faster and easier to implement than other features. There are many ways to implement this extraction and one of the most well known way is how Muda showed in her paper [3].

A really important researcher in the field of Automatic Speech Recognition (ASR), Douglas Reynolds, developed a statistical model called Gaussian Mixture Model [4] and used it to perform the identification and verification processes in a speaker recognition system [5].

2.3 Rationale

Exploring sound waves can lead us to greater understanding on how much information can we obtain and use from it on our routine. If one takes into consideration that a blind person is capable of determining many aspects of his surrounding by the sound, which are crucial for them in a daily basis, how good can we analyze sound in order to acquire knowledge and how can we use this knowledge to model systems that can and are being used to improve our life.

By considering the increasing number of automated systems that work based on activation via human responses such as fingerprint, eye's retina and brain waves, the project aimed to explore this facet on those systems by using the voice as a simple, non-intrusive response that can be analyzed and taken into consideration when designing an automated system.

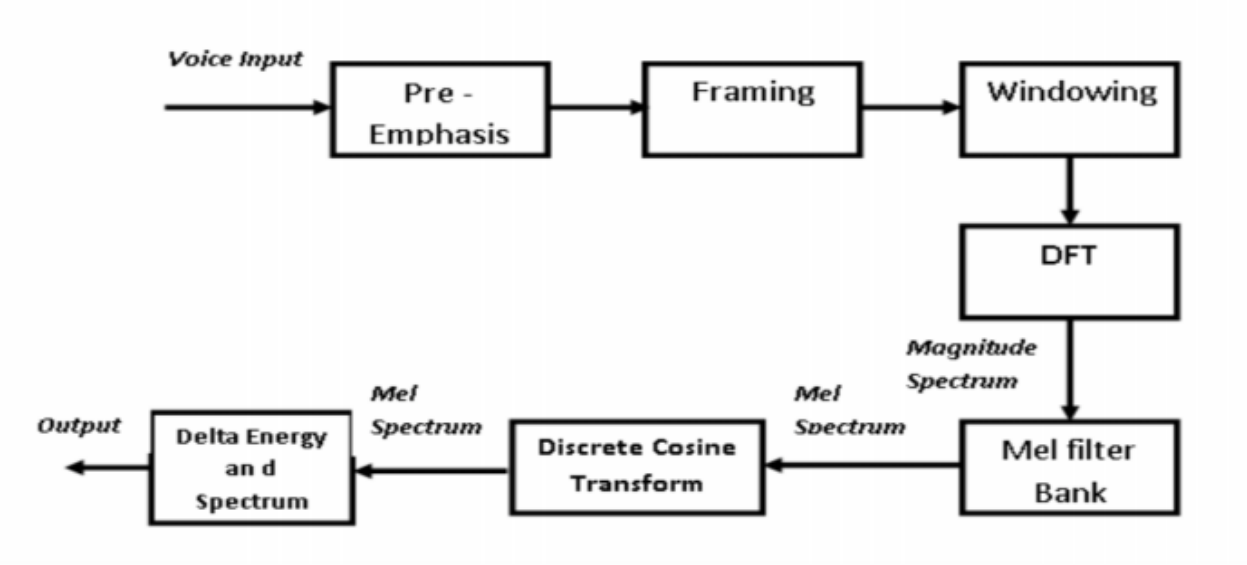
3 Objective

Develop a secure and reliable software that is able to identify if a voice belongs to someone in the database, and whose voice it is. The system should also be text-independent.

4 Methodology

The first step was to take some decisions of certain characteristics of the system. The first big decision was between a text-dependent or text-independent algorithm to train the models. The text-independent system was chosen because of more flexibility in the usage. The best model found for a text-independent system was the Gaussian Mixture Model; on the other hand, algorithms like Dynamic Time Warping (DTW) [3] were ignored for being useful only in a text-dependent context.

The first step to use the GMM is to extract the Mel Frequency Cepstral Coefficients, in order to do that we can separate the process in some steps that can be seen in the image.



1. The first step is to pass the signal in the pre-emphasis filter, that have the purpose of emphasis the higher frequency of the signal that is suppressed during the act of speech, that way we will have more data to differentiated every person.
2. We now have to sample the speech, dividing the signal in frames of 25 ms with overlap of 10 ms. This give us a good trade-off between time and frequency resolution.
3. Then the frames get windowed to smoothes a possible transiction between phonemes forcing the signal to be periodic. We had used the Hamming Window.
4. After we passe the signal to the frequency domain by the Fast Fourier Transform.
5. With the signal in the frequency domain we need to translate between Hertz to Mel Scale of frequency, we do that because the Mel scale is linear for the lower frequencies and this help the extraction to be more accurate since the human ear identify more precisely the change of tones in the lower frequencies. Then, we need to transform the signal from the frequency domain to cepstrum domain, in order to do that we pass the signal to a bank of triangular filter.
6. The bank of filters is composed by 26 triangular filters that covers all the spectrum of the input signal (0 to Fs/2), for each filter the signal will have a coefficient as

a response to the filter. The filter computes the spectral mean around the central frequency with increasing bandwidth and the filters are given by:

$$H'_m[k] = \begin{cases} 0, & \mathrm{k} < f_{mel}[m-1] \\ \frac{f_{hz}[k]-f_{mel}[m-1]}{f_{mel}[m]-f_{mel}[m-1]}, & \mathrm{f}_{mel}[m-1] \leqslant k \leqslant f_{mel}[m] \\ \frac{f_{mel}[m+1]-f_{hz}[k]}{f_{mel}[m+1]-f_{mel}[m]}, & \mathrm{f}_{mel}[m] \leqslant k \leqslant f_{mel}[m+1] \\ 0, & \mathrm{k} > f_{mel}[m+1] \end{cases} \quad (4)$$

Also, the matrix of the filters is absolutely summable and equal to 1.

7. After the transformation from the frequency to cepstrum domain, we take the energy of each output of the filter we do this by summing the energy of each frame response to a particular filter. Then we took the Discrete Cosine Transform of the log-energy, and then we have the 26 MFCC coefficients. On the literature it is widely used only 13 of that coefficients in order to characterize the human voice, they are more than sufficient for that. For the purpose of the project we only took the last 13 coefficients.

Now that the system has features to work with, it will model each speaker in the database with one GMM. The system create a model with 32 different gaussians and train each one with the iterative algorithm called Expectation-Maximization (EM) [6]. This algorithm has as input the coefficients that the system retrieved before, and gives as outputs the parameters that describe the corresponding GMM.

After modeling the users the system is ready to be used. Supposing that the database is already created. it is only necessary an input test (speech) that will be identified and verified. The program will get this input, extract the MFCCs and use them along with the Maximum Likelihood Estimation (ML) [7] to find the probability of this test to belong to each user in the database. This method will give a score and the user that had the bigger one is most likely to be the one speaking in the test speech.

With the probability in hands is still necessary to decide if that input belong to the most likely user in the database or if it belongs to someone not registered. To do that , a

threshold must be established, which means if the most likely user in database has score bigger than the threshold, it probably belongs to that person, otherwise it is from a non-registered user. There is no exact value in the literature, so we ran multiple tests with member of the database and non registers. We found out that in 90% of the tests, registered members had score bigger than 400 and unregistered less de 300, therefore we choose the value of 350.

The entire system was coded in Python. The choice was made because it is an Open Source language and has many libraries that support and implement the structures and algorithms needed. The system relies in the following libraries: NumPy, SciPy and Scikit-learn.

In order to analyse the software, we had developed a small Database that includes 3 persons and collected 20 samples of speech for each of them using Audacity software, sampling in 16Kz. Each of the samples is a small phrase of approximately 10 seconds in different days and circumstances.

After that we collect new samples and try to analyse if the software could correctly identify the person who had been recorded.

And finally we record samples from persons outside the Database and try to set a security threshold that will guaranty that the software indicates if the closest voice in the database belong to the person.

Initially we planned to implement a noise cancellation system to minimize the effects of the environment in the result, but we had could not find any reference that we could implement, for the lack of knowledge in the probability theory that was necessary to implement this feature.

5 Results

It is possible to sort the result in two different scenarios. The first one is when the recorded voice is in the same environment and computer that we used to create our database. The second is when the voice was recorded in other computer.

When the first scenario was analysed, it was possible to conclude that the software had its performance achieving as much as 70% of accuracy - that is, it has provided a correct identification in 70% of the cases. In the second scenario, the software was not reliable and resulted in about 20% of accuracy.

6 Conclusion

The speaker recognition system has shown a performance that can be deemed good, with an above half rate of success when recording and verification steps are held under the same environment and with the same equipment - that is, characteristics such as ambient noise and reverberation and noise-cancelling features from a microphone are the same. It is, however, not recommended the use of this software alone to grant access to sensitive data, such as a bank account or a house lock.

6.1 Future steps

- Develop a more reliable database that includes more peoples, and more samples of each people, in even more different scenarios;
- Implement a noise cancellation system to act in the frequency of the 20 Hz to 16,000Hz;
- Set a reliable threshold; and
- Optimize the software to be able to handle with more data.

7 References

[1] T. Kinnunen and H. Li, 'An overview of text-independent speaker recognition: From features to supervectors', *Speech Communication*, vol. 52, no. 1, pp. 12-40, 2010.

[2] U. Shrawankar and D. Thakare, 'Techniques for Feature Extraction In Speech Recognition System : A Comparative Study', *International Journal Of Computer Applications In Engineering, Technology and Sciences*, no. 0974-3596, pp. 412-418, 2013.

[3] L. Muda, 'Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques', *JOURNAL OF COMPUTING*, vol. 2, no. 3, pp. 138-143, 2010.

[4] Reynolds, D.A.: Gaussian Mixture Models. Encyclopedia of Biometric Recognition. Springer, Heidelberg (2008).

[5] D. Reynolds, 'Speaker identification and verification using Gaussian mixture speaker models',*Speech Communication*, vol. 17, no. 1-2, pp. 91-108, 1995.

[6] T. Moon, 'The expectation-maximization algorithm', *IEEE Signal Process. Mag.*, vol. 13, no. 6, pp. 47-60, 1996.

[7] Scholz, F. W. 2006. Maximum Likelihood Estimation. Encyclopedia of Statistical Sciences. 7.