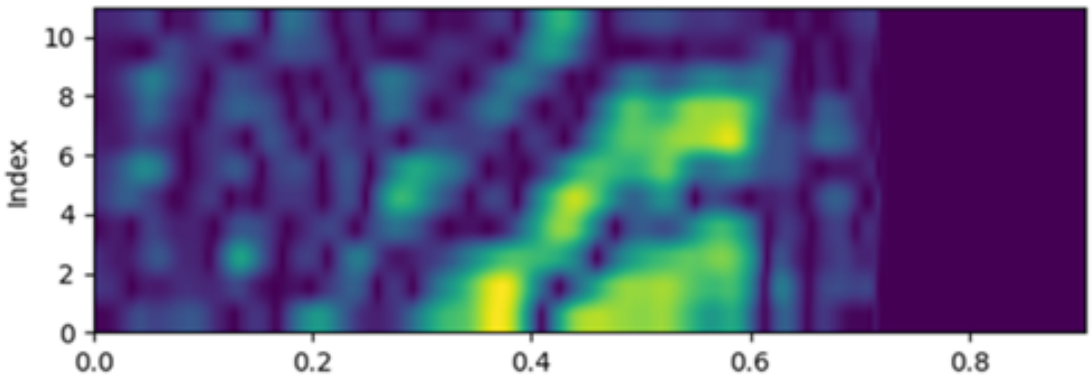
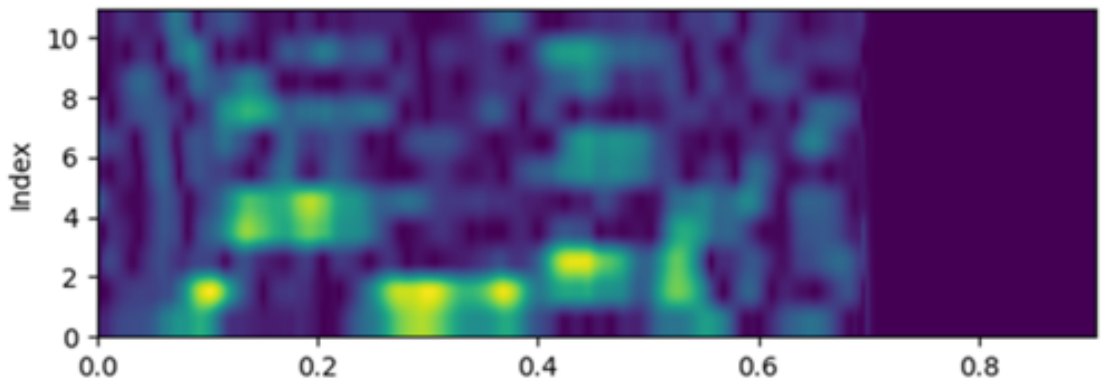
Image and Data Processing Report - Automatic Voice Recognition

October 2022

By Misha Kennerly



and Zoe Tanner



1 Introduction

The ever growing shift towards a digital society, where vast amounts of personal data is stored online, has raised many justified concerns over data security. There are many different methods of securing property. However, many traditional options have major issues; pins can be hacked and ID cards can be stolen [1]. Instead, by considering biometric identifiers, we could utilise a users features as a password. There are a range of biometric identifiers which are unique to every person, for instance; fingerprints, facial features and voice.

This project aims to investigate the methods of securing a device using automatic voice recognition (AVR). AVR operates on the premise that no two people have the same voice, meaning that it is possible to create a unique digital 'voice fingerprint' based on a persons voice. A user should then be able to speak to their device and be matched to their digital voice fingerprint to successfully be identified and allowed access to the device. If an imposter instead tries to gain access, their voice will not match up to the digital fingerprint, therefore they will be rejected. This provides a secure way of protecting property, which is why many large corporations are interested in using this technique.

2 Background

Using the human voice as a security system benefits from a large number of key identifies. Every human possesses a different voice which consists of varying tone, pitch, rhythm and frequency. It is clear that the average males voice has a lower pitch then the average females, however, the average range of an individuals voice is unique. The overall range of speech that humans can produce is 70-20,000Hz, which provides large variability between people (making the voice an ideal identifier).

By breaking the voice down into components such as pitch, phonantion (the process of the vocal cords producing certain sounds) and volume, we can isolate different identifiers which vary from person to person. Variance arises since sound is produced from the vocal folds, which vibrate to modulate the airflow at specific frequencies related to the geometry and tension of the surrounding cartilage. The vocal tract then attenuates or amplifies certain frequencies to output the sounds that we hear when we speak. Since the geometry of the vocal system varies between individuals, so does the sound they produce [2].

The way in which humans perceive sound is not linear, which poses a problem when using linear frequency scaling - such as Hertz. Instead, we have a logarithmic perception of sound, which can be visualised using a mel-frequency spectrum, which is a time-frequency spectrum with frequency in units of mel instead of Hz. The mel scale accurately describes how our ears perceive sound, which is why almost all methods of sound analysis use mels. The conversion between frequency in Hz and mel is as follows:

$$m = 2595 \log_{10}(1 + \frac{f}{700}) \qquad (1)$$

where m is frequency in units of mel and f is frequency in units of Hz [3].Therefore, the perceived difference between 2 and 3 mel will be the same as the difference between 3 and 4 mel, which cannot be said in units of Hz. In addition, volume is also perceived on a logarithmic scale, so amplitudes of frequency are conventionally converted to decibels.

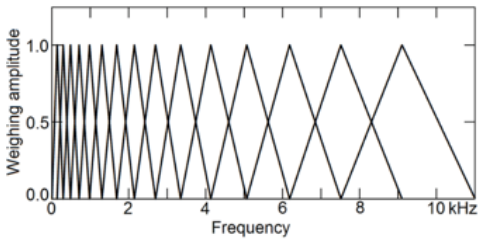


Figure 1: showing the construction of mel bands [4]

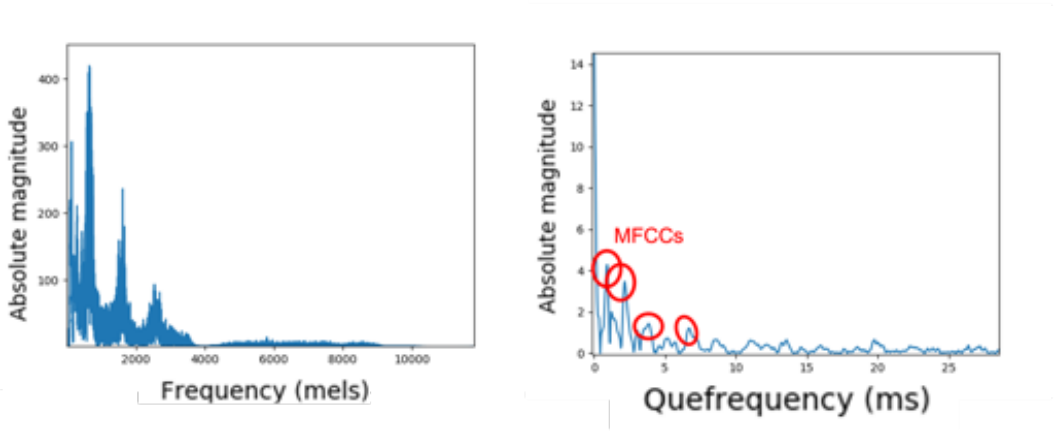


Figure 2: showing the mel-frequency plot of the a speech signal resulting cepstrum. Note the cepstrum was determined from the inverse fourier transform of the mel-frequency spectrum whereas the MFCCs are obtained discretely using the discrete cosine transform. The quefrency is the effective frequencies needed to reconstruct the mel spectrum.

In order to construct a mel frequency spectrum, the frequencies need to be sampled in terms of mel-bands. These are constructed by splitting the highest and lowest frequencies into equal units of mel. Then, frequency is converted back into Hz and triangular windows are created, centred at the respective mel frequencies in Hz with widths corresponding the centres of the adjacent mel frequency bands (as illustated in figure 1).

The signal can be weighted through these windows consecutively and the fourier transform of these weighted signals can be taken, before they are stacked on top of each other to produce the mel-frequency spectrum.

The frequency spectrum can be characterized in terms of formants. These are frequencies representing local maxima of the frequency spectrum and can be extracted from vowel sounds, as vowels are made from treating the vocal tract as tube without any obstruction. Obstruction due to the glottis in the vocal tract leads to a consonant sound.

Due to the geometry of the vocal cords and length of vocal tract, it can be said the distribution of formants from person to person for the same vowel is different[5]. This will be one of the parameters investigated in this project to distinguish between people.

The distribution of these formants (highlighted in red within figure 2b) can be characterised in terms of mel-frequency cepstrum coefficients (MFCC). Using the discrete cosine transform of a frequency spectrum, the coefficients showing what frequency cosines are required to reproduce the mel-frequency spectrum can be obtained.

With these, more variables can be determined such as the first and second derivative MFCCs (ΔMFCCs and $\Delta\Delta$MFCCs respectively) with respect to the adjacent cosines.

The other variable explored in this project is perceived pitch of voice. This is determined by the fundamental frequency, f_0, of speech whereas other dominant frequencies in the spectrum are formed from multiples of f_0. The range of f_0 typically spans from 85-120Hz in males and 165-255Hz in females[6]. This provides another parameter to distinguish peoples voices.

There are two main types of voice recognition; text dependant and text in-dependant. Text dependant voice recognition relies on keywords/phases while text independant does not and is more flexible. This project relies on text dependant voice recognition, where algorithms were built to distinguish between different speakers through the repetition of a key word.

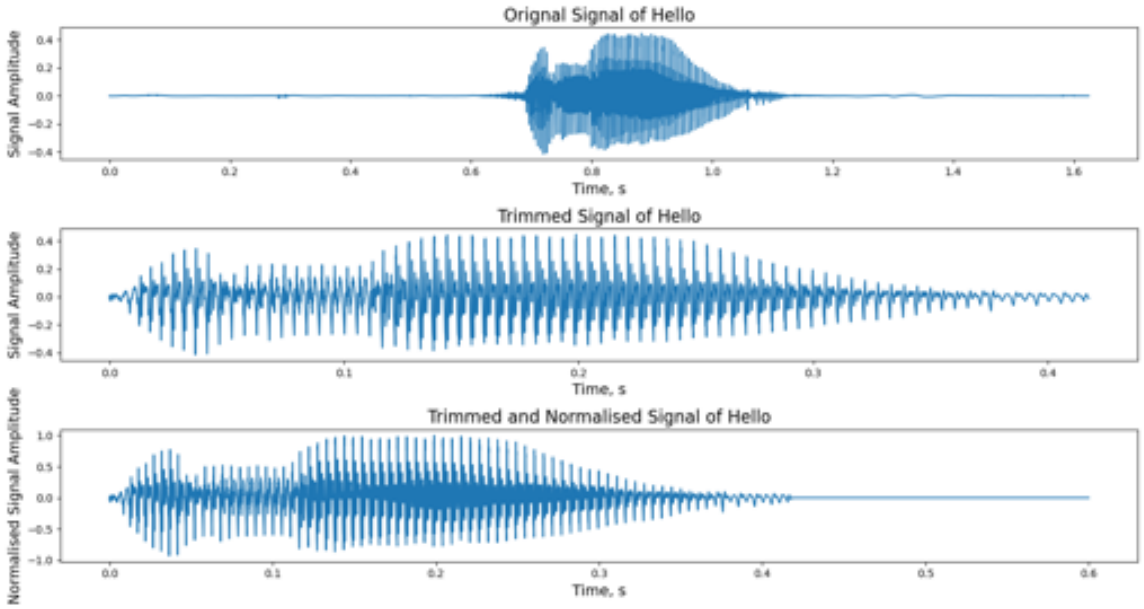


Figure 3: a) Original signal of the word 'Hello'. b)

There are many different approaches that could be considered when building a voice identification system - such as linear prediction coding (LPC) or using MFCCs [7]. In this project, it was decided that MFCC's would be used since they best represents the human perception of sound.

As f_0 and quantities from MFCCs are independent variables, orthogonal axis can be created, forming a plot of 'voice space'. For each speech a point on voice space can be appended and from multiple plots of the same person a centroid can be determined. After building up a representation of voice space for different peeople, if a test speech sample is added, an effective distance to each centroid can be determined. That test sample will be assigned to the nearest speaker group, represented by its centroid.

3 Methodology

3.1 Pre-processing: signal cleaning

First, to make things simple, a single syllable word was considered - this being 'air'. By using a word like 'air', it was possible to isolate just one phonation. Using a microphone, with a sampling rate of 22,050 Hz in a quiet room, to reduce noise, 4 people were asked to repeat the word 'air' twice. Once all the repeats were collected, the sound files were saved into a .wav file. This was later done for the word 'Hello', which was the word used for the majority of the project because it was two syllables long, meaning more features could be extracted than could be from 'air'.

Before any analysis could take place on the collected sound files, the files were cleaned. In order to clean up the sound files, several things were considered. These were the start and end points of the word, the amplitude of the signal and the length of the signal.

It was necessary to first find the start and end points of a word so that the word itself could be isolated (since the background noise before and after the the word was spoken were not of interest). Meaning that, any signals before and after the word could be cut. This process of 'trimming' is illustrated in figure 3a. As seen in the figure, the full original signal spans a total of 1.6 seconds. Then in figure 3

b, the signal of interest has been isolated and plotted, which now spans only 0.4 seconds.

There are many methods that could be used to find the start and end points of a word. In this project, since the background noise was relatively low, it was decided that the simplest way to do this was to first find the average absolute amplitude of the signal - by summing the absolute amplitude values for all data points and then dividing this by the total number of data points. A loop was then set up to filter through each data point, i, and average the absolute signal between i and i+100. Then if this average was below half of the total average signal, then it was considered as equal to nan. By repeating this over the whole signal, two tails were obtained. A set of nan data points at the start of the signal, and a set of nan values at the end of the signal. By cutting out these nan values, the isolated signal of interest could be found, as seen in figure 3 b.

Once trimmed, the signals were put through a further processing step in order to make sure each signal could be compared to one another. Firstly, the amplitudes were normalised, to ajust for louder and quieter sound files. Varying volumes may have occurred due to differences in the distance between the microphone and speaker, or due to variances in projection while speaking. While there are many methods that could be used to normalise the input signals - like integrating over the signal and setting the result to a constant value - the method used was dividing the signal by the maximum signal amplitude. This was decided as it was a simple but effective method, which could be implemented since the maximum signal amplitude was not due to random fluctuations. It also meant that all signals ranged in amplitude from -1 to 1, which made all the files comparable with one another. The final step in normalising the signal was to make sure all the signals were the same length (and contained the same number of data points). In order to do this, a new end point was defined, which was slightly longer than all of the trimmed signals (in the case of the 'Hello' signal in figure 3 this was chosen to be 0.6 seconds). Then, between the end of the trimmed signal and 0.6 seconds, a list of zeros was added (which had the same sampling rate as the rest of the signal).

These steps resulting in the trimmed and normalised signal illustrated in figure 3 c.

3.2 Extracting characteristics

After signal cleaning, the characteristics which define an individuals voice can be extracted.

In order to obtain these characteristics, the cleaned signals were first adapted to reflect the way in which humans perceive sound. To do this mel-spectrums were obtained using Librosas pre-defined mel spectrum function, which used the same theory as described in the background. 2048 time windows were used to determine the time resolution in the mel-spectrum and 90 mel-bands were used for the frequency resolution. After this mel frequencies for each time window were obtained and stacked to produce the mel-frequency spectrum for each speech sample.

To get MFCCs, the discrete cosine transform was applied to the mel spectrum giving the frequency cosines that are needed to reproduce the mel spectrum. Due to the information redundancy of high frequency cosines only the first 13 MFCCs were analysed.

From this the Δ MFCC values were calculated by taking the n+1th indexed MFCC from the nth MFCC in the specific time window. Δ MFCCs are more visually intuitive compared to MFCCs as shown in figure 4.

This is because the Δ MFCC values which correspond to speech silence are reduced to zero as only noise is picked up here hence the regions corresponding to speech are more readily highlighted. From this 12 coefficients from the 13 MFCCs per time window could be extracted.

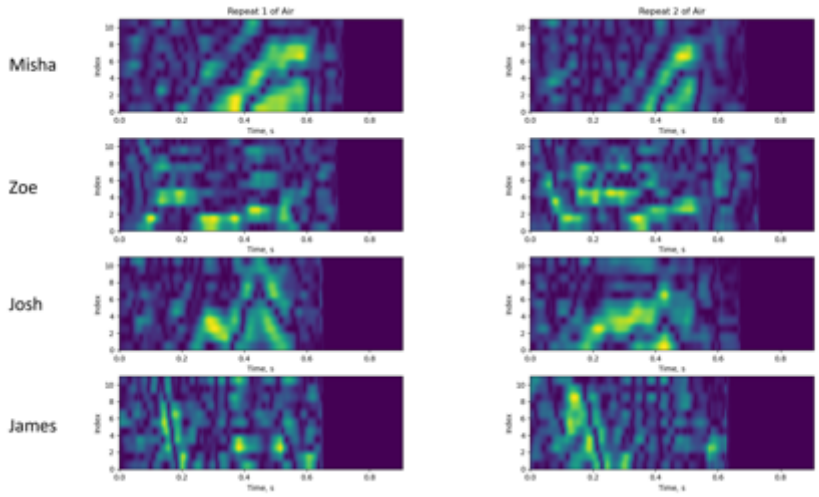


Figure 4: Showing the first 11 Δ MFCCs values across time for saying the word 'air', repeated twice for each speaker.

In figure 6, the three different speech characteristics reflected in the three different axes were 'Max MFCC value - Index 2', 'average pitch' and 'rms value between MFCCs and Δ MFCCs'. The meanings and methods of finding these values will be explained shortly. However it is key to note that these axes are not constrained to these characteristic identifiers, for example average word length, average volume or max MFCC value of a different index could have been used. In this project, 3 dimensions were used because it is simple to graphically represent 3D speech space, also only a small number of speakers were analysed, making it easier to differentiate between groups.

1: Max MFCC value - index 2

Separate MFCC plots could be obtained for each sound file by using the method outlined previously. Then, by isolating just the second index, a list of MFCC values (of index 2) over time could be obtained. There is a characteristic peak of MFCC values in time for each different speaker. By taking the maximum value in the isolated list described, the characteristic peak could be obtained. This method was implemented for all the repeats of all four speakers.

2: Average Pitch

Average pitch was obtained by taking the fundamental frequency of sound of the person speaking. This correlates to the perceived pitch of the person speaking. In a frequency spectrum distinct peaks will appear correlating to the dominant frequencies produced by human speech (f_0, f_1,f_2 etc.)

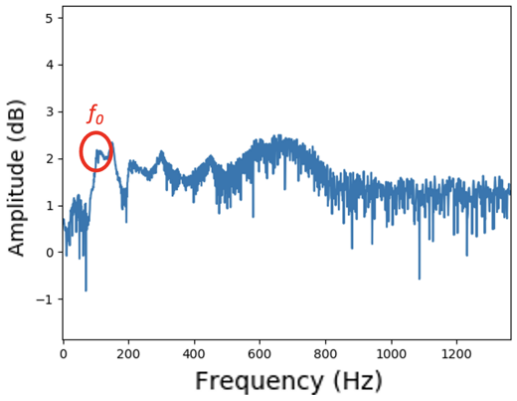


Figure 5: Illustrating the location of the peak that needs to be extracted from the algorithm

Before extracting the peaks, this data must be smoothed to avoid any peaks appearing due to noise which can be done using a

savgol filter. The smoothing was set to be done over 29 index values and with a 2nd order polynomial fit.

After smoothing the peaks the first peak in the spectrum can be extracted. Because the lowest fundamental frequency from a human voice ever recorded was 75Hz [8], the spectrum was analysed from this point onwards, so any peaks before could be neglected as they were not competitors for f_0

To select the peaks, find peaks was used from scipys libraries. Choosing the parameters to be of width minimum 20 Hz and minimum prominence of 0.015dB from the rest of the spectrum enables the algorithm to select the most important peaks in terms of their index values. From this the frequency corresponding to the first peak could be indexed.

3: RMSE Value Between MFCCs and Δ MFCCs

Root mean square error (RMSE) represents how closely two data sets fit each other, a value of zero meaning a perfect fit. If it can be hypothesised that each person has different RMSE value between the MFCCs and the Δ MFCCs, then this can be applied as an axis. By using a RMSE function from sewar libraries, these values can be extracted for each speech sample. For this to work, the matrices representing the MFCCs and Δ MFCCs need to be the same size, so the matrix was reshaped to only highlight the first 11 values

When these characteristics are obtained they can combined to form a set of orthogonal axes representing voice space, with each speech samples average pitch, RMSE between MFCCs and Δ MFCCs and max MFCC value plotted as a data point.

5 speech samples each from 4 people where used to make a up a training data set and from this a centroid could be taken representing the mean values on each axis for that person as shown in figure 6.

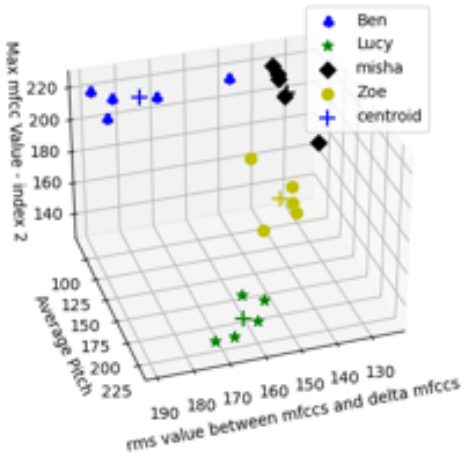


Figure 6: Plot illustrating 4 different speakers in 3D speech space. The 'plus' symbols represent the centroid of each speaker

3.3 Testing phase

Once the all the sound files were read in and plotted accordingly, the test phase could begin. To test the voice recognition system, each of the speakers were asked to say 'Hello' 5 times into the microphone. These samples were again processed in the same way, and by extracting the 3 key voice characteristics from each test data file, the test data could also be plotted onto the 3D voice space. A method known as vector quantization was then used to assess which group the test sound belonged to. The method of vector quantization involved finding the vectors between the test data point and the 4 centroids using equation 2.

$$D = \sqrt{(x_c - x_t)^2 + (y_c - y_t)^2 + (z_c - z_t)^2} \quad (2)$$

Where D represents the distance between the test data point and a given centriod,x,y and z are the axis of voice space and $(x, y, z)_c$ represents a centroid value for a given axes and $(x, y, z)_t$ represents the test value a given axes. The distance D can be illustrated by the figure 7.

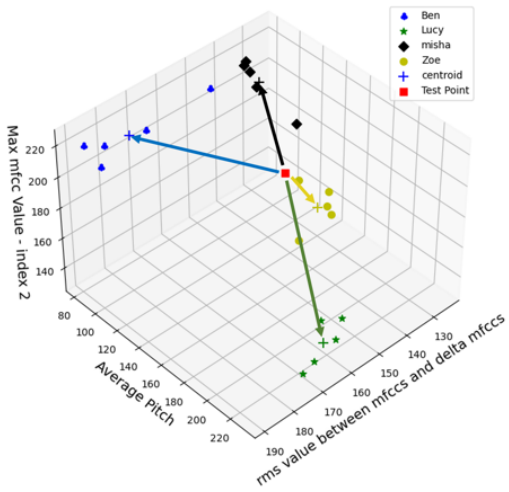


Figure 7: illustration of how vector quantization was implemented. Each arrow illustrates the effective distance between the test point and the chosen centroid.

By finding the minimum distance D, it was possible to predict which group the test point belonged to and therefore find the corresponding speaker it originated from. In the example shown in figure 7, it can be concluded that the yellow arrow is the shortest, meaning that the test point will be assigned to the yellow group (which corresponds to Zoe).

4 Results

By using the framework described in the methodology section, the following centroid data was found for the four speakers.

Speaker	RMS between MFCCs and delta MFCCs	Average Pitch (Hz)	Max MFCC Value
Ben	175.2	88.8	215.1
Lucy	164.6	227.5	142.5
Misha	132.7	93.7	206.5
Zoe	147.7	181	186.1

Table 1: Centroid data of the 4 speakers for the 3 chosen dimensions.

The results from the test data can been seen in table 2. By using the parameters defined in table 1 and one of the tests (selected from table 2), by using equation 2, the smallest vector D could be obtained. This value corresponded to a speaker which is represented in column 6 of table 2. It was found that the voice recognition system was able to accurately determine the true speaker 75% of the time. It also appeared to accurately find the difference between the two female voices (Zoe and Lucy) better than it could distinguish the two male voices (Ben and Misha). The reasons for which this may have occurred are highlighted in the discussion section.

Test no.	RMS between MFCCs and delta MFCCs	Average Pitch	Max MFCC Value	Distance to nearest centroid	Nearest centroid name	Correct person?
1	152	173.3	225.9	40.8	Zoe	Yes
2	156.6	175	214.9	30.7	Zoe	Yes
3	156.1	232.5	140.3	10.07	Lucy	Yes
4	174.2	88.3	203.6	11.6	Ben	No - Misha
5	152.3	94.2	204.1	19.7	Misha	Yes
6	154.6	95.8	207.7	22.1	Misha	Yes
7	172.8	88.3	207.1	8.4	Ben	No - Misha
8	148.4	178.3	204.8	18.9	Zoe	Yes

Table 2: Results table for testing phase. Illustrating the locations of test points in 'voice space' and whether the correct speaker has been matched to the test data.

5 Discussion

From this project a voice recognition software has been created, which results in 75% accuracy. There is an inconsistency between identifying different speakers.

One of the factors that limit accuracy is the number of dimensions in voice space. Despite 3 dimensions giving reasonable spacing between the centroids of the 4 speakers, there needs to be other ways to characterise the speakers voice patterns to space out the effective distances between the centroids. An example where this is evident is where Misha and Bens average pitch values are similar, so another axis would be useful to provide greater distinction between the voices. Some examples of other axis to include could be 2nd highest MFCC value or highest $\Delta\Delta$ MFCC value. This will increase the accuracy of the software especially if one of the data points has anomalous data on one of the axes. The effect of this will be diluted if all the other distances from the other axis fit closer to the centroid.

There were also errors in the average pitch algorithm. This came from the algorithm not recognising what the first fundamental frequency peak was. Even though visually the peak would be obvious, adjusting the parameters so that the rule for one would work for all of them was challenging as some fundamental frequency peaks would be less wide than the threshold but still more prominent leading to it being excluded. Therefore the parameters applied were used as a rule of thumb. Machine learning can be used to determine which peaks are relevant to the f_0 value.

Despite this, the algorithm still manages to find the fundamental frequency consistently showing distinction especially between male and female speakers. Also like volume, pitch can vary with each repeat depending on the tone of voice the speaker chooses, which will mean a larger spread of pitch values compared to other variables.

Furthermore more training data points can be added which will increase the accuracy of the centroid value, for instance 20 data points can be collected instead of 5.

Another way to get more axes would be to characterise the vowels used in a phrase, as different vowels have different formant frequencies [9] and these can be analysed individually. For this a larger phrase will be needed like 'my voice is my password' and a way to identify the vowels spoken.

As illustrated in Figure 4, it is easy see that each speaker has a very similar Δ MFCC value for each speech pattern which can be verified by eye. Another method of voice recognition would be to read in one of the second repeats (for example, Misha's second repeat), and take away each element in the image from each corresponding element in the four training data sets to find a difference between two images. By taking the mean of the differences, a value which represents similarity could be obtained (where 0 represents two identical images and ∞ represents two non correlated images). Therefore, in the example of using Misha's second repeat as the test image, it was found that the smallest 'similarity value' corresponded to Misha's training data, meaning the test data was correctly matched to the training data. This is a simple and effective method of voice recognition but is limited to only one variable to identify speakers. If the utterance has a different time length then this will mean a less valid comparison can be made as the elements corresponding to each phonation will be at different time points.

A way to improve on this would be to use a machine learning algorithm, which uses many repeats to learn how to identify a speaker. This method would become more accurate than just using the method described above as would utilise more sets of training data. We can also theorise that this would work because as mentioned, a human can identify the two matching repeats by eye, therefore it should be possible to create a machine learning algorithm to do the same.

6 Conclusion

This project successfully created a voice recognition system that identified 3 key characteristics from 4 different speakers voices. From this, a speaker could be correctly identified during testing 75% of the time. With further research more key characteristics could be utilised to more accurately predict the correct user during the testing phase.

Disclaimer: All participants have signed a consent form for their voices to be used within this project. Permission forms can be found within the project submission folder.

References

[1] Muhammad Nizam Kamarudin. Biometric voice recognition in security system. 02 2014.

[2] Zhang Z. Mechanics of human voice production and control. *J Acoust Soc Am.;140(4):2614.*, 10 2016.

[3] D. O'Shaughnessy. *Speech Communication: Human and Machine*. Addison-Wesley series in electrical engineering. Addison-Wesley Publishing Company, 1987.

[4] Julian Chen. *Elements of Human Voice*. 08 2016.

[5] Daniel Aalto, Jarmo Malinen, and Martti Vainio. Formants, 06 2018.

[6] J.L. s Fitch and Holbrook. *Modal Fundamental Frequency of Young Adults in Archives of Otolaryngology*. 1970.

[7] Sohaib Tallat. Voice identification and recognition system. 2014.

[8] Feinberg O'Connor JJ, Bennett PJ. Preferences for very low and very high voice pitch in humans. *PLoS One*, 7 2012.

[9] Emmanuel Ferragne and François Pellegrino. Formant frequencies of vowels in 13 accents of the british isles. *Journal of the International Phonetic Association*, 40(1):1–34, 2010.