Voiceprint System for Biometric Authentication

Geetha Bothe, Jonathan Leet, Rahul Raj, Ravi Ray, Mahesh Sooryambylu, Sreeram Vancheeswaran, and John V. Monaco

Pace University Seidenberg School of CSIS, White Plains, NY 10606

Abstract
Voice based authentication is one of the forms of biometric authentication. This approach will verify an individual speaker against his/her stored voice pattern. In this study we are using a specified common phrase for all users, which are segmented and analyzed using speech processing tools to compare a given voice sample against the corresponding stored sample. The voice samples are converted into a Spectrogram, which is then segmented using energy thresholds and compared against a stored voice sample. This paper describes the implementation details of techniques used - Fast Fourier Transformation, Mel frequency bank calculation, cepstral normalization, utterance segmentation and dynamic time warping.

1. Introduction

The concept of authentication goes hand in hand with most of the security related problems. Authentication is the act of confirming the identity of an entity. We can authenticate an identity in three ways: by something the user knows (such as a password or personal identification number), something the user has (a security token or smart card) or something the user is (a physical characteristic, such as a fingerprint, called a biometric). Biometric authentication has been widely regarded as the most foolproof - or at least the hardest to forge or spoof.

Voice recognition is one of the methods of performing biometric authentication. This paper describes voice print systems, which uses voice recognition for verifying a user is who he claims to be.

Some relevant work has been done in voice based biometric authentication systems:

2. Roy Morris (2012) [26] – In this project voice samples are segmented using energy values calculated manually from spectral information.

2. Overview

Voiceprint systems use the following four types of passphrases for user authentication:

1. A user-specified phrase, like the user's name
2. A specified phrase common to all users
3. A random phrase that the computer displays on the screen
4. A random phrase that can vary at the user's discretion.

This study focuses on using the specified, specially-designed passphrase, "My name is", which is common to all users. The reasons for using a common authentication phrase are:

1. It simplifies the segmentation problem. Earlier work indicated that features extracted from the individual phonetic units of an utterance increased authentication performance (Trilok, 2004), and this requires segmenting the utterance into phonetic units. Segmenting one known utterance into its phonetic units is much easier than segmenting many unknown utterances into their phonetic units.

2. It allows for the careful selection of the common phrase to optimize the variety of phonetic units for their authentication value. The "My name is" phrase contains seven phonetic units: three nasal sounds, the two [m]'s and one [n]; three vowel sounds, [aI], [eI], and [I]; and one fricative [z]. The nasal and vowel sounds characterize the user's nasal and vocal tracts, respectively, and the fricative characterizes the user's teeth and front portion of the mouth.

3. It facilitates testing for imposters since the common utterance spoken by non-authentic users can be employed as imposter utterances.

4. It permits the measurement of true voice authentication biometric performance and avoids potential experimental flaws. The combination of using the same authentication utterance for all users,
and one that consists of frequently used words that are easy to pronounce, avoids many of the experimental flaws that Maxion (2011) describes in measuring the performance of biometric systems.

A common authentication phrase for a person will constitute distinct voice components called phonemes. A phoneme is the smallest contrastive unit in the sound system of a language. Each phoneme has a pitch, cadence, and inflection, giving each person a unique voice. Based on the individual’s vocal tract, including tract length, ratio of larynx to sinuses cadence, pitch, tone, frequency, range, and duration of voice, the frequency values for the voice sample will vary. The similarity in voice comes from cultural and regional influences in the form of accents.

The frequency response generated from the common phrase utterance of an individual is called Spectrogram. The Spectrogram comprises of Frequency and Amplitude plotted over time. Frequency is the number of times per second a wave repeats in itself (or cycles). The amplitude measures the amount of air pressure variation. Once the individual phonemes or words from the common phrase being uttered by an individual are identified (segmented from the entire sample), the frequency and amplitude values for each such segment is distinct per individual. The spectrographic analysis helps to determine the individual phonemes from a phrase as well as distinctly identify the speaker based on these values. A Spectrogram is generated from the voice sample and helps in distinguishing the phonemes [m], [ai], [n], [ei], [m], [i] and [z] from the “My name is” portion of the utterance. The following tools were used in this study:

**Audacity**: Audacity is an open source audio editing software which is used for the voice recording, manual processing and Spectrogram generation. Audacity is available in all major platforms, including Windows, Mac and Linux.

**Matlab**: Signal Processing Toolbox in MATLAB provides digital signal processing functions for estimating the FFT, computing the energy values in mel-frequency bands etc., which are used for speech processing.

### 3. Methodology

The overall process involves managing speech samples in a repository, preprocessing and generating the spectrum, calculating Mel Filter Banks and MFCC, automatically segmenting the speech samples, using Dynamic Time Warping algorithm to segment the phonemes, and extracting the feature vectors.

### 3.1. Speech Repository

A centralized speech repository is created to maintain the speech samples. The speech repository provides the functionality to upload/download speech samples over http and to search the contents by user name. A text file containing phoneme marking is maintained along with each speech file to verify the processing results. For e.g., a marking file with the content “15214, 18786, 23593, 29767, 32457, 39242, 42868, 48554” identifies sample numbers corresponding to the beginning of [m], [i], [n], [ae], [m], [i], [z] and [z] end.

### 3.2. Preprocessing and Spectrogram generation

A waveform visually represents an audio signal. A waveform is made of samples, which is a value at a point representing the audio. Digital audio is sampled at discrete points. Sample rate is measured in Hz and represents the number of digital samples recorded per second to capture the wave form. An audio waveform changes constantly, so to simplify the processing we assume that on short time scales the audio signal doesn't change much or is quasi-stationary. A speech signal analyzed over a sufficiently short period of time gives stable acoustic characteristics. Shorter frames do not contain enough samples to get a reliable spectral estimate whereas in a longer frame signals change significantly.

Short-term analysis of a sound waveform is accomplished by buffering samples into 20-40ms frames. Buffering required to generate a spectrogram is shown in **Figure 1: Speech Frames**. The input stream is collected into consecutive frames containing 1024 samples. The direct buffering method is improved by overlapping the frames where each frame incorporate a set of samples from the previous frame. Successive frames slide the stream by 512 bytes, creating 50% overlap between two adjacent frames.

**Figure 1: Speech Frames**

Since each frame consists of 1024 entries per frame and 44100 Samples per second, it gives frame size of 44100/1024 = 43.06 Frames / Sec. So, one Frame size is approximately 23 milliseconds (1000 milliseconds/43.06). Window functions are used to smoothen the overlap between frames. A Hamming window function given below is often used to taper the sound frames towards the frame edges.
Human cochlea translates physical vibrations caused by a sound wave into electrical information the brain can recognize as distinct sound. A spectral estimate on each frame performs a similar task of identifying the frequencies that are present in the frame. A spectrogram is computed from a finite-length digital wave using the fast Fourier transform (FFT) algorithm. FFT is applied on each frame to generate its frequency components. A spectrum can be plotted by colorizing the spectral amplitudes. Higher amplitudes will be shown in hotter colors of selected color map. shows the spectrogram for the ‘My name is’ portion of an utterance.

Filter banks are prepared from the frequency values listed above. The first frequency band starts at the 1st point, reaches maximum at the 2nd point and then returns to zero at the 3rd point. Second band starts at 2nd point, peaks at 3rd and return to zero at the 4th point. The formula below derives the Mel triangle for each of the banks.

\[
f(x) = \begin{cases} 
0 & \text{when } k < l(m - 1) \\
\frac{k - l(m - 1)}{l(m) - l(m - 1)} & \text{when } l(m - 1) \leq k \leq l(m) \\
\frac{l(m + 1) - k}{l(m + 1) - l(m)} & \text{when } l(m) \leq k \leq (m + 1) \\
0 & \text{when } k > l(m + 1) 
\end{cases}
\]

3.3. Building Mel-Frequency Filter bank

A spectral estimate contains a lot of information that is not required for Speech Processing. Human cochlea cannot discern the difference between two closely spaced frequencies. This effect becomes more pronounced as the frequencies increase. For this reason, clumps of periodogram bins is summed to find the energy present in that region. This is performed using a Mel filterbank approach.

Mel frequencies is calculated from linear frequency as below

\[
m(f) = 2595 \times \log_{10}(1 + \frac{f}{700})
\]

Linear frequency is calculated from Mel frequency as below

\[
l(m) = 700 \times (10^{\frac{m}{2595}} - 1)
\]

The following set of linear frequency values measured in Hz were obtained using the above equation for the sampling rate of 44100 Hz. For feature extraction, the lower 13 frequencies from the 26 mel values are considered.

\[
100,215,346,496,667,863,1087,1343,1636,1970,2353,2791,3291,3862,4516
\]

The features vectors are extracted by passing each frame through the Mel filter banks.

3.4. Mel Frequency Cepstral Coefficients (MFCC)

MFCC is computed by applying Discrete Cosine Transformation of Logarithm of the Mel values. The speech segment is then represented as a sequence of Cepstral Vectors to be used in the next step, DTW comparison.

13 Mel Values/Frame → Logarithm → Discrete Cosine Transform

3.5. Utterance Segmentation

B1.3
Voiceprint system operates on the initial "My Name Is" portion of the utterance and must be extracted from the rest of the speech. This is accomplished by identifying the beginning of the voiced segment and ending of [z] sound present in ‘is’.

Voiced segment indicates periodicity in time domain and harmonic structure in frequency domain. Unvoiced segment is random noise-like in time domain and spectrum without harmonic structure in frequency domain. The voiced speech segment is characterized by the periodic nature, relatively high energy, less number of zero crossings and more correlation among successive samples.

The voiced speech can be identified by observing the waveform in the time domain due to its periodicity nature. Further, the spectrum will have more energy, typically, in the low frequency region.

The silence region is characterized by the absence of any signal characteristics, lowest energy compared to unvoiced and voiced speech segments, relatively more number of zero crossings compared to unvoiced segment. The silence region can be identified by observing the waveform in the time domain due to absence of any signal. In the frequency domain, the absence of any spectral information is the indication that the segment is silence region. The lowest energy can be observed in terms of absence of any appreciable amplitude values. Even though energy is a good information for identifying silence regions, the robustness can be improved further using the zero-crossings information.

Figure 4: Speech Waveform, clearly indicates the voiced and unvoiced segments visually.

Zero crossing rate is plotted in red and energy is plotted in green. Non-voiced segments captures high zero crossing rate and low energy values. As opposed to this voiced segments indicate low zero crossing rate and high energy values. The vertical blue lines mark the ‘My Name is’ part of the speech segment automatically. Figure 5: Energy vs. Zero Crossing shows energy values compared with the zero crossing rate for the previous wave form.

The end of the utterance "My Name Is" is calculated by locating the [z] sound at the end of "is". Locating the [z] sound is done by thresholding the ratio of the energy in the high frequency bands (sum the energy in the bands over about 5 kHz) to that in the low frequency bands (sum those under 5 kHz).

This process marks the speech segment for ‘My name is’ (voice-beginning at frame 56, 'z’ ending at frame 119) and generates Mel values for the marked segment. Figure 6 represents the automatically marked spectrum, in Matlab. The first vertical line indicate beginning of voiced sample and the second vertical line marks the end of ‘z’ sound.

3.6. Cepstral Mean Normalization

To reduce the effect of multiplicative noise on the feature vectors, the energy in a band in every frame is normalized by the mean energy in that specific band across the utterance. Thus, the normalized energy values across thirteen Mel-frequency bands for all the frames is obtained as follows

\[ C_i = c_i - \frac{1}{N} \sum_{k=1}^{N} c_{i_k} \]

Where, \( c_i \) is the \( i_{th} \) feature element in the feature vector and \( c_{i_k} \) is the \( i_{th} \) feature element at frame \( k \). \( N \) is the number of total input frames of data.

The values generated in this process is the input to the subsequent step Dynamic Time Warping (DTW) and phoneme segmentation.

3.7. DTW and automatic phoneme segmentation

Dynamic time warping (DTW) is a technique that finds the optimal alignment between two time series if one time series may be “warped” non-linearly by stretching or shrinking it along its time axis. The warping is used to
determine the corresponding regions between two time series or to check the similarity between them. The two samples being compared is represented as $X = x_1, x_2, \ldots, x_n$; $Y = y_1, y_2, \ldots, y_m$ where $x_i$ represents a sample extracted from the time series $X$ whereas $y_j$ represents the sample extracted from time series $Y$. The Warp Path $W = w_{1,1}, \ldots, w_{i,j}, \ldots, w_{n,m}$ where $i$ is an index in time series $X$, and $j$ is an index in time series $Y$. So, a point in the warp path $k$ can be represented as follows $w_k = (i,j)$ and next point $w_{k+1} = (i',j')$ if $i \leq i' \leq i+1$, $j \leq j' \leq j+1$.

DTW arranges the sequences on sides of a grid starting at bottom left of the grid. The best alignment between the two sequences is identified by finding the optimal path through the grid. Dynamic programming is used to find the minimum distance warp path where solutions to sub-problems (sections of time series) are identified and used to solve slightly larger problem until the final problem solution is derived. The warp path starts at $(1,1)$ and finishes at the end of both the series $(X,Y)$. Another constraint over the warp path forces $(i,j)$ to monotonically increase in the warp path.

A two dimensional $|X| \times |Y|$ cost matrix $D$ is computed from the two time series $X$ and $Y$. The absolute value of difference between the two samples are used to derive the cost matrix. Using dynamic programming if $D(i, j)$ is the minimum distance warp path of the series at $(i,j)$ then minimum distance warp path must be known for all data points that are single point away from $(i,j)$. Since the next point in the warp path must either be incremented by one or must stay along $(i,j)$, the decision to find the next point in the warp $W(i, j)$ is:

$$D(i, j) = Dist(i, j) + \min(D(i - 1, j), D(i, j - 1), D(i - 1, j - 1))$$

If the warp path passes through $D(i, j)$ then the sample $X_i$ is warped to the point $Y_i$. If there is a vertical section in the warp path, a single point in series $X$ is warped to multiple points of series $Y$.

Figure 7: Sample Warp Path represents the cost matrix and the warped path for the two time series represented along the axes.[7]

The input voice is compared against a reference speech sample with the seven phonemes from the phrase ‘My Name Is’ marked. DTW comparison between the two speech samples identify the warping path between them. This warp path is used for marking the seven phonemes of input sample by aligning with the corresponding marks in the reference speech sample.

Figure 8: Warp Path shows warp path between two such speech signals plotted in Matlab. The speech sample along y-axis is the reference template which we selected for the utterance ‘My Name Is’ whereas the speech sample along x-axis is the input voice for the same phrase which is to be compared with the reference template. The reference template has 62 frames between beginning of phoneme [m] and end of phoneme [z] whereas the input voice has 66 frames between beginning of [m] and end of [z].

Table 1: Warp Path captures the warp path. The first column represents the sample number of the reference speech template and second column captures the sample number of the input voice.
Table 1: Warp Path

<table>
<thead>
<tr>
<th>Warp Path</th>
<th>Warp Path Ctd.</th>
<th>Warp Path Ctd.</th>
<th>Warp Path Ctd.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample</td>
<td>Input</td>
<td>Sample</td>
<td>Input</td>
</tr>
<tr>
<td>16384</td>
<td>30720</td>
<td>24064</td>
<td>37376</td>
</tr>
<tr>
<td>16896</td>
<td>30720</td>
<td>24576</td>
<td>37376</td>
</tr>
<tr>
<td>17408</td>
<td>30720</td>
<td>25088</td>
<td>37888</td>
</tr>
<tr>
<td>17920</td>
<td>30720</td>
<td>25600</td>
<td>38400</td>
</tr>
<tr>
<td>18432</td>
<td>30720</td>
<td>26112</td>
<td>38912</td>
</tr>
<tr>
<td>18944</td>
<td>31232</td>
<td>26624</td>
<td>39424</td>
</tr>
<tr>
<td>19456</td>
<td>31744</td>
<td>27136</td>
<td>39936</td>
</tr>
<tr>
<td>19968</td>
<td>31744</td>
<td>27648</td>
<td>39936</td>
</tr>
<tr>
<td>20480</td>
<td>31744</td>
<td>28160</td>
<td>40448</td>
</tr>
<tr>
<td>20992</td>
<td>31744</td>
<td>28672</td>
<td>40960</td>
</tr>
<tr>
<td>21504</td>
<td>32256</td>
<td>29184</td>
<td>41472</td>
</tr>
<tr>
<td>21504</td>
<td>32768</td>
<td>29696</td>
<td>41984</td>
</tr>
<tr>
<td>21504</td>
<td>33280</td>
<td>30208</td>
<td>41984</td>
</tr>
<tr>
<td>21504</td>
<td>33792</td>
<td>30720</td>
<td>41984</td>
</tr>
<tr>
<td>21504</td>
<td>34304</td>
<td>31232</td>
<td>42496</td>
</tr>
<tr>
<td>22016</td>
<td>34816</td>
<td>31232</td>
<td>43008</td>
</tr>
<tr>
<td>22528</td>
<td>35328</td>
<td>31232</td>
<td>43520</td>
</tr>
<tr>
<td>23040</td>
<td>35840</td>
<td>31232</td>
<td>44032</td>
</tr>
<tr>
<td>23040</td>
<td>36352</td>
<td>31232</td>
<td>44544</td>
</tr>
<tr>
<td>23040</td>
<td>36864</td>
<td>31232</td>
<td>45056</td>
</tr>
<tr>
<td>23552</td>
<td>37376</td>
<td>31744</td>
<td>45568</td>
</tr>
</tbody>
</table>

The reference template is marked at sample numbers 15214, 19456, 23593, 29767, 32457, 39242, 42868, and 48554 to indicate the beginning of phonemes [m] [i] [n] [ae] [m] [i] [z] and end of [z]. A look up on Table 2: Warp Path maps the sample numbers of input voice to nearest frame corresponding to the marked sample number in reference template [30720, 31744, 37376, 41984, 47616, 51712, 57856, 61952].

3.8. Feature Vectors

Feature vectors represent the discriminative speech features suitable for detection. The following feature vectors are extracted.

Means of the energy in each of the 13 mel-frequency bands over the entire utterance: - This represent the average of energy values across the entire utterance, for each of the mel-frequency bands.

Variance of the energy in each of the 13 frequency bands over the entire utterance: - This represent the variance of energy values across the entire utterance, for each of the mel-frequency bands.

Means of Energy values for each of the phonemes across the 13 frequency Bands: - This represents the average of energy values for the frames representing a particular phoneme across the 13 mel-frequency bands.

Manual marking of phonemes: - This represent the manual markings of the phonemes, i.e., the frame number which denote the beginning and the frame number which denote the end of the phoneme, in the utterance.

Voice fundamental Frequency (F0):- The fundamental frequency is estimated by finding the most dominant frequency of repetition on a portion of the voice sample. A reliable way of obtaining an estimate of the dominant fundamental frequency for long, clean, stationary speech signals is by using the cepstrum. The cepstrum is a Fourier analysis of the logarithmic amplitude spectrum of the signal. If the log amplitude spectrum contains many regularly spaced harmonics, then the Fourier analysis of the spectrum will show a peak corresponding to the spacing between the harmonics: i.e. the fundamental frequency. Effectively we are treating the signal spectrum as another signal, then looking for periodicity in the spectrum itself.

Fundamental frequency is obtained from the cepstrum by finding the peak in the quefrency region corresponding to typical speech fundamental frequencies and detecting the peak in the cepstrum between 1 and 20ms. The cepstrum works best when the fundamental frequency is not changing too rapidly, when the fundamental frequency is not too high and when the signal is relatively noise-free.
Voice formant Frequency ($F1 - F5$):- The formant frequencies are properties of the vocal tract system and need to be inferred from the speech signal rather than just measured. The spectral shape of the vocal tract excitation strongly influences the observed spectral envelope, such that we cannot guarantee that all vocal tract resonances will cause peaks in the observed spectral envelope, nor that all peaks in the spectral envelope are caused by vocal tract resonances.

![Figure 9: Portion of the voice sample in waveform, spectrum and the Cepstrum plotted using Matlab.](image)

The dominant method of formant frequency estimation is based on modelling the speech signal as if it were generated by a particular kind of source and filter:

![Signal with Flat Spectrum System having only simple resonances Observed speech signal](image)

This type of analysis is called source-filter separation which takes into account the modelled system and the frequencies of its resonances. The best matching system is obtained using Linear Prediction Analysis (LPC). Linear prediction models the signal as if it were generated by a signal of minimum energy being passed through a purely-recursive Infinite Impulse Response (IIR) filter. LPC is used to find the best IIR filter from a section of speech signal and the filter's frequency response is plotted.

To find the formant frequencies from the filter, the locations of the resonances that make up the filter are identified. This involves treating the filter coefficients as a polynomial and solving for the roots of the polynomial.

4. Experiment Results

The experiments involved voice sample collection from different persons, manual and automated segmentation of the "My name is" portion from the voice sample, manual and automated segmentation of the seven phonemes "M", "Ai", "N", "Ei", "m", "i" and "z", spectrogram generation, MFCC generation and feature extraction.

Speech samples were collected from 30 persons. Each person was asked to repeat the utterance "My name is" followed by the person's name -- for example, "My name is John Smith.". 20 samples each from the 30 persons were recorded. The samples were recorded using 'Audacity'. The samples were collected in mono channel at a sampling rate of 44100 samples/sec. The speaker metadata i.e., name, gender, and age were maintained in the Speech Repository.

The manual segmentation of speech samples were accomplished using Audacity by identifying and marking the beginning and end of "My name is" portion, beginning and end of the seven phonemes "M", "Ai", "N", "Ei", "m", "i" and "z". These markings were captured in a text file and maintained in the speech repository. The manually marked values are used to validate the automated segmentation.

From each of the 600 samples collected and processed, two sets of features were extracted. Pace University Biometric Authentication tests reported the following results.

1. Features taken from each of the seven phonemes in the entire phrase 'My Name Is' reported Equal Error Rate of 1.05%
2. Features taken from the entire phrase 'My Name Is' reported an Equal Error Rate of 1.95%

Matlab code to perform the voice sample segmentation achieved the following results.

1. Achieved 98% accuracy in locating the beginning of phrase 'My Name is' from the voice samples. The 2%
failures are attributed to feeble energy levels in the samples.
2. Achieved 94% accuracy in locating the end of 'My Name is' from the speech samples. The 6% errors are due to those specific samples where higher frequency components lacked the 'z' sound.
3. Achieved 94% accuracy in automatically segmenting the seven phonemes based on a segmented voice template using Dynamic Time Warping module

5. Conclusion
A methodology for the voice based authentication system is described using Fast Fourier Transformation, Mel frequency bank calculation, cepstral normalization, utterance segmentation and dynamic time warping. A system for identifying the beginning and end of the common phrase from the voice sample is described and also an approach to automatically segment the phonemes from the specific phrase is defined. The feature vectors extracted from various voice samples belonging to different subjects is used to perform Biometric Authentication.

6. References
[1] Kevin Gold (kevin.gold@yale.edu) and Brian Scassellati (scac@cs.yale.edu) - Audio Speech Segmentation Without Language-Specific Knowledge
[3] Nitin N Lokhande, Dr.Navnath S Nehe and Pratap S Vikhe - Voice activity detection algorithm for Speech Recognition Applications
[18] Lena Maier-Hein Speech Recognition Using Surface Electromyography
[23] Mohammad T. Shami and Mohamed S. Kamel - Segment-based approach to the recognition of emotions in speech
[28] The MFCC by Aldebaro Klautau - 11/22/05