Speech Based Biometric System Using GFCC Features

Ravneet Singh¹, Navpreet Kaur², Parminder Singh³

1 pursuing M.Tech, Electronics and Communication
2 Assistant Professor, ASRA College of Engg. And Tech
3 Assistant Professor, ASRA College of Engg. And Tech
Bhawanigarh, Punjab

Abstract: Numerous Biometric modalities have emerged recently as researchers are continuously striving to achieve higher security standards for applications that use biometrics to provide access to users in critical environment. Gait of a person, Thermo gram, Near Infrared Images, Smile Recognition, Lip Movement recognition, Thermal Palm, Hand Finger knuckle, Finger veins, Nail ID, Skin Spectroscopy, Electrocardiogram, Dental Biometrics and DNA are some of the biometric techniques being used recently. Speech is simplest, nonintrusive, unimodal biometric parameter. However acoustic degradation, background noise, channel noise, health and ageing of an individual affect the quality of voice which makes Speaker Recognition System not very efficient. Hence, preprocessing and voice enhancing techniques and strategies need to be considered. The vulnerability lies when the testing and enrolment conditions are not perfectly matched which is often the case in real time environment. Degraded speech signals hence limit the effectiveness of Speaker identification and verification. In the proposed work, speaker identification has been carried out using gammatone frequency cepstral coefficients (GFCC) as these found robust to noise than other popular speech features i.e. Mel-frequency cepstral coefficients (MFCC). To divide the speech features in small space, clustering has been carried out using Gaussian mixtures which provide combined feature-set properties of a speech sample. Further these features are fed to pattern recognition based neural network which classifies the data into particular speaker identification. Experimental results found 92% of accuracy for the collected dataset.

Keywords: GFCC, speaker identification, Gamma tone filters

1 Introduction

With the intensification of identification technology research, speaker recognition systems in the laboratory environment have achieved high recognition rates [1]. However, a practical application environment, in which additive noise, room reverberation, and channel/handset variations exist and have to be considered, poses considerable challenges to such systems. The features can perform well in clean environments but not in adverse environments or under mismatched training and testing conditions. In the SPHINX system, for example, when researchers used a different microphone for model training and recognition, the recognition rate dropped from 85.3 to 18.6 % and 76.5 to 36.9 %, respectively [2]. As another example, under clean testing conditions, the MFCC features achieved over 96 % accuracy, when the signal-to-noise ratio (SNR) of the testing condition drops to 6 dB, the accuracy of the mel-frequency cepstrum coefficient (MFCC) features decreases to 41.2 % [3]. The fundamental reason for this is that the distortion of speech in different environments causes a mismatch between the training environment and the testing environment. As a result, the speech information obtained by training data does not correctly reflect the data of the testing environment, which exerts a great influence on the recognition results. The source of being not robust for speech recognition lies in the mismatch between the training and the testing environments. The purpose of noise robustness in speech recognition is to eliminate the mismatch between the training and the testing environment that is caused by noise. There are mainly four kinds of methods: robust feature extraction, speech enhancement, model compensation, and microphone arrays. This paper studies robust feature extraction, in particular, auditory feature extraction. There are many feature extraction methods based on human ear hearing today. For example, MFCC [4], perceptual linear predictive (PLP) [5], relative spectra-perceptual linear predictive (RASTA-PLP) [6], and cochlear cepstrum coefficient (CFCC) [3] have become the mainstream techniques employed by the robust feature extraction method.

2 Feature extraction based on the gammatone filter

The cochlea is the vital organ of the human ear hearing system. The basilar membrane is an important structure of the cochlea that receives sound. Sound waves spread in the form of a traveling wave
along the basilar membrane. The dynamic response process of sound generated by the basilar membrane is similar to the process of shaking a ribbon. The basilar membrane has important frequency-selective characteristics. For a pure tone input whose frequency is high, the location of the maximum amplitude is near the bottom of the basilar membrane. On the other hand, when the frequency is low, the location of the maximum amplitude is at the top of the basilar membrane [7]. The basilar membrane also has a spectrum analysis feature: for complex sound input, the basilar membrane can convert different frequencies into different locations of the basilar membrane. The intensity of different sounds is converted into different vibration amplitudes of the basilar membrane. Thus, different frequency components and the corresponding amplitudes of a compound sound are separated. The response process of the basilar membrane to sound in different positions is equivalent to the filter’s response process. In order to better understand the mechanism of the basilar membrane, it is critical to seek the type of filter that can best simulate the filtering characteristics of the basilar membrane. This paper uses the gammatone filter bank [8] to realize the cochlear model.

3. Proposed speaker identification system using Gammatone filter banks

Derived from psychophysical observations of the auditory periphery, the gammatone filter is a standard cochlea auditory filter. It uses a causal impulse response function to describe the characteristics of the filter. This impulse response function was earliest used to describe the physiological impulse response characteristics of the cat’s auditory nerve [8]. The impulse response of a gammatone filter centered at \( f_i \) is

\[
g_i(t) = A t^{-n} \exp(-2\Pi b_i t) \cos(-2\Pi f_i + \phi_i) \tag{1}
\]

where \( A \) is the gain of the filter. It is a constant, usually equals to 1. \( t \) refers to time, \( n = 4 \) is the order of the filter, \( f_i \) is the central frequency of the \( i^{th} \) filter, \( \phi_i \) is the phase shift, and \( U(t) \) is the unit step function. In the simplified model, \( b_i = 1.019 ERB(f_i) \) is the attenuation factor of the \( i^{th} \) filter, determines the decay rate of the impulse response with respect to the filter bandwidth, whereas the bandwidth of each filter is related to the human auditory critical band. From auditory psychology, the mathematical expression for the attenuation factor is

\[
ERB(f_i) = 24.7 \times \left( 4.37 \times \frac{f_i}{1000} + 1 \right) \tag{2}
\]

The bandwidth of each filter is determined by (2), in which the mathematical expression of the center frequency \( f_i \) is given by the following formula:

\[
f_i = (f_H + 228.7) \exp\left(-\frac{V}{9.26}\right) - 228.7 \quad 1 \leq i \leq N \tag{3}
\]

The value of \( N \) is 128, which indicates that the cochlear filter model is achieved by stacking 128 gammatone filters. \( f_H \) is the filter cutoff frequency. The center frequency of each gammatone filter is distributed between 30 and 5500 Hz, and the center frequency of each filter is equally distributed in the domain of the ERB (equivalent rectangular bandwidth). The frequency coverage of all the filter banks ranges from 80 to 8000 Hz. The distribution of the gammatone filter banks’ frequency response is shown in Fig. 1. It is easy to observe, filter’s band lying at the high- frequency is wider than filter’s band lying at the low frequency. Curves distribute denser at low- frequency segment than that at high- frequencies segment, which implies that the frequency analysis is better at low- frequency than that at high- frequency and also highly resembles the frequency analysis of basilar membrane. Meanwhile, it is easy to note that the gradients is larger in both sides of central frequency \( f_i \), which indicates that Gammatone filter has sharp frequency selective property, the decrease on the edge of the filter is slow, which avoids the energy leakage between the adjacent frequency-bands. These properties of Gammatone filters are consistent with filtering property of basilar membrane while weaken the influence of offset of resonance peak on speech characteristics parameter extraction when different person pronounces the same tone.
Fig. 1: Impulse frequency response of 128
channels' gammatone filter banks (each curve
represents four channel). The bandwidth of the
$\text{th}$ filter is determined by the attenuation factor ($b_i$) of
the $i$th filter.

1) Pre-emphasis
In general, the digitized speech waveform
has a high dynamic range and suffers from additive
noise. In order to reduce this range, pre-emphasis is
applied by using a first-order finite impulse response
(FIR) high-pass filter:

$$H_p(z) = 1 - \mu z^{-1}$$

(8)

where the pre-emphasis coefficient is
$\mu = 0.97$. Boosting the high-frequency energy
makes the information from these higher frequencies
available to the acoustic model.

2) Framing and windowing
First, we split the signal into several frames
so that we can analyze each frame in a short time
instead of analyzing the entire signal. Speech ranging
from 10–30 ms is basically short-term stationary.
Each time frame is windowed with a Hamming
window to eliminate discontinuities at the edges. The
filter coefficients $w(n)$ of a Hamming window of
length $n$ are computed according to the following
formula:

$$w(n) = \begin{cases} 
0.54 - 0.46 \cos \left( \frac{2\pi n}{2N-1} \right), & 0 \leq n \leq N-1 \\
0, & \text{otherwise} 
\end{cases}$$

(9)

where $N = 256$ is the total number of
samples in each frame.

3) Fast Fourier transform (FFT)
Spectral analysis shows that different
timbres in speech signals correspond to different
energy distributions over frequencies. Therefore, FFT
is usually performed to obtain the magnitude
frequency response of each frame[9]. We applying
FFT to a windowed speech frame for obtaining its
frequency spectrum, $X(k)$.

4) Frequency response of Gammatone filter banks
The frequency response of the $l$th
gammatone filter $H_l(k)$ is obtained by the
following Eq. (10).

$$H_l(k) = \text{DFT} \left[ g_l(n) \right], \ l = 1,2,3,...,L$$

(10)

Where $L$ represents the number of filter’s
channel. $g_l(n)$ is the impulse response in time—
domain. It is obtained by Eq. (7).

5) Equal-loudness pre-emphasis
The study of physiological acoustics shows
that human auditory system sensitivity to different
frequencies is different at the same sound intensity.
Humans are more sensitive to medium and high
frequency, and so a loudness level transformation has
to be applied to the critical-band spectrum. The
function $E_l$ is an approximation of the equal
loudness of human hearing at different frequencies.
The equal loudness weight of the $l$th filter, $E_l$

$$E_l = \frac{\omega_l^2 \times (\omega_l^2 + 1.44 \times 10^6)}{(\omega_l^2 + 1.6 \times 10^5) \times (\omega_l^2 + 9.61 \times 10^6)}$$

(11)

where $\omega_l$ is the central frequency
corresponding to the $l$th filter.

6) Gammatone filter-banks filtering
The short-time energy spectrum is calculated
by squaring the frequency spectrum $X(k)$, after
which gammatone filter banks are used to filter the
energy spectrum of each frame signal. The filter
output after equal loudness, $X_{\text{pe}}(k)$, is the product of
the filter outputs. It is given by
\[ Y_l = \sum_{k=0}^{N} X(k)^2 \times HE_l(k) \times I \leq l \leq L \]

Where \( L \) represents the number of filter’s channel. \( N \) is the number of \( FFT \) points in each windowed frame. \( X(k)^2 \) is the energy spectrum.

Firstly, a signal is passed through gammatone filter bank. Equal loudness is applied to each filtered output. Further, a logarithmic and Discrete Cosine transform is applied to get GFCC features of input signal.

GFCC has a fine resolution at low frequency as compared to MFCC[10]. MFCC works with a log while GFCC works with cube root. Cube roots provide more robustness to GFCC as compared to that of logs in MFCC. Hence, GFCC is more robust in noisy environment than MFCC.

**Figure 2: Block diagram of proposed feature selection method**

### 4. Results and discussions

After feature extraction phase, artificial neural network has been applied to get speaker identification base on testing and training of collected datasets. measure of how well the neural network has fit the data is the confusion plot. Here the confusion matrix is plotted across all samples. The confusion matrix shows the percentages of correct and incorrect classifications. Correct classifications are the green squares on the matrices diagonal. Incorrect classifications form the red squares. If the network has learned to classify properly, the percentages in the red squares should be very small, indicating few misclassifications. In proposed work, only seventeen samples out of 192 samples misclassified with wrong speaker. Confusion matrix shows 92.2 % percentage sensitivity in overall performance.

**Classification Accuracy**

The classification accuracy is the extent to which the classifier is able to correctly classify the exemplars and is summarized in the form of confusion matrix to the test data. This is defined as the ratio of the number of correctly classified patterns (TP and TN) to the total number of patterns (species) classified. It is defined as the summation of diagonal elements in the confusion matrix divided by total number of classes. It is the fraction of the total number of face correctly classified. Therefore from confusion matrix we have plotted the graphs as below

<table>
<thead>
<tr>
<th>Data No.</th>
<th>Samples</th>
<th>TP</th>
<th>TN</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data 1</td>
<td>12</td>
<td>12</td>
<td>0</td>
<td>100</td>
<td>2</td>
<td>85.79</td>
</tr>
<tr>
<td>Data 2</td>
<td>12</td>
<td>12</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>100.00</td>
</tr>
<tr>
<td>Data 3</td>
<td>12</td>
<td>11</td>
<td>1</td>
<td>91.70</td>
<td>1</td>
<td>91.70</td>
</tr>
<tr>
<td>Data 4</td>
<td>12</td>
<td>10</td>
<td>2</td>
<td>83.30</td>
<td>0</td>
<td>91.65</td>
</tr>
<tr>
<td>Data 5</td>
<td>12</td>
<td>11</td>
<td>1</td>
<td>91.70</td>
<td>1</td>
<td>91.70</td>
</tr>
<tr>
<td>Data 6</td>
<td>12</td>
<td>10</td>
<td>2</td>
<td>83.30</td>
<td>1</td>
<td>90.90</td>
</tr>
<tr>
<td>Data 7</td>
<td>12</td>
<td>12</td>
<td>0</td>
<td>100</td>
<td>2</td>
<td>85.79</td>
</tr>
<tr>
<td>Data 8</td>
<td>12</td>
<td>12</td>
<td>0</td>
<td>100</td>
<td>1</td>
<td>92.30</td>
</tr>
<tr>
<td>Data 9</td>
<td>12</td>
<td>10</td>
<td>2</td>
<td>83.30</td>
<td>0</td>
<td>91.65</td>
</tr>
<tr>
<td>Data 10</td>
<td>12</td>
<td>10</td>
<td>2</td>
<td>83.30</td>
<td>1</td>
<td>90.90</td>
</tr>
<tr>
<td>Data 11</td>
<td>12</td>
<td>11</td>
<td>1</td>
<td>91.70</td>
<td>0</td>
<td>91.65</td>
</tr>
<tr>
<td>Data 12</td>
<td>12</td>
<td>12</td>
<td>0</td>
<td>100</td>
<td>2</td>
<td>85.79</td>
</tr>
<tr>
<td>Data 13</td>
<td>12</td>
<td>8</td>
<td>4</td>
<td>66.67</td>
<td>1</td>
<td>88.50</td>
</tr>
<tr>
<td>Data 14</td>
<td>12</td>
<td>12</td>
<td>0</td>
<td>100</td>
<td>1</td>
<td>92.30</td>
</tr>
<tr>
<td>Data 15</td>
<td>12</td>
<td>12</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>100.00</td>
</tr>
<tr>
<td>Data 16</td>
<td>12</td>
<td>12</td>
<td>0</td>
<td>100</td>
<td>2</td>
<td>85.79</td>
</tr>
</tbody>
</table>

**Table 2: Overall accuracy**

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>92.3906</td>
<td>92.1875</td>
<td>92.3938</td>
</tr>
</tbody>
</table>

As seen above high values of sensitivity, specificity and accuracy has been noticed when system tested for individual queries from the dataset.
In this speaker dependent data has been used, same algorithm can be tested for speaker independent data.

5. Conclusion

The basic of speaker recognition system is to decode the speech signal in a sequential manner based on the acoustic features of the signal between acoustic features and phonetic symbols. Here a database of speech signal of a number of people are used to train and gather some training data for future use of recognizing any specified person. There are number of feature extraction algorithms available for speech i.e. LPC’s, MFCC’s, GFCC’s, pitch, energy in which MFCC and GFCC are widely used as their effectiveness in unique feature extraction based on human audible frequencies. GFCC features found more robust in noisy signals and are proved more robust than the other. Hence GFCC features has been used in the proposed speaker identification system. As a long feature vector originates which takes time for training in classifier, an intermediate phase has been included to deduct the input feature space. Gaussian mixtures has been chosen for this which has property to cluster the data around individual Gaussians depending upon the input data and gives better features for classification phase. Neural network based classifier has been implemented for training and testing of the input features. It has been found that proposed system gives 92% accuracy in identification of speaker ID’s.

References:


