Gammachirp filterbank based speech analysis for speaker identification

MOUSLEM BOUCHAMEKH, BOUALEM BOUSSEKSOU, DAOUD BERKANI
Signal and Communication Laboratory – Electronics Department
National Polytechnics School,
10 Avenue Hacen BADI. 16200 Algiers,
ALGERIA.
mbouchamekh@gmail.com

Abstract: - Many modern speaker recognition systems use a bank of linear filters as the first step in performing frequency analysis of speech and extracting the acoustics parameters that allow characterizing the speaker identity. In this paper we illustrate the use of novel feature set extracted from speech signal. The new technique for extracting these parameters is based on the human auditory system characteristics and relies on the gammachirp filterbank to emulate asymmetric frequency response and level dependent frequency response. For evaluation a comparative study was operated with standard MFCC.

Key-Words: - Speaker identification, MFCC, Gammachirp, triangular filterbank

1 Introduction

Feature extraction is the key to the front-end process in speaker identification systems. The performance of the identification is highly dependent on the quality of the selected speech features. Most of the current proposed speaker identification systems use mel frequency cepstral coefficients (MFCC) and linear predictive cepstral coefficients (LPCC) as feature vectors. It is known that speech auditory frequency selectivity is largely determined by signal processing in cochlea [4, 5, and 7]. The basilar membrane inside the cochlea is usually conceived (in psychoacoustical auditory masking models) as a bank of band-pass filters that have increasing bandwidth. Irino and Patterson [4, 5] have developed a theoretically optimal auditory filter, the gammachirp, whose parameters can be chosen to fit observed physiological and psychoacoustical data.

In this work, a new approach for speech analysis based on gammachirp filters is shown. After extracting parameters we are interested to compare their performance with standard MFCC for text-independent speaker identification system, the evaluation is conducted on a database of 168 speakers extracted from TIMIT. Our speaker identification system is based on Gaussian Mixture Model (GMM) classifier [1].

2 The standard MFCC

The spectral based features of the Mel-Frequency Cepstral Coefficients have been proven to provide an accurate depiction of the spectral information of the human vocal tract. The Mel-Cepstral features are calculated by taking the cosine transform of the real logarithm of the short-term energy spectrum expressed on a mel-frequency scale.

After pre-emphasizing the speech using a first order high pass filter and windowing the speech segments using a Hamming window of 20 ms length with 10 ms overlap, the Discrete Fourier Transform is taken of these segments. The magnitude of the Fourier Transform is then passed into a filter bank comprising of twenty five triangular filters. The start and end points of these filters were calculated firstly by evenly spacing the triangular filters on the Mel-Scale and then using equation 1 to convert these values back to the linear scale.

$$Mel(f) = 2595. \log_{10} \left(1 - \frac{f}{700}\right)$$  (1)

The resulting filters used in our experiments are shown in fig.2.
Lastly, the Cepstral Coefficients were calculated from the log-energy outputs of these filters by the equation:

$$MFCC_i = \sum_{k=1}^{35} X_k \cos \left( i \left( k - \frac{1}{2} \right) \frac{\pi}{35} \right)$$  \hspace{1cm} (2)$$

Where \(i\) is the number of the coefficients and \(X_k\) is the log energy output of the \(k\)th filter.

### 3 Gammachirp filter

The gammachirp filter is a gamma distribution modulated at central frequency \(f_c\). It has an impulse response given by the following function [5]:

$$g_c(t) = \alpha t^{n-1} e^{-2\pi B(f_c)t} e^{-j(2\pi f_c t + \phi + c \log t + c \phi)}$$  \hspace{1cm} (3)$$

With \(\alpha > 0\)

\(n\): a parameter defining the order of the corresponding filter.

\(f_c\): The frequency of modulation of the function gamma.

\(\phi\): The initial phase.

\(\alpha\): Amplitude normalization parameter.

The term \(B(f_c) = b\). ERB \((f_c)\) characterizes the equivalent rectangular bandwidth (ERB) of the filter and \(b\) is a parameter defining the envelope of the gammachirp filter. The function \(ERB(f_c)\) is defined by the expression:

$$ERB(f_c) = 24.7 + 0.108.(f_c) \text{ Hz}$$  \hspace{1cm} (4)$$

\(c\): a factor introducing the asymmetry of this filter. Psychoacoustics studies show that \(c\) is strongly dependent on the signal power in the frequency bandwidth centered in \(f_c\).

$$c = 3.38 - 0.107 Ps$$  \hspace{1cm} (5)$$

With \(Ps\) is signal power.
5 Modeling by Gaussian Mixture Model (GMM)

In the speaker identification system under investigation, each speaker enrolled in the system is represented by a Gaussian mixture model (GMM). The idea of GMM is to use a series of Gaussian functions to represent the probability density of the feature vectors produced by each speaker. The mathematical representation is [1]:

\[ P(\tilde{x}|G_5) = \sum_{i=1}^{M} w_i G_i(\tilde{x} | \mu_i, \Sigma_i) \]  

(7)

Where \( M \) is the number of mixtures, \( \tilde{x} \) is the feature vector, \( w_i \) is the weight of the i-th mixture in the GMM, \( \mu_i \) is the mean of the i-th mixture in the GMM, and \( \Sigma_i \) is the covariance matrix of the i-th mixture in the GMM. The Model parameters \((w_i, \mu_i, \Sigma_i)\) characterize a speaker voice in the form of probability density function. They are determined by the Expectation maximization (EM) algorithm.

In the identification phase, the log-likelihood scores of the incoming sequence of feature vectors as subjected to each speaker model are calculated by:

\[ L(X, G_5) = \sum_{i=1}^{M} P(\tilde{x}_i|G_i) \]  

(8)

Where \( X = \{ \tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_T \} \) is the sequence of speaker feature vectors, and \( M \) is the total number of feature vectors. A GMM that generates the highest \( L(X, G_5) \) score is identified as the producer of the incoming speech signal. This decision method is called maximum likelihood (ML).

6 Experimental Evaluation

Three experiments have been conducted on 168 speakers database extracted from TIMIT, the first experience is conducted on original speech sampled at 16 KHz, and the last two experiments are conducted on downsampled version of speech at 8 KHz, the downsampling is done after filtering the speech in the band of [0 3400] Hz and applying a decimation of factor 2. The speech signal was extracted by using an energy based algorithm (the silences durations are excluded). The analysis of speech signal was conducted over speech frames of 20 ms with overlapping of 10 ms. In TIMIT, each speaker produces 10 sentences, 7 arbitrary sentences were used for training, and the last 3 sentences were used for testing, the average length of sentence is 3 seconds. In other word there was 21 seconds of speech for training and 9 seconds for 3 tests with 3 seconds for each test.

The classification engine used in this work was based on 32 mixtures GMM classifier initialized by vector quantization [10].

The results obtained in the first experiment are summarized in table 1.

<table>
<thead>
<tr>
<th>Number of coefficients</th>
<th>Mel triangular filterbank</th>
<th>Gammatone filterbank</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>68.85</td>
<td>86.87</td>
</tr>
<tr>
<td>4</td>
<td>92.06</td>
<td>90.48</td>
</tr>
<tr>
<td>6</td>
<td>96.43</td>
<td>96.21</td>
</tr>
</tbody>
</table>
As it is shown in figure 8, the identification rate is increasing with the coefficients number for the both parameters, standard MFCC and gammachirp filterbank based coefficients. We can also remark that for a number of coefficients lower than 12 the standard MFCC are slightly efficience, and contrarily for more than 12 coefficients.

In the second experiment, the speech database is filtered in the band of [0, 3400] Hz, and downsampled to 8 KHz by decimation. The identification results are summerized in table 2.

<table>
<thead>
<tr>
<th>Number of coefficients</th>
<th>Mel triangular filterbank</th>
<th>Gammachirp filterbank</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>49.41</td>
<td>51.19</td>
</tr>
<tr>
<td>4</td>
<td>83.53</td>
<td>84.33</td>
</tr>
<tr>
<td>6</td>
<td>91.47</td>
<td>91.27</td>
</tr>
<tr>
<td>8</td>
<td>95.24</td>
<td>95.64</td>
</tr>
<tr>
<td>10</td>
<td>97.62</td>
<td>97.02</td>
</tr>
<tr>
<td>12</td>
<td>97.42</td>
<td>98.02</td>
</tr>
<tr>
<td>14</td>
<td>98.61</td>
<td>98.02</td>
</tr>
<tr>
<td>16</td>
<td>98.41</td>
<td>98.21</td>
</tr>
<tr>
<td>18</td>
<td>98.21</td>
<td>98.02</td>
</tr>
<tr>
<td>20</td>
<td>97.02</td>
<td>98.41</td>
</tr>
</tbody>
</table>

Table 2 : Identification rate (%) obtained on 8 KHz database.

To evaluate the performances in case of presence of noise, in the third experiment we evaluate the speaker identification system on noisy database, our database is noised with additif white gaussian noise. The obtained results are summerized in the table 3.

<table>
<thead>
<tr>
<th>SNR (dB)</th>
<th>Mel filterbank</th>
<th>Gammachirp filterbank</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>81.27</td>
<td>49.60</td>
</tr>
<tr>
<td>5</td>
<td>83.53</td>
<td>83.14</td>
</tr>
<tr>
<td>10</td>
<td>93.25</td>
<td>92.46</td>
</tr>
<tr>
<td>15</td>
<td>94.84</td>
<td>95.04</td>
</tr>
<tr>
<td>20</td>
<td>96.43</td>
<td>96.63</td>
</tr>
<tr>
<td>25</td>
<td>96.53</td>
<td>96.63</td>
</tr>
<tr>
<td>30</td>
<td>96.63</td>
<td>98.21</td>
</tr>
</tbody>
</table>

Table 3 : Identification rates (%) for noisy speech.

The identification rate increases with speech quality, for higher signal to noise ratio we have higher identification rate, the gammachirp filterbank based parameters are slightly more efficience than standard
Conclusion
In this paper we have exposed a new method of speech analysis based on the human auditory system characteristics and rely on the gammachirp filter. The extracted coefficients are evaluated using GMM classifier, and compared with standard MFCC parameters for text independent speaker identification. The obtained results show that the new technique is very useful for noisy speech, and get more good rates than standard MFCC.

References