Text Independent Speaker Identification and Verification System Using Bessel Features

Mayank Gupta and Suryakanth V. Gangashetty

Abstract—In this paper, we explore the use of Bessel features derived from speech utterances, to develop Gaussian mixture speaker models for text independent Speaker Identification. The proposed approach to speaker identification is based on existing methods that employ Gaussian mixtures for the modeling of speakers. However, we have developed the speaker models from the Bessel features derived from the speech utterances, as an alternative to Mel-frequency cepstral coefficients for developing the speaker models. The proposed approach is tested on two databases of ten and twenty speakers respectively and their performance is evaluated. Finally, we have made some suggestions for future work involving the use of Bessel features for text independent speaker identification and verification.

Index Terms—Bessel functions, Gaussian mixture models, Text independent, Speaker identification.

I. INTRODUCTION

Speaker recognition may be defined as a process in which the identity of a person is established through his/her voice. The ability of a machine to correctly recognize the speaker can be put to various uses like access control systems, retrieval of sensitive information from a database, financial transactions on the telephone etc. Here we would like to emphasize on two closely related fields, namely speaker identification and speaker verification. Speaker verification concerns primarily with deciding whether a person is actually the one that he/she claims to be, from his/her voice sample. On the other hand, speaker identification basically attempts to determine the best possible match from a group of certain speakers, for any given input speech signal. Almost all speaker recognition schemes involve the collection of speech utterances from the speakers. The next step involves the extraction of features from these speech utterances, which are then used to develop models that can adequately capture speaker specific information. In general, a separate model is constructed for each speaker. The identification stage involves extracting features from the test utterance, evaluating the probability of the feature vectors to belong to the different speaker models, and finally deciding in favor of the speaker model that gives maximum probability.

Gaussian mixture speaker models have been widely used for speaker recognition and verification [1], [2], [3]. Speaker recognition by using Gaussian mixture speaker models involves developing the individual speaker models from training data set, essentially by using the Mel-Frequency Cepstral Coefficients (MFCC)[4] extracted from the training speech samples. In this paper, we propose using Bessel features [5] extracted from the training speech utterances to develop the Gaussian Mixture speaker models. Bessel function based expansion of speech has been used for speaker identification in [6], [7], [8]. The rest of the paper is organized as follows. Section 2 discusses the development of Bessel feature extraction from the speech utterances. Section 3 elaborates about the databases used in the experiments. Section 4 gives a brief overview of Gaussian mixture speaker models based speaker recognition systems. In Section 5, we have presented details of the proposed approach. Section 6 presents the studies on speaker identification and verification.

II. EXTRACTION OF BESSEL FEATURES

Bessel functions of first kind, arise as solutions to the wave equation inside cylindrical tubes [5], and can be used as basis functions to represent non stationary signals like speech signals [5], [9]. We can model the vocal tract as an organ pipe, which has cylindrical structure. In this representation, we can assume that there is a sound source at one end of the tube (the larynx or voice box) and the tube is open at other ends (the lips or nose). Thus there is a good motivation to choose Bessel functions of the first kind, given their naturalness, for representing the sounds produced in the vocal tract, which could be approximated as an acoustic tube for short-time intervals analysis [9]. In our work, we have used the zero-order Bessel series expansion as mentioned in [10], for representing speech signals. We may express such a signal by

\[ s(t) = \sum_{m=0}^{Q} C_m J_0(\frac{\lambda m}{a}t) \]  

(1)

where \( \{\lambda_m, m = 1, 2, 3,\ldots\} \) are the ascending order positive roots of \( J_0(\lambda) = 0 \) where as \( J_0(\frac{\lambda m}{a}t) \) are the zero order Bessel function. \( Q \) is the order of Bessel expansion. The coefficients \( C_m \), appearing in (1) are computed by using the following equation

\[ C_m = \frac{2 \int_0^a t x(t) J_0(\frac{\lambda m}{a}t) dt}{a^2 [J_1(\lambda m)]^2} \]  

(2)
We refer to these coefficients simply as the Bessel features of the signal \( s(t) \) in this network.

III. DEVELOPMENT OF THE SPEECH DATABASE

We constructed two databases of ten and twenty different speakers respectively. All the speech samples were recorded in laboratory conditions, and the same microphone was employed for all the recordings. Each speaker was asked to read random (and different) printed content for a minute. In the database of 10 speakers, there were 5 male speakers and 5 female speakers. Similarly, the 20 speaker’s database had 10 speakers of each of the two genders.

IV. GAUSSIAN MIXTURE SPEAKERS MODEL

The weighted sum of a certain number of component densities, \( M \), is used to represent a Gaussian mixture density [2]. We can denote such a mixture density by the equation

\[
p(\tilde{x} / \lambda) = \sum_{i=1}^{M} p_i b_i(\tilde{x})
\]

Here, is a D-dimensional random vector, \( i = 1, \ldots, M \) are the component densities, while \( p_i, i = 1, \ldots, M \) are the mixture weights. Each component density is a multivariate (in this case D-variate) Gaussian function of the form

\[
b_i(\tilde{x}) = \frac{1}{(2\pi)^{D/2} |\sum_i|^{1/2}} \exp\left\{ -\frac{1}{2} (\tilde{x} - \mu_i)'^{\top} \sum_i^{-1} (\tilde{x} - \mu_i) \right\}
\]

In (4), \( \mu_i \) is the mean vector, and \( \sum_i \) is the covariance matrix. Moreover, the mixture weights satisfy the constraint \( \sum_{i=1}^{M} p_i = 1 \).

Any Gaussian mixture density is completely represented by the mean vectors, covariance matrices and mixture weights of all component densities. We concisely represent such a density by the notation

\[
\lambda = \left\{ p_i, \mu_i, \sum_i \right\} i = 1, \ldots, M.
\]

When using GMM for speaker identification, we represent each speaker by a separate GMM, i.e. for any speaker \( s \), we have a model parameterized by. A discussion on the finer details regarding the choice of covariance matrices that can be used in the GMMs can be found in [2]. In this paper, we have restricted our approach to nodal [11], diagonal covariance matrices [2].

V. PROPOSED APPROACH FOR SPEAKER IDENTIFICATION

The following sub-sections describe the approach that we have used in our speaker recognition experiments. First, the extraction of the Bessel features for developing the Gaussian mixture speaker models is discussed. Then, details of how we carried out the recognition experiments are discussed.

A. Development of Gaussian Mixture Speaker Model From Bessel Feature

In the first stage of our experiment, we constructed Gaussian mixture speaker models for the ten speakers’ database. From the sixty seconds of speech of each speaker, we used the first 30 seconds for training purpose. First, 30 seconds of each speech utterance was split into frames of 20 milliseconds (320 samples), and the overlap between the successive frames was kept at 10 milliseconds (160 samples), we used a 320 point Hamming window for framing. Then, the Bessel features for each frame (that appear as \( C_m \) in equation (1)) were found out. We restricted the value of Q to 320 for each frame. Next, the Bessel features for each frame were arranged in descending order of magnitude, starting from the highest magnitude feature. Going on the lines of MFCC based Gaussian mixture speaker models as discussed in [2], we retained the first 12 highest magnitudes Bessel features from each frame for developing the speaker models. Each of the 10 individual speaker models were constructed using the Bessel features set derived from the first 30 seconds of speech of each speaker using the Expectation-Maximization (EM) algorithm [12]. To observe the effect of the model order M (i.e. the number of component densities) on the performance, we constructed speaker models for the same speaker using different values of M [2].

B. Speaker Identification from Gaussian Mixture Models

Now, we discuss the functioning of a Gaussian mixture speaker model based identifier as mentioned in [2]. Consider a group of S speakers \( S \). Each of the speakers is represented by his/her respective GMM: \( \lambda_1, \lambda_2, \ldots, \lambda_S \). To perform identification, the objective translates to finding the speaker model which has the maximum a posteriori probability for a given observation sequence. This is expressed as Assuming

\[
\hat{S} = \arg\max_{1 \leq i \leq S} Pr(\lambda_i / X)
\]

\[
= \arg\max_{1 \leq i \leq S} \frac{p(X \setminus \lambda_i) Pr(\lambda_i)}{p(X)}
\]

Assuming that the likelihood of different speakers are equal (i.e.) and taking note of the fact that p(X) is same for all speakers, (6) reduces to

\[
\hat{S} = \arg\max_{1 \leq i \leq S} \sum_{t=1}^{T} \log(X / \lambda_i)
\]

C. Speaker Identification Tests

Here we are using 30 sec speech which is not used in training. Then we can calculate the probability of each speaker. So the speaker having the least value is identified as speaker. It has also been observed that MFCC feature vector perform poorly in noisy environment. The following table shows the comparative study between the identification system using BFCC and MFCC.
TABLE I: COMPARATIVE STUDY OF SPEAKER IDENTIFICATION USING MFCC AND BFCC

<table>
<thead>
<tr>
<th>No. of mixtures</th>
<th>SNR</th>
<th>MFCC</th>
<th>BFCC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>17.2181</td>
<td>17.2284</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>36.7985</td>
<td>37.7428</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>72.0826</td>
<td>74.2124</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>91.6814</td>
<td>95.0896</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>96.38</td>
<td>96.9943</td>
</tr>
<tr>
<td>Clean Speech</td>
<td>97.3167</td>
<td>98.5489</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>10</td>
<td>16.9053</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>16.8369</td>
<td>29.7177</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>40.023</td>
<td>48.5974</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>77.7709</td>
<td>78.7652</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>94.8714</td>
<td>97.6353</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>99.6577</td>
<td>99.7308</td>
</tr>
<tr>
<td>Clean Speech</td>
<td>99.4933</td>
<td>99.8695</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>10.1536</td>
<td>15.0198</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>29.5847</td>
<td>31.1992</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>50.5937</td>
<td>67.184</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>85.709</td>
<td>93.2553</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Clean Speech</td>
<td>100</td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>

D. Speaker Verification System

For verification system we are using universal back ground model. For making this model we are taking 8 speakers. Out of which 4 are males and 4 are females. First we extract the feature from the given test speech signal. Next we calculate the probability for each of the speaker model and background model. Then we calculate the maximum likelihood ratio with the help of following formula

\[ \Lambda(X) = \log(X \mid \hat{\lambda}_{\text{typ}}) - \log p(X \mid \hat{\lambda}_{\text{typ}}) \]  

(8)

Afterwards, we set the threshold value experimentally to determine whether the speaker should be accepted or rejected. We observed that it is much easier to set the threshold value in case of BFCC than of MFCC.

VI. SUMMARY AND CONCLUSION

In this paper we have used Bessel Functions to calculate the Cepstral Coefficients. Through the various observations made, we infer that Bessel functions give more accurate results when compared to that of the MFCCs. However, to improve the accuracy, these Bessel Features with different modeling schemes other than Gaussian Mixtures can also be used.

REFERENCES