Text Independent Speaker Identification Using Bessel Features

Shivesh Ranjan, Viresh Ranjan, Chetana Prakash, 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**

***Index Terms*—Gaussian mixture models, Bessel functions, 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 involve 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].However,[6] uses a neural network based approach while [7] uses a Vector Quantization[8] based recognition approach, as compared to our Gaussian Mixture Models (GMMs)[1] based approach.

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.

II. Extraction of Bessel Features

Bessel functions of first kind, $J_n(x)$ arise as solutions to the wave equation inside cylindrical tubes [5], and can be used as basic 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. In this sense, our approach is different from [6],[7],[9] all of which have used $J_1(x)$,as the basis functions, as compared to $J_0(x)$ in our approach. Specifically, consider a speech signal s(t) defined over some arbitrary interval (0,a). We may express such a signal by

$$s(t) = \sum_{m=1}^{Q} C_m J_0\left(\frac{\lambda_m}{a} t\right) \qquad (1)$$

where $\{\lambda_m, \mathrm{m} = 1,2,3, \ldots\}$ are the ascending order positive roots of $J_0(\lambda) = 0$, whereas $J_0\left(\frac{\lambda_m}{a} t\right)$ are the Zero-Order Bessel functions. Q is the order of the Bessel expansion. The speech signal s(t) in (4) is represented as a linear combination of orthogonal Bessel functions. The orthogonality of Bessel functions $J_0(x)$, on the interval

Manuscript received September 12, 2012; revised October 23, 2012.

Shivesh Ranjan is with the Department of E.C.E., Manipal Institute of Technology, Manipal-576104, India (e-mail: vireshranjan@yahoo.co.in)

Viresh Ranjan is with the Department of E.C.E., Birla Institute of Technology, Mesra-835215, India (e-mail: chetana@research.iiit.ac.in)

Suryakanth V. Gangashetty is with International Institute of Information Technology, Hyderabad-500032, India (e-mail: svg@iiit.ac.in)

$0 \leq t \leq a$, with respect to the weight t is expressed by the following expression

$$\int_0^a t J_0\left(\frac{\lambda_m}{a}t\right) J_0\left(\frac{\lambda_n}{a}t\right) dt = 0, \quad for \quad m \neq n \qquad (2)$$

The coefficients C_m , appearing in (1) are computed by using the following equation

$$C_m = \frac{2\int_0^a tx(t)J_0\left((\lambda_m/a)t\right)dt}{a^2[J_1(\lambda_m)]^2} \qquad (3)$$

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

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 speakers database had 10 speakers of each of the two genders

IV. Gaussian Mixture Speaker Models

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

$$p(\overrightarrow{x}/\lambda) = \sum_{i=1}^{M} p_i \; b_i(\vec{x}) \qquad (4)$$

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

$$b_i(\overrightarrow{x}) = \frac{1}{(2\pi)^{\frac{D}{2}} \; |\sum_i|^{1/2}} exp\left\{-\frac{1}{2}(\vec{x} - \overrightarrow{\mu_i})' \sum_i^{-1} \; (\vec{x} - \mu_i)\right\} \qquad (5)$$

In (5), $\vec{\mu}_i$ 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

$$\lambda = \{p_i, \overrightarrow{\mu_i}, \sum_i\} \text{ i= 1, ..., M} \qquad (6)$$

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 λ_s . 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 Models from Bessel Features

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 Speaker 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={1,2, ...,*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

$$\tilde{S} = arg \quad max_{1\leq k\leq S} \Pr(\lambda_k/X)$$
$$= arg\, max_{1\leq k\leq S} \frac{p(X/\lambda_k)\Pr(\lambda_k)}{p(X)} \qquad (7)$$

Assuming that the likelihood of different speakers are equal (i.e.$Pr(\lambda_k) = 1/S$) and taking note of the fact that p(X) is same for all speakers, (7) reduces to

$$\tilde{S} = \arg max_{1\leq k\leq S} \; p(X/\lambda_k) \qquad (8)$$

We can exploit logarithms and the independence between observations [2], so that (8) translates to

$$\tilde{S} = \arg \; max_{1\leq k\leq S} \sum_{t=1}^{T} \log p(\overrightarrow{x_t}/\lambda_k) \qquad (9)$$

The term $p(\overrightarrow{x_t}/\lambda_k)$ appearing in (9) is computed from the expression given in (4).

C. Speaker Identification Tests

To test the performance of proposed speaker recognition system, we used the last 30 seconds of speech from each

speaker's recorded speech (the 1st 30 seconds were used to develop the Gaussian mixture speaker models).First we tested our approach on the database of 10 speakers. For this, we extracted the Bessel features from each of the 10 speech utterances, in the same way as discussed previously.

In the next stage, the a *posteriori* probabilities for each of the speaker models were found and a decision was made according to (9). This was repeated for all the 10 speakers of the first database (of 10 ten speakers).

The same sets of steps were repeated for the second database of 20 speakers. Our main motivation behind testing 2 different databases was to observe the change in performance of the recognition scheme due to an increase in number of speaker models.

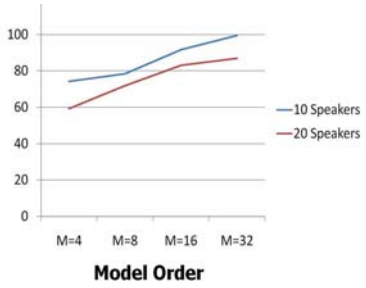

Fig. 1. Increase in performance with model order for the bessel features based speaker identification system.

TABLE I: RESULTS OF SPEAKER IDENTIFICATION EXPERIMENTS CARRIED OUT ON 10 AND 20 SPEAKERS DATABASE.

Model Order M	Identification Results (in %)	
	10 speakers	20 speakers
4	74.3	59.4
8	78.5	72
16	91.8	83.15
32	99.6	87

VI. STUDIES ON SPEAKER IDENTIFICATION

Rather than going for a single recognition experiment for each speaker, we evaluated our approach in the following manner, which is same as discussed in [2].The Bessel features from the test speech utterances were used to produce a sequence of feature vectors$\{\overrightarrow{x_1}, \overrightarrow{x_2}, \ldots \overrightarrow{x_t}\}$. We then constructed T length overlapping segments in the following way

Segment 1: $\{\overrightarrow{x_1}, \overrightarrow{x_2}, \ldots \overrightarrow{x_T}\}$
Segment 2: $\{\overrightarrow{x_2}, \overrightarrow{x_3}, \ldots \overrightarrow{x_{T+1}}\}$

In our experiments, for each segment, we chose T=500 feature vectors (where each vector was 12 Dimensional), and identification was performed on a total of 400 such segments.

Table I shows the success percentage for a total of 400 recognition attempts for each speaker, on the database of 10 speakers, and similar results for the database of 20 speakers . Fig. 1 shows the increase in performance with increase in model order M for both the databases, for the Bessel feature based Gaussian mixture speaker identification scheme. We will discuss the significance of these results in the next section, and also make some suggestions for future work.

VII. SUMMARY AND CONCLUSIONS

In this paper, we have proposed an approach for speaker identification using Bessel features. As evident from our results, the Bessel features can be successfully used for speaker recognition with reasonable accuracy. Here we would like to emphasize that as compared to the MFCCs which have been designed to exploit the nature of the human auditory perception, the Bessel features are rather 'raw' in the sense that we have not incorporated any use of the human auditory perception information while deriving them. An attempt has been made in [7] to incorporate this information into Bessel features for Speaker identification. However, the work in [7] was not able to give very impressive performance. An appropriate modeling of Bessel Features, taking in to account the human auditory perception, is yet to be developed. The readers are encouraged to explore this possibility.

Table I emphasizes that there is a decrease in performance with an increase in number of speakers in the database. It is observed from Figure 1 that there is a marked improvement in the performance of the Bessel features based speaker identification system with the increase in model order. This improvement in performance, with increase in Model Order, is similar to that obtained in MFCC based speaker recognition experiments [2].

The readers are also encouraged to compare the performance of the proposed approach against that using MFCC for speech utterances corrupted by noise.

REFERENCES

[1] D. R. O. Hart, E. Peter, and G. S. David, Pattern Classification, *John Wiley and Sons Asia Pte. Ltd*., Singapore, 2006
[2] A. R. D. Rose and C. Richard, "Robust text-independent speaker identification Using Gaussian mixture speaker models," *IEEE Trans. Speech and Audio Processing*, vol. 3, no. 1, pp.72-83, January 1995
[3] A. R. Douglas, "Speaker Identification and verification using Gaussian mixture speaker models," *Speech Communication*, vol. 17, no. 1, pp. 91-108, 1995.
[4] B. D. Steven and M. Paul, "Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357-366, 1980
[5] S. Jim, "Signal processing via Fourier-Bessel series expansion," *Digital Signal Processing*, vol. 3, pp.112-124, 1993
[6] K. Gopalan and T. R. Anderson, "Speaker identification using Bessel function representation and a back-propagation neural network," in *Proceedings of the IEEE International Symposium on Industrial Electronics*, vol. 1, pp. 381-383, 1995.
[7] K. Gopalan, R. A. T. Cupples, and J. Edward, "A comparison of speaker identification using features based on cepstrum and Fourier-Bessel expansion," *IEEE Trans. Speech and Audio Processing*, no. 3, pp. 289-294, May 1999
[8] R. Lawrence and B. H. Juang, "Fundamentals of Speech Recognition," *Pearson Education (Singapore) Pte. Ltd*., 2005
[9] F. S. Gurgen and C. S. Chen, "Speech enhancement by Fourier-Bessel coefficients of speech and noise," *IEEE Proceedings*, vol. 137, no. 5, pp.290-294, 1990
[10] P. R. Bilas, "Discrimination between ictal and seizure-free EEG signals using empirical mode decomposition," *Research Letters in Signal Processing*, pp.1-5, 2008.
[11] J. Oglesby and J. Mason, "Radial Basis Function Networks for Speaker Recognition," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 393-396, 1991
[12] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. 39, pp.1-38, 1977