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 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], [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.

Manuscript received August 18, 2012; revised October 8, 2012.

Mayank Gupta is with National Institute of Technology, Hamirpur (H.P.) India in Electronics & Communication Engineering, India (email- mayank01.gupta@gmail.com).

Suryakanth V. Gangashetty is with Language Technolgy Research Center in International Institute of Information Technology, Hyderabad, India. (email- svg@iiit.ac.in)

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

$$s(t) = \sum_{m=1}^{Q} C_m J_0(\frac{\lambda m}{a} t) \quad (1)$$

where $\{\lambda_m, m = 1, 2, 3...\}$ 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

$$C_m = \frac{2\int_0^a t x(t) J_0((\frac{\lambda m}{a} t))dt}{a^2 \left[J_2(\lambda m)\right]^2} \quad (2)$$

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

$$p(\vec{x}/\lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x}) \tag{3}$$

Here, 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

$$b_i(\overrightarrow{x}) = \frac{1}{(2\pi)^{\frac{D}{2}} \left|\sum i\right|^{1/2}} \exp\left\{-\frac{1}{2}(\vec{x}-\overrightarrow{\mu_l})'\sum_i^{-1}(\vec{x}-\mu_i)\right\} \tag{4}$$

In (4), μ_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

$$\lambda = \left\{p_i, \overrightarrow{\mu_i}, \sum i\right\} \quad i=1, \ldots, M. \tag{5}$$

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: $\lambda_1, \lambda_2, \ldots, \lambda_5$. 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

$$\hat{S} = \arg\max_{1\le k\le S} \Pr(\lambda_k / X)$$

$$= \arg\max_{1\le k\le S} \frac{p(X \setminus \lambda_k)\Pr(\lambda_k)}{p(X)} \tag{6}$$

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

$$\hat{S} = \arg\max_{1\le k\le S} \sum_{t=1}^{T} \log(X/\lambda_s) \tag{7}$$

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

No. of mixtures	SNR	MFCC	BFCC
2	0	10	10
	10	17.2181	17.2284
	20	36.7985	37.7428
	30	72.0826	74.2124
	40	91.6814	95.0096
	50	96.38	96.9943
	Clean Speech	97.3167	98.5489
4	0	10	16.9053
	10	16.8369	29.7177
	20	40.023	48.5974
	30	77.7709	78.7652
	40	94.8714	97.6353
	50	99.6577	99.7308
	Clean Speech	99.4933	99.8695
8	0	10.1536	15.0198
	10	29.5847	31.1992
	20	50.5937	67.184
	30	85.709	93.2553
	40	100	100
	50	100	100
	Clean Speech	100	100

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

$$\wedge(X) = \log(X \mid \lambda_{hyp}) - \log p(X \mid \lambda_{\overline{hyp}}) \qquad (8)$$

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

[1] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, John Wiley & Sons Asia Pte. Ltd., Singapore, 2006

[2] D. A. Reynolds and R. C. Rose, "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] D. A. Reynolds, "Speaker Identification and verification using Gaussian mixture speaker models," *Speech Communication* vol.17, no.1, pp. 91-108, 1995.

[4] S. B. Davis and P. Mermelstein, "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] J. Schroeder, "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, July 10-14, 1995

[7] K. Gopalan, T. R. Anderson, and E. J. Cupples, "A comparison of speaker identification using features based on cepstrum and Fourier-Bessel expansion," *IEEE Trans. Speech and Audio Processing*, vol. 7, no. 3, pp. 289-294, May 1999

[8] L. Rabiner 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, October 1990

[10] R. B. Pachori, "Discrimination between ictal and seizure-free EEG signals using empirical mode decomposition," *Research Letters in Signal Processing*, pp.1-5, Hindawi Publishing Corporation, January 2008

[11] J. Oglesby and J. Mason, "Radial Basis Function Networks for Speaker Recognition," *Proceedings of IEEE International Conference on Acoustics ,Speech, and Signal Processing*, pp. 393-396, May 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.

Mayank Gupta completed his Bachelors of Technology in Electronics & Communication Engineering from National Institute of Technology, Hamirpur (H.P.), India in May 2012. His major interests are in the field of signal processing, free space optical communication systems, biomedical signal processing and Embedded Systems. He has successfully completed many projects in these areas from the research institutes of great repute like Indian Institute of Science, Bangalore, International Institute of Information Technology, Hyderabad and National Institute of Technology, Hamirpur.

Dr. Suryakanth V. Gangashetty joined IIIT Hyderabad on 23rd August 2006 as an Assistant Professor. He completed his B.E (Computer Science and Engineering) from Govt. College of Engineering Davangere in 1991, M.Tech (Systems Analysis and Computer Applications) from REC Surathkal in 1998 and Ph.D (Neural Network Models for Recognition of Consonant-Vowel Units of Speech in Multiple Languages) from IIT Madras in 2005. He did his post-doctoral studies at Carnegie Mellon University, Pittsburgh during April 2007 to July 2008. He is the author of about 90 papers which have been published in national as well as international conferences and journals. He has co-authored four book chapters in edited volumes published by Springer and World Scientific publishing company. His research interests include Speech Processing, Neural Networks, Multimedia Signal Processing, Pattern Recognition, Soft Computing, Machine Learning, Image Processing, Natural Language Processing, Artificial Intelligence and Fuzzy Logic.