Features Based on Fourier-Bessel Expansion for Application of Speaker Identification System

Saransh Chhabra[1], Ronak Bajaj[1], R. B. Pachori[2], R. N. Biswas[3]

[1]CVEST, International Institute of Information Technology, Hyderabad, India
[2]Department of Electrical Engineering, Indian Institute of Technology, Indore, India
[3]Mentor Professor, NIIT University, Neemrana, Rajasthan, India

Abstract

A compact representation of speech is possible using Bessel functions because of the similarity between voiced speech and the Bessel functions. Both voiced speech and the Bessel functions exhibit quasi-periodicity and decaying amplitude with time. In this paper, we have developed various feature extraction techniques using zero-order Bessel functions as basis functions for the task of closed-set text-independent speaker identification system. The features are tested on TIMIT, CHAINS and IIIT-Hyderabad speech databases. The performance of the proposed feature extraction techniques is compared with the results obtained using MFCC features. A generic Gaussian Mixture Model (GMM) classification system is used for speaker modeling. The proposed extraction techniques provide results comparable to the widely used MFCC.

1 Introduction

Speaker identification is a decision making process of who of the registered speakers is most likely the author of the unknown speech sample. The system is usually composed of a feature extraction module and a statistical modeling module. Feature extraction module extracts useful features from speech signal and these are given as input to the modeling module which computes a statistical model from those features.

Recent advances in using AM-FM modelling of speech signals [1], [2], [3] inspires the use of AM-FM speech modelling for the task of speaker identification [4]. Separation of speech resonances using FB expansion without band-pass filtering has been done in [1]. The purpose of this paper is to propose new feature extraction techniques based on FB expansion and AM-FM modelling of speech signals and evaluate their performance for the task of speaker identification.

The rest of the paper is organized as follows: Section 2 describes the proposed feature extraction techniques. In Section 3 and 4, we describe the rest of the features/parts of the SID system. Results are given in Section 5. We conclude the paper in Section 6.

2 Feature Extraction Techniques

The zero-order FB series expansion of a discrete-time signal x[n] considered over some arbitrary interval [0, N] is expressed as:

$$x[n] = \sum_{m=1}^{M} C_m J_0\left(l_m n/N \right) \tag{1}$$

where $J_0(.)$ are the zero-order Bessel functions. The coefficients C_m are computed by relation,

$$C_m = \frac{2\sum_{n=0}^{N} nx(n) J_0(l_m n/N)}{N^2 (J_1(l_m))^2} \tag{2}$$

where $J_1(.)$ are the first-order Bessel functions, and $\{\lambda_m: m=1,\ldots,M\}$ are the ascending order positive roots of $J_0(\lambda)=0$.

Instantaneous Parseval energy (E_m) and instantaneous frequency (f_m) corresponding to each FB coefficient is obtained using relations,

$$E_m = C_m^2 \frac{N^2}{2} [J_1(l_m)]^2 \tag{3}$$

$$f_m = \frac{l_m}{2pN} \tag{4}$$

, where *N* is the number of sample points in a frame.

There is a one-to-one correspondence between the frequency *f* of signal and the order, *m* of FB coefficient

$\frac{f}{\left(f_s/2 \right)} = \frac{m}{n}$, where *N* is the number of sample points in a frame.

Based on the value *m* and the frequency it corresponds, FB coefficients are divided in segments such that each segment has FB coefficients capturing the desired frequency range of original signal. We call this process, **segmentation**.

In this paper, we focus on developing different feature extraction techniques and their experimental evaluation. We use speech data sampled at a frequency of 16000 Hz.

2.1 Features Based on FB Coefficients (C_m)

An energy measure in a narrow band of frequencies in each segment is obtained. The magnitude of the basis signal amplitude at each indices of the segment are added to get the energy measure as,

$$e_i = \sum_{m=m_{i1}}^{m_{i2}} |C_m| \quad i = 1,2,...,26 \tag{5}$$

where, m_{i1} is the starting value of m in i^{th} segment, m_{i2} is the last value of m in i^{th} segment.

Each e_i represents approximately the energy measure in each segment. These feature vectors are analogous to the log energy output of the band-pass filters in the Mel-cepstral domain representation.

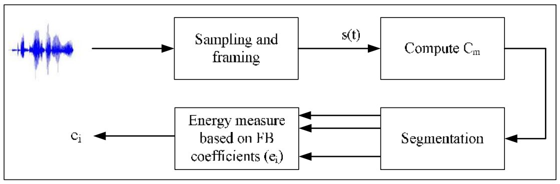

Figure 1. Feature extraction based on FB coefficients.

2.2 Features Based on FB Coefficients Energy (E_m)

Human auditory system perceives information based on the energy in a band of frequencies rather than that at a single frequency. Motivated by this fact we use the Parseval energy (E) for each segment as features.

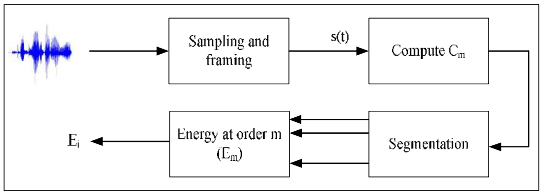

Figure 2. Feature extraction based on FB coefficients' energy.

E in the narrow band of frequencies in each segment is obtained using (6) and a vector of 26 such energy measures is used as a feature vector.

$$E_i = \sum_{m=m_{i1}}^{m_{i2}} C_m^2 \frac{N^2}{2} [J_1(l_m)]^2 \quad ,i = 1,2,...,26 \tag{6}$$

where, m_{i1} is the starting value of m in i^{th} segment, m_{i2} is the last value of m in i^{th} segment.

2.3 Features based on FB expansion and DESA [1] method (F_i)

FB coefficients in each segment are used to obtain band-pass signals using FB expansion (1), so we have 26 waveforms. Since each waveform captures a narrow band of approximately 106 Mel, we assume presence of at most one formant in each segment. As we model each formant as an AM-FM signal, DESA is applied on each waveform to get its amplitude envelope (AE_i) and instantaneous frequency (IF_i).

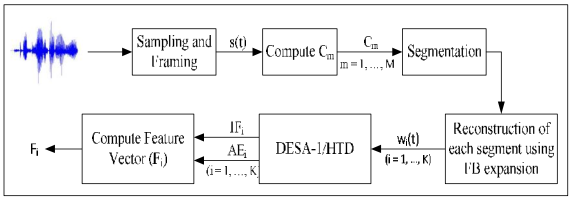

Figure 3. Feature extraction based on FB expansion and DESA

Instantaneous frequency and amplitude envelope are combined together to obtain a mean-amplitude-weighted short-time estimate (F_i) of the instantaneous frequency for each reconstructed waveform

$$F_i = \frac{\sum_{m=1}^{N} IF_i(m) AE_i^2(m)}{\sum_{m=1}^{N} AE_i^2(m)} \quad ,i = 1,2,...,26 \tag{7}$$

The adoption of a mean amplitude weighted instantaneous frequency (7) is motivated by the fact that it provides more accurate frequency estimates and is more robust for low energy and noisy frequency bands when compared with an unweighted frequency mean.

2.4 Features based on mean frequency (f_{mean})

E_m and f_m for each FB coefficients in each segment is calculated. E_m and f_m are combined together to obtain a mean-energy weighted short-time estimate, f_{mean} of the instantaneous frequency for each segment.

$$f_{mean} = \frac{\sum_{m1}^{m2} f_m E_m}{\sum_{m1}^{m2} E_m} \tag{8}$$

, N is number of sample points in a frame.

26 dimensional f_{mean} is used as a feature vector.

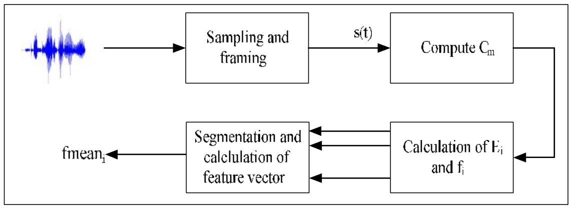
Sampling and framing
s(t)
Compute C_m
Segmentation and calclulation of feature vector
Calculation of E_i and f_i
$fmean_i$

Figure 4. Feature extraction based on mean frequency Note is that the figure is centred.

The information extracted from speech signal using f_{mean} is similar to the FB-DESA (F_i) technique. However, this technique requires less computation. No reconstruction of signal from FBC is required.

3 Gaussian Mixture Speaker Model

The above set of feature vectors were evaluated using Gaussian mixture model (GMM) classifier consisting 32 mixtures. The GMM, like a parametric model has structure and parameters that control the behaviour of the density in known ways, but without the constraint that the data must follow a specific distribution [5]. We have used a random mean selection, followed by a single iteration k-means clustering for initialization. With the same classifier used on all the features and for all the databases, the effectiveness of each feature extraction technique can be compared.

4 Databases

The experiments were evaluated on the TIMIT database, the CHAINS corpus and IIIT-Hyderabad database.

TIMIT contains a total of 6300 sentences, 10 sentences spoken by each of 630 speakers. A set of 22 speakers is taken from database. Out of 10 sentences, we selected 7 for training and 3 for testing.

The CHAracterizing Individual Speakers (CHAINS) corpus [6] contains recordings of 36 speakers obtained in two different sessions with a time separation of about two months. For our experiments, we use the SOLO condition database recorded in first session in a sound proof environment as the training data while the FAST condition data recorded in second session in a quite office environment is used as testing data.

IIIT-Hyderabad database (prepared at IIIT-H) contains 24 speakers, 33 sentences spoken by each speaker in a normal office environment using PHILIPS SHM3100 headphones with cable length of 2 meters.

5 Results

To explore the performance, speaker identification system using each of the proposed feature extraction techniques are implemented on MATLAB. SID system using widely used MFCC is also implemented and used as a base for comparison. Tests are carried out on all the three databases. Length of training and testing material used is 60 sec and 15 sec respectively. Results obtained are summarized in table 1.

S. No.	Feature Extraction Technique	TIMIT (22 Speakers)	CHAINS (16 Speakers)	IIIT-H (24 Speakers)
1	Based on MFCC	22	14	22
2	Based on FB Coeff.	22	13	22
3	Based on FB Coeff. Energy	22	12	21
4	Based on FB-DESA technique	22	12	22
5	Based on mean frequency	22	14	24

Table 1 – Performance of SID system.

6 Conclusion

Our experimental evaluation indicates that the features based on FB expansion gives comparable results to the widely used technique MFCC, with great reduction in computational complexity. Because of the decaying behavior of Bessel functions, which are more analogues to speech signal, results to better representation of the signal and less number of coefficients are required.

Out of the four proposed techniques, the technique in which mean-energy weighted short-time estimate of frequency are used as feature vectors gives best results.

7 References

[1] R. B. Pachori and P. Sircar, “Speech analysis using Fourier-Bessel expansion and discrete energy separation algorithm,” Proc. IEEE Digital Signal Processing Workshop, and workshop on Signal Processing Education, pp. 423-428, 24-27 Sept. 2006, Wyoming, US.

[2] R. B. Pachori and P. Sircar, “Analysis of multicomponent AM-FM signals using FB-DESA method”, Digital Signal Processing, vol. 20, pp 42-62, 2010.

[3] R. B. Pachori, “Discrimination between ictal and seizure-free EEG signals using empirical mode decomposition,” *Research Letters in Signal Processing*, vol. 2008, article id: 293056, 2008.

[4] M. Grimaldi and F. Cummins, “Speaker identification using instantaneous frequencies,” *Proc. IEEE Trans. On Audio, Speech, and Language Processing*, vol. 16, no. 6, Aug 2008.

[5] D. A. Reynolds and R. C. Rose, “Robust text-independent speaker identification using Gaussian mixture speaker models,” IEEE Trans. on Speech and Audio Processing, vol. 3, no. 1, Jan. 1995.

[6] Cummins F, Grimaldi M, Leonard T, Simko J (2006) The CHAINS corpus: CHAracterizing Individual Speakers. In: Proceedings of SPECOM’06, pp 431–435.