Engineering Science and Technology, an International Journal 17 (2014) 145–151

HOSTED BY

ELSEVIER

Contents lists available at ScienceDirect

Engineering Science and Technology, an International Journal

journal homepage: http://www.elsevier.com/locate/jestch

Full length article

Auditory ERB like admissible wavelet packet features for TIMIT phoneme recognition

P.K. Sahu [a], Astik Biswas [a,*], Anirban Bhowmick [b], Mahesh Chandra [b]

[a] Dept. of Electrical Engineering, National Institute of Technology, Rourkela, India
[b] Dept. of ECE, Birla Institute of Technology, Mesra, Ranchi, India

ARTICLE INFO

Article history:
Received 15 December 2013
Received in revised form
24 April 2014
Accepted 24 April 2014
Available online 28 May 2014

Keywords:
Speech recognition
Wavelet packets
ERB scale
WERBC
WMFCC
Phoneme recognition

ABSTRACT

In recent years wavelet transform has been found to be an effective tool for time–frequency analysis. Wavelet transform has been used as feature extraction in speech recognition applications and it has proved to be an effective technique for unvoiced phoneme classification. In this paper a new filter structure using admissible wavelet packet is analyzed for English phoneme recognition. These filters have the benefit of having frequency bands spacing similar to the auditory Equivalent Rectangular Bandwidth (ERB) scale. Central frequencies of ERB scale are equally distributed along the frequency response of human cochlea. A new sets of features are derived using wavelet packet transform's multi-resolution capabilities and found to be better than conventional features for unvoiced phoneme problems. Some of the noises from NOISEX-92 database has been used for preparing the artificial noisy database to test the robustness of wavelet based features.

1. Introduction

Speech as a medium of human to machine or machine to machine communication has been gaining popularity since the last few decades. Artificial intelligence cannot cultivate significantly without the improvement of automatic speech recognition. Most of the systems developed till now are based on the frequency domain analysis of the speech signal in a laboratory environment. However, speech recognition accuracy still degrades significantly in adverse real time situation and sensor mismatch conditions.

Automatic Speech Recognition (ASR) system comprises front end processing and back end processing. Front end encompasses various feature extraction and noise compensation techniques. Back end have different types of acoustic, language and pronunciation. Feature extraction is a technique of extracting optimum maximal information from a phoneme which gives maximum discrimination between phoneme classes. Feature extraction technique should be robust enough to perform well in different environmental conditions as well as sensor mismatch conditions. Apart from wavelet based feature extraction techniques some of the commonly used feature extraction techniques are Mel Frequency Cepstral Coefficients (MFCCs) [1], Linear Prediction based Cepstral Coefficients (LPCCs) [2], Gammatone Feature Cepstral Coeffficients (GFCC) [3,4], perceptual linear prediction [5]. Feature extraction techniques should be preceded by Fourier Transform (FT) in order to obtain its speech spectrum. Having a uniform resolution over the frequency plane windowed FT or the Short Time Fourier Transform (STFT) technique is not suitable to recognize some of the phonemes such as stops. It is difficult to detect a short event like burst in a slowly time varying signal by using STFT technique. To overcome this problem, wavelet packets (WPs) and local cosine transforms have helped in feature extraction [6–8].

Wavelet Packets (WPs) [9–11] are considered to have important signal representation schemes impacting compression, detection and classification [12,13]. WPs are extensively used in the analysis of pseudo-stationary time series processes and quasi-periodic random fields, such as the acoustic speech process [14,15]. WPs can be used effectively to describe a rich coverage of signal-space decomposition as well as providing a way for generating sub-band dependent partitions of the observation space. In conclusion, WPs induce a family of structural filter-banks with rich coverage of time–frequency characteristics that has the potential for enriching the way conventional MFCC features describe the short term behavior of the acoustic speech process.

WPs and multi-rate filter bank analysis have been adopted to improve the performance of conventional features by dividing the

* Corresponding author.
E-mail addresses: pksahu@nitrkl.ac.in (P.K. Sahu), astikbiswas@live.com (A. Biswas), anirban.bhowmick@outlook.com (A. Bhowmick), shrotriya69@rediffmail.com (M. Chandra).
Peer review under responsibility of Karabuk University.

http://dx.doi.org/10.1016/j.jestch.2014.04.004

frequency axis analogs to MEL scale frequency resolution in the context of ASR [7,16–19]. They used the Daubechies (db) two channel filter (TCF) which is reported to enhance the recognition performance for specific phone subcategories (stops and unvoiced speech) in a portion of the TIMIT. Choueiter and Glass (2007) [15] explored the problem of two-channel filter-bank design and they proposed the novel framework of rational filter-banks. Main focus of this work was to improve the frequency selectivity with respect to the conventionally adopted Daubechies WPs by designing a type of MEL frequency filter-bank structure. Improved performances were achieved in a simplified phone-segmented classification task with respect to MFCCs. Farooq et al. (2010) [17] used wavelet transform-based feature extraction technique by taking into account temporal as well as frequency band energy variations for Hindi phoneme recognition. This feature extraction technique performed better than MFCC features in a simplified phone classification. Litvin and Cohen (2011) [19] have shown that wavelet based bark scale aligned WP decomposition improves the performance of single-channel source separation of audio signals. Recently Pavez and Silva (2012) [18] have shown that wavelet based wavelet Packet Cepstral Coefficients (WPCC's) have shown concrete results that complement the previous work on supporting the use of WPs as a feature extraction techniques for ASR.

In this paper WP based features are wavelet based, in which the frequency axis is divided analogs to the Equivalent Rectangular Bandwidth (ERB) [20] scale frequency resolution. This ERB scale was originally designed to model human cochlear filtering [21]. ERB scale frequency resolution can be used to approximate center frequencies and the bandwidth of each Gammatone filter in GFCC. Frequency axis has been divided according to ERB scale to follow the response of human cochlea. In this paper it has been tried to take the advantage of auditory ERB filter-bank as well as WP to extract the coefficients at a certain frequency of interest. This technique attempts to reduce the articulation effect in the phoneme features. Recently we have shown the effectiveness of these ERB features for Hindi consonant recognition applications [22]. The performance of this feature technique have been tested with TIMIT database. Further, these features have proved more robust in presence of babble, volvo, factory and white noises. The performance of the wavelet based feature is compared with wavelet like MFCC(WMFCC) [8,17], MFCC and GFCC.

The rest of the article is organized as follows: Section 2 gives brief overview of wavelet based feature extraction technique. Section 3 provides brief overview of TIMIT database. Section 4 covers the details of experiments performed and result obtained for phone recognition task. Finally, the conclusions of the experiment are drawn in Section 5.

2. ERB like WP decomposition and feature extraction

Refs. [11,16] can be referred for detail description of wavelet analysis. The 24 sub-band wavelet packet tree is derived which approximate the ERB scale division as shown in Fig. 1. The

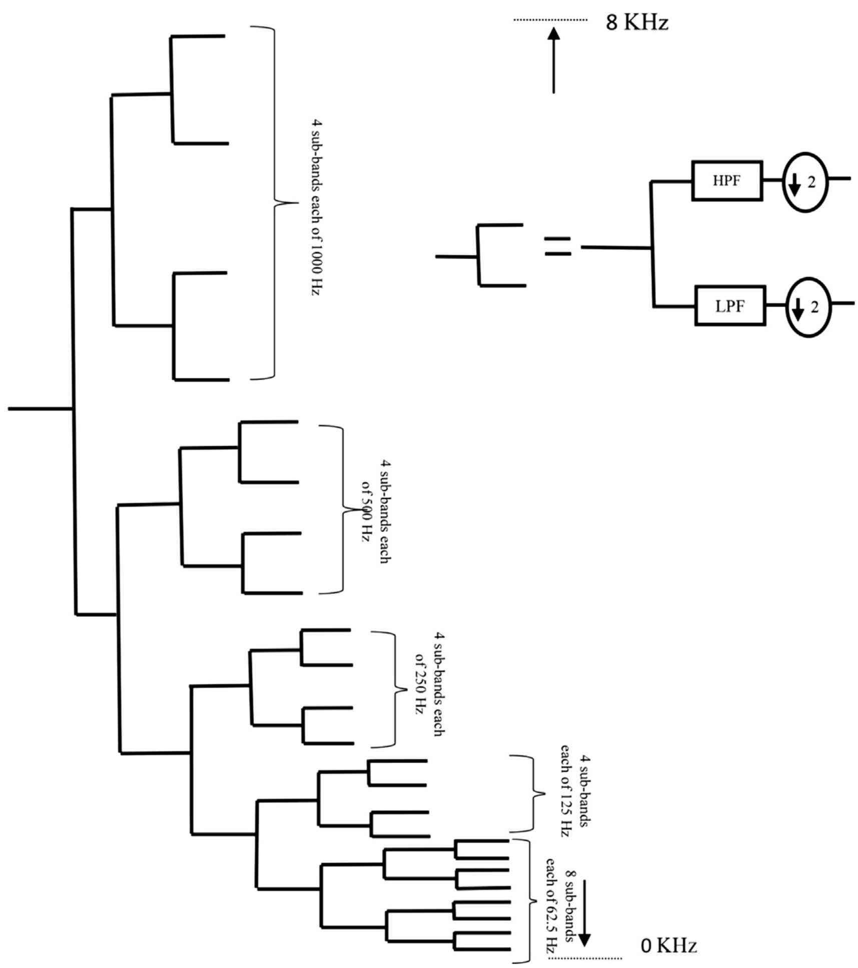

Fig. 1. 24 sub-band wavelet packet tree based on ERB scale.