Engineering Science and Technology, an International Journal xxx (2018) xxx–xxx

HOSTED BY

ELSEVIER

Contents lists available at ScienceDirect

Engineering Science and Technology, an International Journal

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

Full Length Article

Performance evaluation of Hindi speech recognition system using optimized filterbanks

Mohit Dua *, Rajesh Kumar Aggarwal, Mantosh Biswas

Department of Computer Engineering, National Institute of Technology, Kurukshetra, India

ARTICLE INFO

Article history:
Received 27 December 2017
Revised 5 April 2018
Accepted 9 April 2018
Available online xxxx

Keywords:
Automatic speech recognition
MFCC
GFCC
BFCC
Differential evolution

ABSTRACT

An Automatic Speech Recognition (ASR) system implementation uses a conventional pattern recognition technique that stores a set of training patterns in classes and compares the test patterns with training patterns to place them in the best matched pattern class. Most state-of-the-art ASR systems use Mel Frequency Cepstral Coefficient (MFCC) and Perceptual Linear Prediction (PLP) to extract features in training phase of the ASR system. However, sensitivity of MFCC & PLP to background noise has resulted in use of noise robust features Gammatone Frequency Cepstral Coefficient (GFCC) and Basilar-membrane Frequency-band Cepstral Coefficient (BFCC). But many issues associated with these feature extraction methods, like accepted bandwidth and standard number of filters are unresolved till date. This paper proposes a novel approach to use Differential Evolution (DE) algorithm to optimize the number and spacing of filters used in MFCC, GFCC and BFCC techniques. It also evaluates the performance of the said feature extraction methods with and without DE optimization in clean as well as in noisy environments. The results conclude that BFCC based ASR systems performs 0.4% to 1.0% better than GFCC and 7% to 10% better than MFCC in different conditions.

1. Introduction

Speech is used by the human beings as fundamental mean of communication. However, with advances in research it is being used for communicating with machines also. An ASR system takes speech signal as an input and gives text as an output. In the last six decades, many techniques have been proposed to make this speech to text conversion accurate and efficient independent of speaker, device or the environment [1–4]. However, noise intrusion in the speech signals make it tough for scientists to implement a standard ASR system [5]. Intelligibility of the speech signal decreases due to this noise intrusion.

Besides acoustical model adaptation and robust feature extraction approaches, noise reduction is considered as an effective approach for robust ASR system [5]. Speech enhancement is one of the approach, that is used to remove the noise by estimating the noise characteristics. However, speech enhancement algorithms can enhance the quality of input speech signal but not speech intelligibility. Speech intelligibility is computed in the presence of distortion and different types of noises. For the implementation of most of the speech enhancement algorithms exact measurement of spectrum of background noise is required which is practically not possible to compute. Most of the noise estimation algorithms are also not able to evaluate error-free or detailed noise spectrum. It is seen that sometimes the speech signal gets corrupted and also, speech signal similar to noise gets discarded while implementing such algorithms.

Noise robust feature extraction techniques also play a crucial role in implementation of error-free ASR systems [6,7]. An ideal feature vector should carry accurate information from the recorded speech signal and should be robust against noise. Hence, the development of noise robust feature extraction methods has been area of prime research in ASR over the last five decades [8–10]. Linear predictive cepstral coefficients (LPCC) [11], Temporal Patterns (TRAPs) [12], MFCC [13], Perceptual linear prediction (PLP) [14] and wavelets [15] are some of the feature extraction methods proposed by scientists in the last sixty years [16]. Out of these proposed approaches, MFCC is known to be the most accurate for speech recognition systems. MFCC has low computation overhead and performs quite well in clean environments. However, it does not perform well in the presence of additive noise. Hence, lack of robustness of MFCC to background noise has resulted in use of noise robust features GFCC [8] and BFCC [9,17]. All the three

* Corresponding author.
E-mail addresses: er.mohitdua@nitkkr.ac.in (M. Dua), mantoshbiswas@nitkkr.ac.in (M. Biswas).

Peer review under responsibility of Karabuk University.

https://doi.org/10.1016/j.jestch.2018.04.005
2215-0986/

Please cite this article in press as: M. Dua et al., Performance evaluation of Hindi speech recognition system using optimized filterbanks, Eng. Sci. Tech., Int. J. (2018), https://doi.org/10.1016/j.jestch.2018.04.005

techniques use different filterbanks for feature extraction. MFCC uses Mel-scale filter-bank [13], GFCC uses Gammatone filter-bank [18] and BFCC uses Gammachirp filter-bank [18], where Gammachirp is an extension of Gammatone filter [9,17,18]. GFCC is designed to simulate the process of human hearing system as it uses bank of non-linear filters while performing analysis of speech signal [18]. However, MFCC uses bank of linear filters during speech signal analysis. Hence, researchers have exploited the GFCC and BFCC features' noise characteristics to develop ASR systems [8,9,17,19–22].

Although many noise robust feature extraction methods have been proposed by researchers in the last few decades, issues like accepted bandwidth and standard number of filters are still matter of research till date. Hence, with development of new feature extraction methods, efforts have also been made to optimize the feature extraction techniques. In [23], the authors proposed an evolution strategy (ES) [24] to optimize two complementary filter banks (CFB) based feature extractors and proved that the proposed system provides optimal cepstral representation for speaker verification than the traditional Liner frequency cepstral coefficients (LFCC) or MFCC feature extraction methods. Similar to works in [23], the authors of [25] performed speaker-specific filterbank optimization for text-independent speaker verification by applying the Artificial Bee Colony (ABC) algorithm [26]. Here also, the proposed method outperforms the conventional Mel and linear scales based methods. In [27], optimization of adaptive bands filter bank (ABFB) is proposed by using genetic algorithm (GA) to optimize its design parameters for robust speech recognition. The result analysis of the proposed work shows that the optimized ABFB performs significantly better than the classical Bark-scale filter bank. A novel framework for Genetic algorithms (GA) based MFCC filterbank optimization has been proposed by the authors of [28] for speaker diarization. The authors in [29] optimized the filter-bank of the MFCC features by applying the Principal Component Analysis (PCA) approach to improve the recognition accuracy in noisy environments. Later, one of the authors from the same group applied modified PCA and Linear Discriminant Analysis (LDA) to solve the problem of negative filter coefficients in his work proposed in [30]. The experimental evaluations prove that the proposed novel filter-banks show better performance in clean as well as in noisy environments. Recently, some work has been done to refine the features by applying different optimization approaches like Particle swarm optimization (PSO) [31], Differential evaluation (DE) [32] and Genetic algorithm (GA) [10,33,34]. All these methods are population based search techniques. The particles in the population change their values by observing the values of other particles of the population. Researchers have used GA and PSO in ASR for optimization of feature vector [10]. GA uses selection, crossover, and mutation for optimizing features, whereas PSO does population initialization, calculates individual best and global best and updates swarm for optimization. GA and PSO both have some strengths and weaknesses. The researchers have also used PSO after including a crossover operator in it [35,36]. However, it has been shown that DE outperforms GA in against convergence speed and optimal quality parameters [10,37].

Therefore, the proposed work mainly uses DE algorithm to optimize the MFCC, GFCC and BFCC features to enhance performance of the HINDI language ASR system. Initially, the performance evaluation of all three-feature extraction methods is done in clean and noisy environments. The results show that GFCC and BFCC outperform MFCC in noisy environments. Further, these feature vectors are optimized using DE optimization method. The results describe that DE optimized BFCC features show significant improvement over MFCC and GFCC features. The developed ASR system uses Hidden Markov Model Tool Kit (HTK) [38] 3.5 beta-2 version and MATLAB version 15 for its implementation.

The remaining part of the paper is organized as: Section 2 briefly describes the fundamentals of feature extraction, filterbanks and optimization methods. Section 3 gives details of the proposed architecture, Section 4 deals with some details of the Hindi language and speech corpus, Section 5 gives the simulation and experiment analysis, and Section 6 concludes the proposal.

2. Feature extraction methods & differential evolution

2.1. Filterbank and feature extraction

With the ever changing technology and research methods, the speech recognition lies on the frontier of filter bank to optimize the ASR system and produce the efficient output. The objective of feature extraction is to detect a set of variables from the speech signal that are correlated acoustically. Such variables are termed as features. Feature extraction removes unwanted and redundant information. The proposed work uses three feature extraction techniques MFCC, GFCC and BFCC. This section describes the fundamentals of these feature extraction methods and filterbank used by these methods.

2.1.1. Mel filterbank and MFCC

Researchers have been using MFCC as an established and proven method to extract distinct characteristics of input speech signal [5]. MFCC uses some parts of speech production and speech perception to extract a feature vector that contains all information about the speech signal. The process for MFCC feature extraction includes following steps:

- Pre-emphasis of input speech signal is performed to amplify the energy at high frequencies [39]. It not only, reduces the difference in power components of the signal and but also distributes power across the relative frequencies. As a result, the high frequencies are more prevalent in the pre-emphasized signal. The input signal is divided into frames which contain arbitrary number of samples. Each time frame is then distributed in different Hamming window to eliminate discontinuities at the edges. The operation is performed using Eq. (1).

$$W(n) = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{N} - 1\right) & 0 \leqslant n \leqslant N-1 \\ 0 & Otherwise \end{cases} \quad (1)$$

  where $W(n)$ is Hamming window, N denotes the total number of samples, n refers to the current sample.
- After windowing the Discrete Fourier Transform (DFT) is applied to segregate the energy comprised into each frequency band. FFT is calculated for each frame to extract frequency components of the input speech signal. This is achieved by reckoning the discrete Fourier transform given by Eq. (2).

$$f_{t,i,0} = \left| \frac{1}{N} \sum_{k=1}^{N-1} \left(e^{-j2\pi \frac{ki}{N}} \right) f_k \right| \quad (2)$$

  where $i = 0, 1, 2, \ldots, (N/2) - 1$, t is the time frame and N is the number of sampling points within a time frame t.
  The spectrum obtained by discrete Fourier transform is filtered with different band pass filter and the power of individual frequency band is enumerated. This is needed to estimate the power spectrum. The enumeration of the spectrum band is as follows:

$$f_{t,k,1} = \sum_{i=0}^{\frac{N}{2}-1} c_{k,i} f_{t,i,0} \quad (3)$$

Please cite this article in press as: M. Dua et al., Performance evaluation of Hindi speech recognition system using optimized filterbanks, Eng. Sci. Tech., Int. J. (2018), https://doi.org/10.1016/j.jestch.2018.04.005