Computers and Electrical Engineering 38 (2012) 1467–1479

Contents lists available at SciVerse ScienceDirect

Computers and Electrical Engineering

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

Average framing linear prediction coding with wavelet transform for text-independent speaker identification system ☆

Khaled Daqrouq [a,*], Khalooq Y. Al Azzawi [b]

[a] Electrical & Computer Eng. Department, King Abdulaziz University, Jeddah, Saudi Arabia
[b] Electromechanical Engineering Dept., Univ. of Technology, Baghdad, Iraq

ARTICLE INFO

Article history:
Received 19 June 2011
Received in revised form 22 April 2012
Accepted 23 April 2012
Available online 17 May 2012

ABSTRACT

In this work, an average framing linear prediction coding (AFLPC) technique for text-independent speaker identification systems is presented. Conventionally, linear prediction coding (LPC) has been applied in speech recognition applications. However, in this study the combination of modified LPC with wavelet transform (WT), termed AFLPC, is proposed for speaker identification. The investigation procedure is based on feature extraction and voice classification. In the phase of feature extraction, the distinguished speaker's vocal tract characteristics were extracted using the AFLPC technique. The size of a speaker's feature vector can be optimized in term of an acceptable recognition rate by means of genetic algorithm (GA). Hence, an LPC order of 30 is found to be the best according to the system performance. In the phase of classification, probabilistic neural network (PNN) is applied because of its rapid response and ease in implementation. In the practical investigation, performances of different wavelet transforms in conjunction with AFLPC were compared with one another. In addition, the capability analysis on the proposed system was examined by comparing it with other systems proposed in literature. Consequently, the PNN classifier achieves a better recognition rate (97.36%) with the wavelet packet (WP) and AFLPC termed WPLPCF feature extraction method. It is also suggested to analyze the proposed system in additive white Gaussian noise (AWGN) and real noise environments; 58.56% for 0 dB and 70.52% for 5 dB. The recognition rates for the whole database of the Gaussian mixture model (GMM) reached the lowest value in case of small number of training samples.

1. Introduction

Automatic speech recognition (ASR) has been studied by a large number of researchers for about four decades [1]. From a commercial viewpoint, ASR is a tool with a potentially large market due to its wide range of application from the automation of operator-assisted service to speech-to-text aiding systems for hearing-impaired individuals [2].

A commonly used technique for feature extraction is based on the Karhunen–Loeve transform (KLT) [10]. These models have been applied to text-independent speaker recognition cases [3] with exceptional results. Karhunen–Loeve transform is the optimal transform according to minimum mean square error (MMSE) and maximal energy packing. Most of the suggested speaker identification systems use Mel frequency cepstral coefficient (MFCC) [5] and linear predictive cepstral

☆ Reviews processed and approved for publication by Editor-in-Chief Dr. Manu Malek.
* Corresponding author. Address: P.O. Box 80204, Jeddah 21589, Saudi Arabia. Tel.: +966 5 66 980400; fax: +966 5 695 268.
E-mail address: haleddaq@yahoo.com (K. Daqrouq).

0045-7906/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compeleceng.2012.04.014

coefficient (LPCC) [6] as features. Although MFCC and LPCC have proved to be two very good features in speech recognition, the disadvantage of the MFCC is that it uses short time Fourier transform, which has a weak time–frequency resolution and an assumption that the signal is stationary. Therefore it is relatively difficult to recognize plosive phonemes by these features. Currently, some researches [7–9] are focusing on the wavelet transform for speaker feature extraction.

Wavelet transform [4,3,11] has been extensively considered in the last two decades and has been widely utilized in various areas of science and engineering. The wavelet analysis process is implemented with dilated and translated versions of a mother wavelet. Since signals of interest can generally be expressed using wavelet decompositions, signal processing algorithms can be implemented by adjusting only the corresponding wavelet coefficients. From a mathematical point of view, the scale parameter of a wavelet can be a positive real value and the translation can be an arbitrary real number [1]. From a practical point of view, however, in order to improve computation efficiency, the values of the shift and scale parameters are often limited to some discrete lattices [12,13].

Wavelet and WP analysis have been proven as effectual signal processing techniques for a variety of digital signal processing problems. Wavelets have been used in two different methods in feature extraction plans designed for the task of speech/voice recognition. Discrete wavelet transform in place of discrete cosine transform is utilized for the feature extraction period in the first method [16]. In the second method, wavelet transform is used directly on the speech/voice signals and either wavelet coefficients containing high energy are extracted as features [8] but suffer from shift variance, or sub band energies are used instead of the Mel filter-bank sub band energies proposed in [17]. Particularly, WP bases are used in [18] as close approximations of the Mel-frequency division using Daubechies orthogonal filters. In [19], a feature extraction method based on the wavelet Eigen function was proposed. Wavelets can offer a significant computational benefit by reducing the dimensionality of the Eigen value problem. A text-independent speaker identification system based on improved wavelet transform is proposed in [9], where learning of the correlation between the wavelet transform and the expression vector is performed by kernel canonical correlation analysis.

The wavelet packets transform (WPT) performs the recursive decomposition of the speech signal obtained by the recursive binary tree. Basically, the WPT is very similar to discrete wavelet transform (DWT). However, WPT decomposes both details and approximations instead of only performing the decomposition process on approximations. WPT features have superior presentation than those of the DWT [19]. Nevertheless, as the number of wavelet packet bases grows, the time required to appropriately classify the database will become nonlinear. Consequently, dimensionality decreasing becomes a significant issue. Selecting a beneficial and relevant subset of features from a larger set is crucial to enhance the performance of speaker recognition [20,21]. A feature selection scheme is, therefore, needed to choose the most valuable information from the complete feature space to form a feature vector in a lower-dimensionality, and take away any redundant information that may have disadvantageous effects on the classification quality. To select an appropriate set of features, a criterion function can be used to provide the discriminatory power of the individual features.

The wavelet packet perceptual decomposition tree was first proposed by Sarikaya [22] and yields the wavelet packet parameters (WPP). In [24], the energy indexes of DWT or WPT were proposed for speaker identification, where WPT was superior in terms of recognition rate. Sure entropy was calculated for the waveforms at the terminal node signals obtained from DWT [25,60] for speaker identification.

Neural network applications for classification have been considered in recent years [30,15]. They are widely applied in data analysis and speaker identification. The advantage of the artificial neural network is that the transfer function between the input vectors and the target matrix (output) does not have to be predicted in advance. Artificial neural network performance depends mainly on the size and quality of training samples [28,29]. When the number of training data is small, not representative of the possible space, standard neural network results are poor. Fuzzy theory has been used successfully in many applications to reduce the dimensionality of feature vector [31]. There are many kinds of artificial neural network models, among which the back-propagation neural network (BPNN) model is the most widely used [32]. The generalized regression neural network (GRN) was introduced by [32]. Ganchev et al. [35] proposed a probabilistic neural network for speaker identification.

In fact, LPC is popular and widely used because its coefficients representing a speaker by modeling vocal tract parameters and the data size are very suitable for speaker and speech recognition. Many algorithms were developed to find a better representation of a speaker by means of a linear predictive coding technique [37,38,23]. The predictor coefficients themselves are rarely utilized as features, but they are transformed into robust and less correlated features such as linear predictive cepstral coefficients (LPCCs) [39], line spectral frequencies (LSFs) [40], and perceptual linear prediction (PLP) coefficients [41]. PLP is known as a state of the art for speech recognition task. Other, somewhat less effective features include partial correlation coefficients (PARCORs), log area ratios (LARs) and formant frequencies and bandwidths [42,56]. In the present work, the focus will be on modifying LPC coefficients and reducing the dimensionality of feature vectors.

In this research, the authors improve an effectual and a novel feature extraction method for text-independent systems, taking in consideration that the size of neural network input is a very crucial issue. This affects quality of the training set. For this reason, the presented features extraction method offers a reduction in the dimensionality of speech signals. The proposed method is based on average framing LPC in conjunction with WT upon suitable level with an appropriate wavelet function (Daubechies-type1, which is known as Haar function). For classification, PNN is proposed to accomplish online operations in a speedy manner.