Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 61 (2015) 416 – 421

Procedia Computer Science

Complex Adaptive Systems, Publication 5
Cihan H. Dagli, Editor in Chief
Conference Organized by Missouri University of Science and Technology
2015-San Jose, CA

A Wavelet Packet and Mel-Frequency Cepstral Coefficients-Based Feature Extraction Method for Speaker Identification

Claude Turner[a], Anthony Joseph[b]*

[a]*Dept. of Computer Science, Norfolk State University, 700 Park Ave, Norfolk, VA 23504*
[b]*Department of Computer Science, Pace University, 163 William St., New York, NY 10038*

Abstract

One of the most widely used approaches for feature extraction in speaker recognition is the filter bank-based Mel Frequency Cepstral Coefficients (MFCC) approach. The main goal of feature extraction in this context is to extract features from raw speech that captures the unique characteristics of a particular individual. During the feature extraction process, the discrete Fourier transform (DFT) is typically employed to compute the spectrum of the speech waveform. However, over the past few years, the discrete wavelet transform (DWT) has gained remarkable attention, and has been favored over the DFT in a wide variety of applications. The wavelet packet transform (WPT) is an extension of the DWT that adds more flexibility to the decomposition process. This work is a study of the impact on performance, with respect to accuracy and efficiency, when the WPT is used as a substitute for the DFT in the MFCC method. The novelty of our approach lies in its concentration on the wavelet and the decomposition level as the parameters influencing the performance. We compare the performance of the DFT with the WPT, as well as with our previous work using the DWT. It is shown that the WPT results in significantly lower order for the Gaussian Mixture Model (GMM) used to model speech, and marginal improvement in accuracy with respect to the DFT. WPT mirrors DWT in terms of the order of GMM and can perform as well as the DWT under certain conditions.

Peer-review under responsibility of scientific committee of Missouri University of Science and Technology

Keywords: Cepstral Coefficients, Speaker Recognition, Wavelet Packets ;

* Claude Turner. Tel.: +1-757-823-8311; fax: +1-757-823-9229.
E-mail address: cturner@nsu.edu

Peer-review under responsibility of scientific committee of Missouri University of Science and Technology
doi:10.1016/j.procs.2015.09.177

1. Introduction

Automatic speaker recognition is the identification of a person from his/her voice (Furui, 1997; Campbell,1997; Bimbot et al., 2004). A typical speaker recognition system consists of two phases: an enrollment (or training) phase, and an authentication (or testing) phase. In the enrollment phase, the user speaks an appropriate phrase into a microphone or similar device attached to the system. The system then extracts speaker-specific information from the speech signal in a process called feature extraction. These features are used to build a model for the speaker during the training process. There are many types of models that could be used in speaker recognition, including Gaussian Mixture Models (GMMs) (Reynolds, 1995), Hidden Markov Models, and vector quantization (VQ). However, GMM has been one of the most popular methods for the modeling process. The purpose of the testing phase is to determine whether the speech samples belong to one of the registered speakers. As in the training phase, speech features are extracted from the speech signal presented. The speaker is then determined by finding the speaker model which yields the maximum posterior probability for the input feature vector sequence (Reynolds, 1995).

Feature extraction is the conversion of raw speech signal to acoustic vectors that characterize speaker-specific information. Feature extraction estimates a set of features from the speech signal that represent some speaker-specific information. The speaker-specific information results from complex transformations occurring at multiple levels of the speech production process: semantic, phonologic, phonetic, and acoustic (Atal, 1976;Campbell, 1997). Despite the variation among the categories of speaker-specific information, there are only a small set of criteria that they must satisfy. These are discussed by Nolan and Wolf (Nolan, 2009;Wolf, 1972). There are a variety of filter bank-based feature extraction methods for feature extraction. However, Mel Frequency Cepstral Coefficients (MFCC) (Davis & Mermelstein, 1980) has been the most widely employed approach (Ganchev et al., 2005). In recent years, numerous variations and improvements of the original MFCC idea have been proposed (Ganchev et al., 2005; Sigurdsson et al., 2005). This is mainly attributable to researchers' efforts to exploit progress made in the area of psychoacoustics (Ganchev et al., 2005).

The testing phase of speaker recognition may be cast as a pattern recognition problem. As such, it can be partitioned into two modules (Jin, 2007): (a) a feature extraction module, and (b) a classification module. The feature extraction module is the same as in the training phase. The classification module can be further divided into two components: pattern matching and decision. The *pattern matching* component is responsible for comparing the estimated features to the speaker models. The decision component analyzes the similarity score(s), which could be either statistical or deterministic, to make a decision. The *decision* process is dependent on the system task. For the closed set identification task, the decision could be to select the identity associated with the model that is most similar to the test sample.

The wavelet packet transform (WPT) (or wavelet packet decomposition) has been employed in speaker recognition applications for over two decades with some success (Almaadeed et al., 2015) (Deshpande & Holambe, 2010) (Hsieh et al., 2003) (Sarikaya et al., (1998). Wavelet packets are an extension of the discrete wavelet transform (DWT). The discrete Fourier transform (DFT) is usually employed to compute the spectrum of the speech waveform during the MFCC feature extraction process. However, over the past few years, the discrete wavelet transform (DWT) has gained remarkable attention, and has been favored over the DFT in a wide variety of applications. The DWT enables the decomposition of a signal at multiple layers of resolution. The wavelet packet transform is an extension of the DWT that adds more flexibility to the decomposition process.

This work is a study of the impact on performance in terms of accuracy and efficiency when the WPT is used as a substitute for the DFT in the MFCC feature extraction process. The novelty of this work stems from its exploration of how the use of different wavelets and different decomposition levels in the WPT influences the performance of the speaker identification process. It is shown that the WPT results in significantly lower order for the GMM used to model speaker features and marginal improvement in accuracy. Specifically, we will compare performance in terms of accuracy and efficiency between the DFT and the WPT for Daubechies's first ten wavelets at six different decomposition levels.

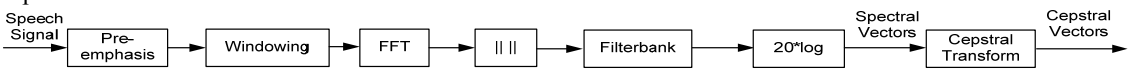

Figure 1 Modular Representation of Feature Extraction