Journal of Computer Science

Original Research Paper

Speaker Identification Using Discrete Wavelet Transform

[1]Shanthini Pandiaraj and [2]K.R. Shankar Kumar

[1]*Department of Electronics and Media Tech., Karunya University, Coimbatore, India*
[2]*Department of Electronics and Communication Engineering, Ranganathan Engineering College, Coimbatore, India*

Article history
Received: 20-06-2013
Revised: 27-04-2014
Accepted: 25-07-2014

Corresponding Author:
Shanthini Pandiaraj
Department of Electronics and Media Tech., Karunya University, Coimbatore, India
Email: pshantini@gmail.com

Abstract: This study presents an experimental evaluation of Discrete Wavelet Transforms for use in speaker identification. The features are tested using speech data provided by the CHAINS corpus. This system consists of two stages: Feature extraction stage and the identification stage. Parameters are extracted and used in a closed-set text-independent speaker identification task. In this study the signals are pre-processed and features are extracted using discrete wavelet transforms. The energy of the wavelet coefficients are used for training the Gaussian Mixture Model. Daubechies wavelets are used and the speech samples are analyzed using 8 levels of decomposition.

Keywords: Speaker Identification, Discrete Wavelets, Gaussian Mixture Model

Introduction

The goal of Automatic Speaker Recognition is to extract, characterize and recognize information about the speaker identity (Da Wu and Fu Lin, 2009). Speaker Recognition is classified as Speaker Identification and Speaker Verification. In the Speaker Identification (SI) system, an unknown speaker is compared with a database of N known speakers. SI is further classified into open-set identification and closed-set identification. The task of identifying a speaker who is assumed to be one of the N registered speakers is known as closed-set speaker identification. If the target speaker is not a member of the set of registered speakers it is known as open-set identification. SI, system can be further classified into text-dependent or text-independent task. If a known test utterance is presented to the recognizer it is a text-dependent task, otherwise it is a text-independent task.

The speaker-specific information is mainly represented by spectral features like Mel-Frequency Cepstral Coefficients (MFCC), Linear Prediction Cepstral Coefficients (LPCCs) and Short-Time Fourier Transforms (STFT). MFCCs that are calculated by taking the Discrete Cosine Transform (DCT) of mel-scaled log filter bank energies, have two drawbacks:

- Since basis vectors of the DCT cover all frequency bands, corruption of a frequency band of speech by noise affects all MFCC
- A frame of speech may contain information of two adjacent phonemes

The LPCC method is based on a linear (all-pole) model of speech production. All the above methods assume the signal to be stationary within a given time frame and lack the ability to analyze localized events correctly.

The speaker identification system is composed of two distinct phases, a training phase and a test phase. In the first phase parameters are extracted from speech.

Wavelet Based Feature Extraction

Wavelets are shifted; scaled version of original or mother wavelets (Yung Lung, 2010). Wavelets split a signal into components that are not pure sine waves. Wavelets have the ability to examine signals simultaneously in both time and frequency. Therefore, wavelet transforms are useful for analyzing noisy, transient signals.

The main advantage of wavelets (Hariharan *et al*., 2013) is that they have a varying window size, being wide for low frequency and narrow for high frequency. This is because low frequency components complete a cycle at large time interval. Therefore, slow varying components can only be identified over long time intervals but fast varying components can be identified over short time intervals. Owing to the fact that windows are adapted to the transients of each scale, wavelets lack of the requirement of a signal to

Science Publications

be stationary during the analysis interval. This leads to optimal time-frequency resolution in all frequency ranges. It can produce high frequency resolution in the low frequency part of the signal, while it has high time resolution in the high frequency part of the signal.

Discrete Wavelet Transform (DWT) decomposes (Claude *et al*., 2011) non-stationary signals at different frequency intervals with various resolutions. The signal passes through a low pass and a high pass filter. Thus, the signal is decomposed into a rough approximation and a detail component. In the orthogonal wavelet decomposition procedure, the generic step splits the approximation coefficients into two parts. After splitting, we obtain a vector of approximation coefficients and a vector of detail coefficients, both at a coarser scale. The information lost between two successive approximations is captured in the detail coefficients. The next step consists in splitting the new approximation coefficient vector; successive details are never re-analysed. Other factors influencing the selection of DWT over conventional methods are that it allows time-frequency localization. It is possible to know simultaneously the exact frequency and the exact time of occurrence of this frequency in a signal.

Wavelet spaces (Daqrouq, 2011) are a series of function spaces that are highly decorrelated from each other and are particularly suitable for the representation of signals and operators at different resolution scales that exhibit speech and speech feature behaviour.

Daubechies' Discrete Wavelet Transforms

The most common family of wavelets, Daubechies, DWT-db has its low-pass filter coefficients determined by solving the following system of Equation 1 (Da Wu and Fu Lin, 2009):

$$\sum_{k=0}^{n-1}(-1)^k h_k k^b = 0, \sum_{k=0}^{n-1} h_k = 2, \sum_{k=0}^{n-1} h_k h_{k+2l} = 2\delta_{0,l} \qquad (1)$$

Materials and Methods

Database Description

The features are extracted using speech data provided by the CHAINS Corpus as mentioned in Table 1. The corpus contains the recordings of 36 speakers obtained in two different sessions with a time separation of about two months. The first recording session provided speech in three different speaking styles (SOLO, SYNCHRONOUS and RETELL). The SOLO condition is used as training and testing material in this study. Sentences s10 to s33 are used to generate the training sets and sentences s1 to s9 are used as speech samples for testing.

Pre-Processing

The sound files contained in the corpus are.WAV files sampled at 44.1 kHz with a resolution of 16 bits. It was down sampled to 22.05 kHz.The speech samples in the frequency range 400 Hz to 8 kHz were used in this study. The speech was pre-emphasized using the factor 0.97. It is done to boost the higher frequencies.

Feature Extraction

The orignal signal is split as shown in Fig. 1. The accuracy of the reference SI system is estimated using DWT for parameterization.

The daubechies wavelets db6 to db10 are used to encode signals. 8 levels of decomposition are used for this study. The total length of the training material is approximately 50 sec per speaker; the number of speakers is 16 while utterances of 10 s are used for testing. The energy of the wavelet coefficients is taken as features for training the Gaussian Mixture Model. The training of the GMM is performed using SOLO recordings from the first recording session. The accuracy of the SI system is expressed as the mean of ten runs.

Results and Discussion

Figure 2 shows the accuracy of the GMM classifier varying the number of Gaussian components. The maximum accuracy is 83.3%. The accuracy of identification reduces as the number of Gaussian components is increased and reaches a minimum for 64 gaussian components. The speaker set consisting of 16 female speakers achieved a maximum accuracy of 74.31% when decomposed to 8 levels using db7.

Four sets of speakers each consisting of 16 speakers are employed. The first set contains eight male and 8 female speakers. The second and the third sets contain 16 male and 16 female speakers. The fourth set contains 16 female speakers speaking the same dialect (Co. Dublin-IE).

Table 1. System parameters

Parameter	Value
Sampling rate	44.1 kHz, 16 bits
Database	Chains corpus
No. of speakers	16
Filter cut-off frequencies	400 Hz and 8 kHz
Window type	Hamming
Analysis frame duration	25 ms
Analysis frame shift	15 ms

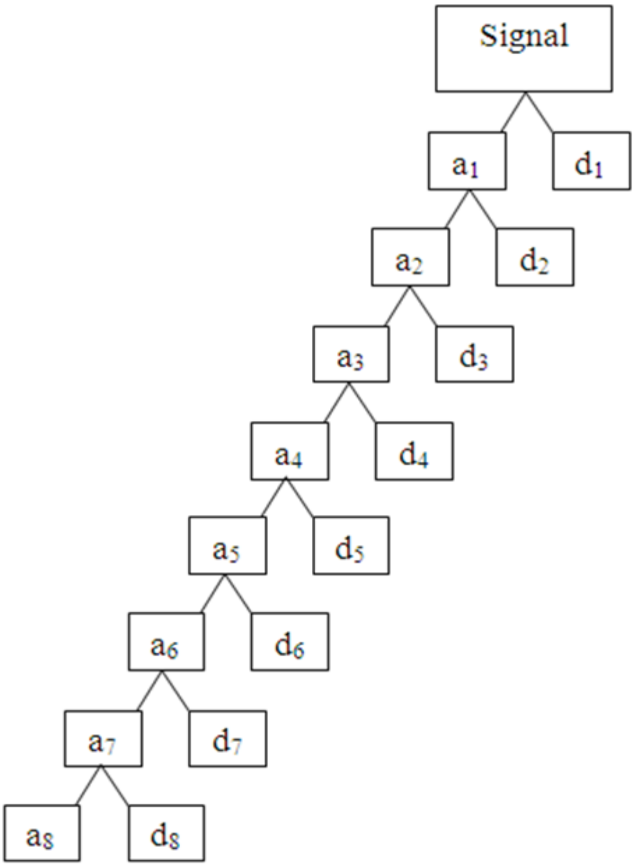

Fig. 1. Tree diagram of DWT at 8 levels

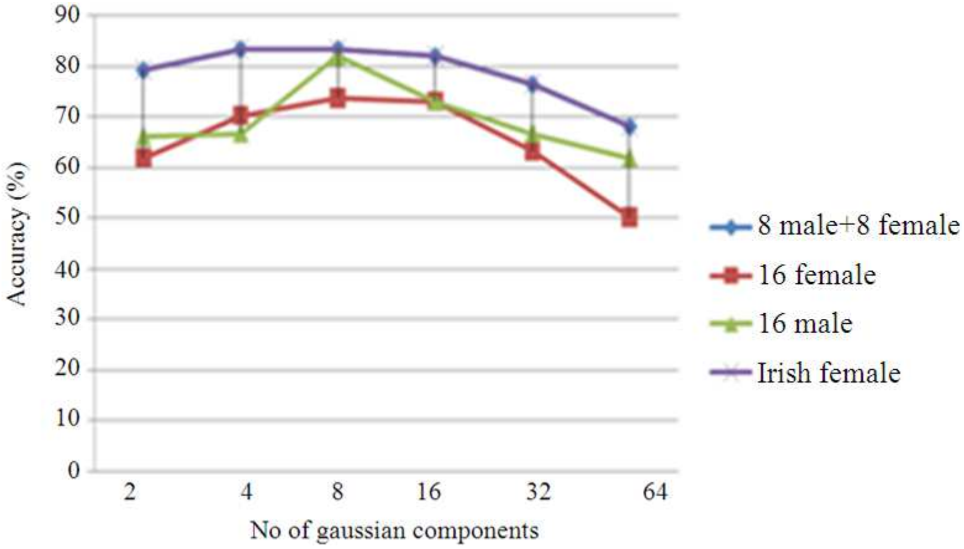

Fig. 2. Performance of GMM classifier

Conclusion

In this study, speaker identification was performed on three sets of data obtained from CHAINS Corpus. Daubechies wavelets and GMM classifier are used for feature extraction and classification, respectively. GMM classifier with eight Gaussian components yielded a maximum efficiency of 83.3% for a dataset consisting of

16 female speakers. In future, speaker identification can be performed fusing MFCC with DWT.

Author’s Contributions

All authors equally contributed in this work.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

References

Claude, T., A. Joseph, M. Aksu and H. Langdond, 2011. The wavelet and fourier transforms in feature extraction for text-dependent,filter-bank based speaker recognition. Procedia Comput. Sci., 6: 124-129. DOI: 10.1016/j.procs.2011.08.024

Da Wu, J. and B. Fu Lin, 2009. Speaker identification using discrete wavelet packet transform technique with irregular decomposition. Expert Syst. Applic., 36: 3136-3143. DOI: 10.1016/j.eswa.2008.01.038

Daqrouq, K., 2011. Wavelet entropy and neural network for text-independent speaker identification. Eng. Applic. Artificial Intelli., 24: 796-802. DOI: 10.1016/j.engappai.2011.01.001

Hariharan, M., C.Y. Fook, R. Sindhu, A.H. Adom and S. Yaacob, 2013. Objective evaluation of speech dysfluencies using wavelet packet transform with sample entropy. Dig. Sig. Process., 23: 952-959. DOI: 10.1016/j.dsp.2012.12.003

Yung Lung, S., 2010. Improved wavelet feature extraction using kernel analysis for text independent speaker recognition. Dig. Sig. Process., 20: 1400-1407. DOI: 10.1016/j.dsp.2009.12.004