Engineering Applications of Artificial Intelligence ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at SciVerse ScienceDirect

Engineering Applications of Artificial Intelligence

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

A clustering based feature selection method in spectro-temporal domain for speech recognition

Nafiseh Esfandian [a], Farbod Razzazi [a,*], Alireza Behrad [b]

[a] Department of Electrical and Computer Engineering, Islamic Azad University, Science and Research Branch, Tehran, Iran
[b] Faculty of Engineering, Shahed University, Tehran, Iran

ARTICLE INFO

Article history:
Received 26 March 2011
Received in revised form
28 January 2012
Accepted 3 April 2012

Keywords:
Speech recognition
Spectro-temporal model
Feature extraction
Clustering
Gaussian mixture models
Weighted *K*-means

ABSTRACT

Spectro-temporal representation of speech has become one of the leading signal representation approaches in speech recognition systems in recent years. This representation suffers from high dimensionality of the features space which makes this domain unsuitable for practical speech recognition systems. In this paper, a new clustering based method is proposed for secondary feature selection/extraction in the spectro-temporal domain. In the proposed representation, Gaussian mixture models (GMM) and weighted *K*-means (WKM) clustering techniques are applied to spectro-temporal domain to reduce the dimensions of the features space. The elements of centroid vectors and covariance matrices of clusters are considered as attributes of the secondary feature vector of each frame. To evaluate the efficiency of the proposed approach, the tests were conducted for new feature vectors on classification of phonemes in main categories of phonemes in TIMIT database. It was shown that by employing the proposed secondary feature vector, a significant improvement was revealed in classification rate of different sets of phonemes comparing with MFCC features. The average achieved improvements in classification rates of voiced plosives comparing to MFCC features is 5.9% using WKM clustering and 6.4% using GMM clustering. The greatest improvement is about 7.4% which is obtained by using WKM clustering in classification of front vowels comparing to MFCC features.

1. Introduction

One of the determinant issues in the performance of speech recognition systems is the process of acoustic representation of speech signals. Successful examples of audio representations are Mel scaled frequency cepstral coefficients (Davis and Mermelstein, 1980) and spectro-temporal features (Chi et al., 2005; Mesgarani et al., 2006; Mesgarani et al., 2008) that are both inspired from human hearing models. In particular, spectro-temporal features use a simplified model of human brain cortical stage after successful modeling of internal ear functionalities. Although there had been some modeling investigations on internal ear (Yang et al., 1992) and auditory cortical system (Wang and Shamma, 1995) for engineering applications; they had not been employed in engineering applications for ten years. Recently, a computational auditory model has been obtained according to neurology, biology and investigations at various stages of the auditory system of brain (Chi et al., 2005) and has been developed in various applications such as phoneme classification (Mesgarani et al., 2008), voice activity detection (Mesgarani et al., 2006; Valipour et al., 2010), speaker separation (Elhilali and Shamma, 2004; Rigaud et al., 2011), auditory attention (Shamma et al., 2011) and speech enhancement systems (Mesgarani and Shamma, 2005) in recent years. This model has two main stages. In the stage of auditory modeling, an auditory spectrogram is extracted for the input acoustic signal. In the next stage, the spectro-temporal features of speech are extracted by applying a set of two dimensional spectro-temporal receptive field (STRF) filters on the spectrogram. STRF filters are scaled versions of a two dimensional impulse response (Chi et al., 2005). It is observed that modified versions of these features are more robust in noisy environments in comparison to cepstral coefficients (Bouvrie et al., 2008). The main drawback of spectro-temporal analysis is the large number of extracted features which may affect the parameter estimation accuracy in the training phase of a speech classifier. Some methods such as PCA, LDA and neural networks are used to reduce the number of features in spectro-temporal domain (Mesgarani et al., 2006; Meyer and Kollmeier, 2011). These methods are general feature selection methods. Therefore, these methods are not exactly compatible with the speech classification problems. In addition, there are some approaches which try to find out the best 2D impulse response (best scale-best rate) to extract the appropriate features (Mesgarani et al., 2008).

This study is motivated by the clustered behavior of information in the spectro-temporal domain. In fact, the phonemes'

* Corresponding author.
E-mail addresses: N.Esfandian@qaemshahriau.ac.ir (N. Esfandian), razzazi@srbiau.ac.ir (F. Razzazi), behrad@shahed.ac.ir (A. Behrad).

0952-1976/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.engappai.2012.04.004