Engineering Applications of Artificial Intelligence 25 (2012) 1194–1202

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
Available online 26 May 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

information is concentrated in the specific parts of the spectro-temporal features space. In other words, it is desirable to represent the phoneme as the parameters of a number of clusters in the spectro-temporal domain. Some studies have shown that the space of MFCC features is not properly clustered (Kinnunen et al., 2001). It means that this space does not have distinct clusters to represent the short time properties of an uttered speech. Therefore, applying clustering methods to MFCC may be just considered as a kind of space coverage. In contrast, there are some domains that clustering results in better secondary features for signal representation and classification (Yu and Kamarthi, 2010; Ahmed and Mohamad, 2008; Yu et al., 2007). In this study, it is tried to study the effect of clustering to represent the speech in spectro-temporal domain. It will be shown that in the new clustered features space, phonemes are more separable.

There are many clustering methods including Gaussian mixture model (GMM) (Duda et al. 2001), *K*-means and weighted *K*-means (WKM) clustering (Kerdprasop et al., 2005), support vector clustering (Ping et al., 2010) and scale space statistics clustering (Sakai and Imiya, 2009). Clustering may be used either as a classification tool for audio and speech signals (Dhanalakshmi et al., 2011) or as a tool to extract and select a set of secondary acoustic features. This study is focused on the second approach. Two clustering methods are studied to reduce the spectro-temporal features into a few effective secondary features for each frame. GMM and WKM clustering algorithms are shown to be useful in many practical image segmentation applications (Blekas et al., 2005; Abras and Ballarin, 2005). Specifically, GMM is a good choice to model irregular data well. Therefore, in this paper, spatial GMM is employed to cluster the feature space as a feature reduction approach. Spatial GMM input vectors include the position attributes in addition to the representation attributes at that point. This makes the vector large and may lead the system to have inaccuracy problems in parameter estimation. To reduce the size of the vectors in the clustering procedure, the vectors should be weighted due to their importance in the representation of the corresponding frame. Therefore, WKM clustering algorithm is investigated as another clustering method which may be useful to cluster the spectro-temporal space.

The organization of the paper is as follows: The spectro-temporal representation is briefly discussed in Section 2. The proposed secondary feature extraction algorithm for phonemes using the behavior of GMM and WKM clusters in spectro-temporal domain is presented in Section 3. The proposed features are experimentally evaluated in the features space and tested on a phoneme classification task in Section 4. The paper is concluded in Section 5.

2. Spectro-temporal feature representation

The auditory model that is described in this section, is a mathematical model of internal ear and the first layer of auditory brain section that are used for speech processing applications in recent years. The block diagram of the auditory model is shown in Fig. 1.

2.1. The primary stage of the auditory model

When an audio signal enters the ear, the neural sensors of the basilar membrane of the cochlea convert one dimensional audio signal into a two-dimensional auditory spectrogram image which the frequency axis of this 2D image is a tonotopic (nearly logarithmic) axis. Basilar membrane can be considered as a band-pass filter bank. This filter bank includes 128 asymmetric band-pass filters with the impulse response $h_{\mathrm{cochlea}}(t;f)$ which are uniformly distributed along the tonotopic axis. The cochlear filter outputs $y_{\mathrm{cochlea}}(t,f)$ are converted into auditory nerve patterns $y_{\mathrm{an}}(t,f)$ by an inner hair cell stage (IHC). IHC stage consists of a high-pass filter in time domain, an instantaneous nonlinear compression $g_{\mathrm{hc}}(.)$ and a time domain low-pass filter $\mu_{\mathrm{hc}}(t)$. The last part of this stage is a model of lateral inhibitory network (LIN) activity, which increases the frequency selectivity of the cochlear filters. LIN is approximated by a first order derivative along the tonotopic axis. The output of this stage $y_{\mathrm{LIN}}(t,f)$ is obtained by using a half wave rectifier to remove the negative outputs and the final output is approximated by integrating $y_{\mathrm{LIN}}(t,f)$ during a short time window with an impulse response $\mu_{\mathrm{midbrain}}(t,\tau)=e^{-t/\tau}u(t)$ with the time constant $\tau\cong 8$ ms. The mathematical formulation of this stage is formulated as below.

$$y_{\mathrm{cochlea}}(t,f)=s(t)^{*}h_{\mathrm{cochlea}}(t;f) \tag{1}$$

$$y_{\mathrm{an}}(t,f)=g_{\mathrm{hc}}\left(\frac{\mathrm{d}}{\mathrm{d}t}y_{\mathrm{cochlea}}(t,f)\right)^{*}\mu_{\mathrm{hc}}(t) \tag{2}$$

$$y_{\mathrm{LIN}}(t,f)=\max\left(\frac{\mathrm{d}}{\mathrm{d}f}y_{\mathrm{an}}(t,f),0\right) \tag{3}$$

$$y(t,f)=y_{\mathrm{LIN}}(t,f)^{*}\mu_{\mathrm{midbrain}}(t;\tau) \tag{4}$$

2.2. The cortical stage of auditory model

The primary auditory stage of the brain analyzes the auditory spectrogram as an image. At this stage, a two-dimensional wavelet transform of auditory spectrogram is calculated. This transform is performed using a spectro-temporal mother wavelet, similar to a two-dimensional Gabor function. In other words, the spectral and temporal modulation contents of the auditory spectrogram are estimated via a bank of modulation-selective 2-D filters. Each filter is tuned to a range of spectral–temporal modulations. Spectro-temporal impulse responses of these filters are called spectro-temporal response fields (STRFs). There are two primitive 2-D STRF types which are named upward (+) and downward (−), respectively which are demonstrated as positive and negative rates, respectively. Therefore, the cortical representation of speech has

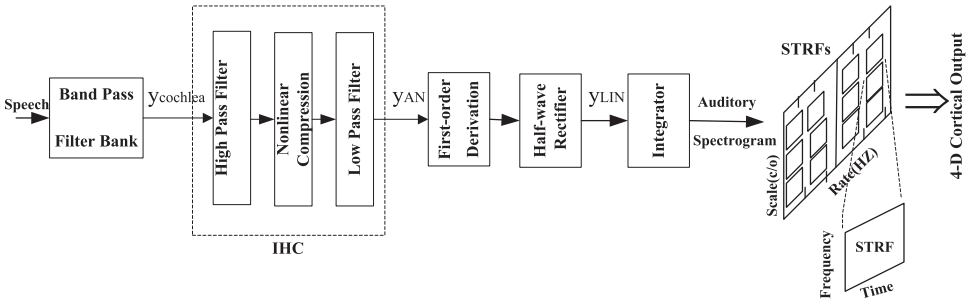

Fig. 1. Block diagram of the auditory model.