Pattern Recognition Letters 33 (2012) 1703–1709

Contents lists available at SciVerse ScienceDirect

Pattern Recognition Letters

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

Discriminative feature extraction for speech recognition using continuous output codes

Omid Dehzangi [a,*], Bin Ma [b], Eng Siong Chng [a], Haizhou Li [a,b,c]

[a] School of Computer Engineering, Nanyang Technological University, Singapore 639798, Singapore
[b] Institute for Infocomm Research, Singapore 138632, Singapore
[c] University of Eastern Finland, FI-80101 Joensuu, Finland

ARTICLE INFO

Article history:
Received 17 December 2010
Available online 29 May 2012

Communicated by S. Sarkar

Keywords:
Speech recognition
Feature transformation
Generalized discriminant analysis
Output coding

ABSTRACT

Feature transformation techniques have been widely investigated to reduce feature redundancy and to introduce additional discriminative information with the aim to improve the performance of automatic speech recognition (ASR). In this paper, we propose a novel method to obtain discriminative feature transformation based on output coding technique for speech recognition. The output coding transformation projects the speech features from their original space to a new one where each dimension of the features captures information to distinguish different phones. Using polynomial expansion, the short-time spectral features are first expanded to a high-dimensional space where the generalized linear discriminant sequence kernel is applied on the sequences of input feature vectors. Then, the output coding transformation formulated via a set of linear SVMs projects the sequences of high dimensional vectors into a tractable low-dimensional feature space where the resultant features are well-separated continuous output codes for the subsequent multi-class classification problem. Our experimental results on the TIMIT corpus show that the proposed features achieve 10.5% ASR error rate reduction over the conventional spectral features.

1. Introduction

In most practical ASR systems, there are two modules involved: the front-end feature extraction and the back-end classification. The feature extraction uses signal processing and analysis methods to reduce the data rate of the signal, and to transform the incoming speech into a form which can represent the speech signal. In practice, achieving accurate recognition requires a careful design of the feature extraction module because this module extracts the discriminative information utilized by the classification module to perform speech recognition.

Acoustic feature vectors conventionally used for ASR systems are not designed by a discriminative measure (Davis and Mermelstein, 1980; Hermansky, 1990). In recent years, there has been much research interest to improve the discriminative capabilities of acoustic features via discriminative transformation. Inspired by discriminative training of acoustic models, previous research as in minimum classification error (MCE) method (Biem and Katagiri, 1993, 1997), feature-space minimum phone error (fMPE) training (Povey et al., 2005), and maximum mutual information (MMI)-splice (Droppo and Acero, 2005) showed that using discriminative training to optimize the feature projection function, as well as employing them to adjust the parameters of the feature extractor, is effective to improve the performance of speech recognition. In general, discriminative techniques require labeled data in their learning mechanism, and employ them in different ways. As such, discriminative feature transformation systems also take various approaches to address the problem of learning the transform function given the annotated data. For instance, Linear Discriminant Analysis (LDA) (Haeb-Umbach and Ney, 1992) as a linear transformation technique finds the optimal solution with the assumption that the distribution of data within each class is a multivariate normal distribution with a single shared covariance matrix for all classes. Such assumption does not always hold when there are rather large number of classes such as in speech recognition. LDA also suffers from insufficient training samples when dealing with high-dimensional data, when the within class scatter matrix is almost singular. Heteroscedastic linear discriminant analysis (HLDA) (Schukat-Talamazzini et al., 1995; Kumar and Andreou, 1998) is an extended version of LDA, in which the limitation implying that all the class covariance matrices are alike is eliminated. In another study (Gopinath, 1998), a maximum likelihood linear transform (MLLT) was introduced which is shown to be very effective. It is also shown that applying MLLT on top of LDA and HLDA leads to improve the classification accuracy as successfully investigated in (Saon et al., 2000). However, all the above-mentioned methods still makes a strong

* Corresponding author. Tel.: +65 97756620; fax: +65 67926559.
E-mail addresses: dehzangi@pmail.ntu.edu.sg (O. Dehzangi), mabin@i2r.a-star.edu.sg (B. Ma), aseschng@ntu.edu.sg (E.S. Chng), hli@i2r.a-star.edu.sg (H. Li).

0167-8655/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2012.05.012

assumption about the distribution of each class of acoustic vectors to be normally distributed.

On the other hand, in the category of nonlinear feature transformation systems, the hybrid connectionist-HMM system (Bourlard and Morgan, 1994) uses discriminatively-trained neural networks to estimate the probability distribution among sub-word units given the acoustic observations. More recently, TANDEM connectionist feature extraction (Hermansky et al., 2000) combines neural-net discriminative feature processing with Gaussian-mixture distribution modeling. These techniques have much less restrictions than the generative transformation models, however, as the transform function becomes more complex, the learning mechanism faces higher computational expense.

Our contribution in this paper is the application of the generalized linear discriminant sequence (GLDS) kernel to measure the similarity between the sequences of feature vectors, and then to formulate the linear transform based on output coding technique using a set of SVMs to derive new features. Differently from LDA and HLDA, the features are extracted via SVM as a distribution free method which is suitable in the high dimensional space. Compared to TANDEM which is a nonlinear transformation using neural networks, we adopt linear SVMs to transform the high dimensional feature vectors for better computational efficiency. The proposed linear feature transformation mechanism encompasses two major modules: (1) application of SVMs with GLDS kernel, (2) output coding technique. Motivated by the prior work in the domain of speaker and language recognition (Campbell et al., 2006), frame-based cepstral features are first expanded using polynomial expansion to a high dimensional space where sequences of context frames are then averaged to form the GLDS kernel. Benefiting from the distribution free properties of SVMs in the high dimensional space, the high-dimension feature vectors are then passed to a set of linear SVMs that project the sequences of expanded vectors into a tractable low dimensional space with the objective to introduce additional discriminative capabilities into the new feature space using output coding technique (Dietterich and Bakiri, 1995). Output coding decompose a multi-class classification problem into multiple binary classification problems. A set of binary classifiers, each trained to distinguish between two disjoint subsets of the labeled data, is constructed to create an output representation for a test instance, referred to as output codes. Each class is assigned a unique binary vector, which represents output codes for the instances in the class, and this representation is referred to as the codeword. For a test instance, classification can be achieved by finding the nearest codeword in Hamming distance. It was shown that the output code representation improved the generalization capability of the learning machines on a wide range of multi-class learning tasks (Masulli and Valentini, 2004). An improved output coding method with continuous relaxation on the output scores was also proposed in (Crammer and Singer, 2000) which is adopted in this paper. We employ linear SVMs as the binary classifiers and investigate the use of output coding technique for feature space transformation in speech recognition task. We also present a scheme to select an effective set of discriminative output codes.

The rest of this paper is organized as follows. In Section 2, the proposed feature transformation as a combination of two techniques namely SVM with GLDS kernels and continuous output coding is presented. In Section 3, the experimental setup and results are presented. Section 4 concludes the paper.

2. The proposed feature transformation system

Fig. 1 shows the architecture of the proposed feature transformation system which is composed of two main stages. In the first stage, polynomial expansion and averaging in the high dimensional space form the GLDS kernel. The second stage is the output coding design to encode discriminative information of the high dimensional space via a set of linear SVM classifiers to form the proposed COC features for speech recognition.

First, MFCC features are expanded into a high dimensional space using polynomial expansion of order d. In the high dimensional space, a sequence of μ successive frames centered around each high dimensional frame vector is formed and then is averaged as the input to the SVM classifiers. Eventually, continuous output codes are designed via a set of C linear SVMs, where the C SVMs have to be optimized on the training data to encode effective phone discriminative information through the generated continuous output codes. Each SVM corresponds in order to a subset of phones and is trained as a projection function in $\{f_1(\cdot), f_2(\cdot), \ldots, f_C(\cdot)\}$. The resulting feature space provides the discriminative information for separating different phones. The COC features are then used as the new input features to the recognizer. In the next sections, the proposed COC feature transformation system is described in detail.

2.1. Support vector machine

The SVM approach offers an effective classification strategy to separate input vectors in 2-class problems and is investigated in many different applications (Vapnik, 1995). SVM projects the input vector $\mathbf{x}$ into a scalar value $f(\mathbf{x})$ as the output score,

$$f(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b \tag{1}$$

where, the vectors $\{\mathbf{x}_i | i = 1, \ldots, N\}$ are the support vectors, N is the number of support vectors, $\alpha_i > 0$ are adjustable weights, $y_i = \{-1, +1\}$, b is the bias term, and the function $K(\mathbf{x}_i, \mathbf{x})$ is the kernel function. For the 2-class classification, the class decision is made based on the sign of $f(\mathbf{x})$. As it can be seen, the classifier is constructed from sums of the kernel function expressed as,

$$K(\mathbf{x}_i, \mathbf{x}) = \phi(\mathbf{x}_i)^t \phi(\mathbf{x}) \tag{2}$$

where $\phi(\mathbf{x})$ is a mapping from the input space to a possibly infinite dimensional space. However, the expansion $\phi(\cdot)$ is implicit and only the dot product of the expansion is known. If the explicit form of $\phi(\cdot)$ is available, then $f(\mathbf{x})$ can be written as,

$$f(\mathbf{x}) = \left[\sum_{i=1}^{N} \alpha_i y_i \phi(\mathbf{x}_i) + \mathbf{b} \right]^t \phi(\mathbf{x}) = \mathbf{w}_{svm}^t \phi(\mathbf{x}) \tag{3}$$

where, $\mathbf{b} = [b\ 0\ 0\ \cdots\ 0]^t$, and $\mathbf{w}_{svm}$ is a weight vector in the high dimensional space. The advantage of this form over the classic one is that it involves only storage and computation of the vector $\mathbf{w}_{svm}$ rather than the support vectors. If there are a significant number of support vectors, using Eq. (3) can significantly reduce the storage space and computation.

2.2. Generalized linear discriminant sequence (GLDS) kernel

The short-time spectral features are not designed using a discriminative measure and involve high overlapping regions. Therefore, directly applying SVMs to the input feature space results in too many support vectors which is difficult to manage during the recognition phase. This problem can be eased by using an explicit kernel for scoring. In (Campbell et al., 2006), the GLDS kernel is proposed for speaker and language recognition in which $\phi_d(\cdot)$ is an explicit polynomial expansion. The polynomial expansion $\phi_d(\mathbf{x})$ is the vector of all monomials of degree d for the original m-dimensional input feature vector $\mathbf{x}$ (e.g., cepstral coefficients). This leads to a vector of $\binom{m+d}{d}$ dimension made up of all possible