Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language 47 (2018) 333–350

www.elsevier.com/locate/csl

Sparse coding based features for speech units classification☆

Pulkit Sharma*, Vinayak Abrol, A.D. Dileep, Anil Kumar Sao

School of Computing and Electrical Engineering, Indian Institute of Technology, Mandi, India

Received 26 May 2016; received in revised form 15 June 2017; accepted 17 August 2017
Available online 1 September 2017

Abstract

In this work, we propose sparse representation based features for speech units classification tasks. In order to effectively capture the variations in a speech unit, the proposed method employs multiple class specific dictionaries. Here, the training data belonging to each class is clustered into multiple clusters, and a principal component analysis (PCA) based dictionary is learnt for each cluster. It has been observed that coefficients corresponding to middle principal components can effectively discriminate among different speech units. Exploiting this observation, we propose to use a transformation function known as weighted decomposition (WD) of principal components, which is used to emphasize the discriminative information present in the PCA-based dictionary. In this paper, both raw speech samples and mel frequency cepstral coefficients (MFCC) are used as an initial representation for feature extraction. For comparison, various popular dictionary learning techniques such as K-singular value decomposition (KSVD), simultaneous codeword optimization (SimCO) and greedy adaptive dictionary (GAD) are also employed in the proposed framework. The effectiveness of the proposed features is demonstrated using continuous density hidden Markov model (CDHMM) based classifiers for (i) classification of isolated utterances of E-set of English alphabet, (ii) classification of consonant-vowel (CV) segments in Hindi language and (iii) classification of phoneme from TIMIT phonetic corpus.

Keywords: Sparse representation; Dictionary learning; Speech recognition

1. Introduction

Feature extraction is an important step for speech recognition, as it involves conversion of the speech signal into a sequence of acoustic features (frame by frame basis) (Rabiner and Schafer, 2010). Here, the extracted features should capture characteristics contributing to the phonetic differences among different speech units, so as to enable discrimination among them. Available features for the tasks in speech recognition are either motivated by speech production or by speech perception mechanisms. Features belonging to the former category are also known as articulatory based features (Saenko et al., 2009; Mitra et al., 2011), and have been demonstrated to give good speech recognition performance, but their estimation from speech signal is difficult (Mitra et al., 2011). Hence, features based on speech perception, such as mel frequency cepstral coefficients (MFCC) (Davis and Mermelstein, 1980),

☆ This paper has been recommended for acceptance by L. ten Bosch.

* Corresponding author.
E-mail address: pulkit_s@students.iitmandi.ac.in (P. Sharma), vinayak_abrol@students.iitmandi.ac.in (V. Abrol), addileep@iitmandi.ac.in (A.D. Dileep), anil@iitmandi.ac.in (A.K. Sao).

http://dx.doi.org/10.1016/j.csl.2017.08.004

perceptual linear predictive (PLP) coefficients (Hermansky, 1990) etc. are very popular for the tasks in speech recognition. Further, few approaches have performed transformations, e.g., neural networks based transformations (Hermansky et al., 2000) and feature-space minimum phone error (fMPE) (Povey et al., 2005), on the standard perception based features to improve the performance of speech recognition systems. Mitra et al. (2011) have demonstrated that MFCC in conjunction with articulatory representations provide better results as compared to individual features, at the expense of increased computational complexity.

Features for speech recognition should highlight the discriminative information among the speech units. Although the speech signal corresponds to a high dimensional data captured using sensors i.e., microphone (Tosic and Frossard, 2011), the number of generating causes is very small as compared to recorded observations. Thus, the information relevant to the underlying process of generating signal (speech in our case) is generally of reduced dimensionality as compared to the recorded observations (Tosic and Frossard, 2011), and can be exploited for estimating efficient representations of the speech signal. Such representations can be achieved either using a compact code or sparse code (Shashanka et al., 2007). In general, compact code ($\mathbf{x}_c \in \mathbb{R}^{N_1}$) for any signal $\mathbf{x} \in \mathbb{R}^N$, is of fewer dimension i.e., $N_1 < N$. On the contrary, the number of elements in sparse distributed code ($\boldsymbol{\alpha} \in \mathbb{R}^N$), are equal to the number of elements in the input, but most of those elements are zero, i.e., only K ($K \ll N$) elements are needed to represent the given input faithfully. However, in sparse code, location of K significant coefficients representing the generating causes may vary for different speech units (Shashanka et al., 2007). This helps in capturing the distinct causes responsible for different speech units. This work is focused on demonstrating the use of sparse code in capturing discriminative information for speech units classification.

In recent years sparse coding based signal processing has been applied to various speech processing applications such as audio classification (Zubair et al., 2013), speaker verification (Haris and Sinha, 2012), speech enhancement (Abrol et al., 2013; Low et al., 2013), speech recognition (Sivaram et al., 2010), speech separation (Xu et al., 2013), speaker tracking (Barnard et al., 2014) and speech coding (Giacobello et al., 2010). In sparse representations (SR) of speech signal, a speech frame is written as a linear combination of atoms of a resource, known as dictionary. The sparse vector obtained for each speech frame, given a dictionary, is used as a feature. The behavior of sparse vector is very much influenced by the choice of dictionary, which could be either analytical or learnt. Analytical dictionaries are easy to implement and have fast transform properties. The learnt dictionaries are derived from the data itself and thus adapt to the variations in data effectively (Tosic and Frossard, 2011).

In this work, novel SR of speech signal, computed using principal component analysis (PCA) based dictionary, is proposed for tasks in speech recognition. The approach presented in this work is similar to the one proposed in Dong et al. (2011), in the context of images. In order to capture the variations present in the speech signal, we have shown that it is preferable to use multiple dictionaries. This is done by first clustering speech frames corresponding to a speech unit into Q different clusters. Eigenvalues and corresponding eigenvectors are computed for each cluster. All the eigenvectors (in decreasing order of eigenvalues) are arranged column-wise to construct dictionary (specifically sub-dictionary) for a cluster. It is studied in the literature of image processing, that the eigenvectors corresponding to the largest eigenvalues give information common to all the training samples, and the least significant information is present in the eigenvectors corresponding to small eigenvalues (Shejin and Sao, 2012; O'Toole et al., 1994, 1997). The eigenvectors corresponding to intermediate principal directions include the discriminative information among the training examples and are demonstrated in the context of face recognition. We have proposed to use a transformation function known as weighted decomposition (WD) (Shejin and Sao, 2012) to suppress the most and least significant components in the proposed PCA-based dictionary.

In the proposed approach, dictionaries belonging to all the clusters along with their centroids are stored for each speech unit (class). For any speech frame, a minimum distance criteria is used to select a suitable dictionary and then the corresponding sparse vector is used as a feature. The extracted features are employed in a continuous density hidden Markov model (CDHMM) based classifier to demonstrate the usefulness of these features for speech units classification tasks. As a comparison, we have also explored K-singular value decomposition (KSVD) dictionary (Aharon et al., 2006), simultaneous codeword optimization (SimCO) dictionary (Dai et al., 2012) and greedy adaptive dictionary (GAD) (Jafari and Plumbley, 2011) to obtain SR for a speech unit in the proposed approach.

This paper is an extension of our existing work published in Sharma et al. (2015), and the contributions of this work are: (i) PCA-based multiple dictionaries to extract SR based feature from speech signals, (ii) emphasizing discriminative information using a dictionary based on weighted decomposition of principal components (PCs)