Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 76 (2016) 28–41

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Phone classification via manifold learning based dimensionality reduction algorithms

Heyun Huang, Louis ten Bosch *, Bert Cranen, Lou Boves

CLST/CLS, Radboud University, Nijmegen, Netherlands

Received 8 May 2014; received in revised form 30 September 2015; accepted 29 October 2015
Available online 7 November 2015

Abstract

Mechanical limitations imposed on the articulators during speech production lead to a limitation of the intrinsic dimensionality of speech signals. This limitation leads to a specific neighborhood structure of speech sounds when they are represented in a high-dimensional feature space. We investigate whether phone classification can be improved by exploiting this neighborhood structure, by means of extended variants of the conventional Linear Discriminant Analysis (LDA) based on manifold learning.

In this extended LDA approach, the within-class and between-class scatter matrices are defined in terms of adjacency graphs. We compare extensions of LDA that use either a full adjacency graph or an adjacency graph defined in the neighborhood of the training observations. In addition, we apply different kernels for weighing the distances in the graphs via different kernels, of which the Adaptive Kernel is proposed in this paper.

Experiments with TIMIT show that while LDA algorithms that use the full adjacency graph do not outperform traditional LDA, the algorithms that exploit only local information provide significantly better results than traditional LDA. These improvements are not uniform across different broad phonetic classes, which suggests that the added value of the neighborhood structure is phone class dependent. The structure is represented by locally different densities in the neighborhood of feature vectors that are representative of a specific phone in a specific context.

Keywords: Phone classification; TIMIT; Manifold learning; Graph embedding framework; LDA-based dimensionality reduction

1. Introduction

The movements of articulators in the human speech production system are subject to mechanical and ballistic constraints. Due to these constraints the effective 'intrinsic' dimensionality of the set of acoustic features of speech signals is limited, even when these signals are represented in a high dimensional space. During the last decade several different attempts have been made to develop acoustic representations of speech signals that benefit from the low intrinsic dimensionality, based on the insight that the local structure is dependent on the speech sound and its acoustic context, as determined by the temporal and spatial limitations imposed by the articulatory system. A number of approaches aimed at reestimating the movements of the vocal tract from the speech signals in the form of articulatory features (Frankel et al., 2007). Another research direction uses explicit parametric trajectories to capture the articulatory dynamics (Gish and Ng, 1996; Gong, 1997; Illina and Gong, 1997; Han et al., 2007; Zhao and Schultz, 2002), especially for vowels. The authors in Kim and Un (1997), Paliwal (1993), Wellekens (1987), Pinto et al. (2008), Russell (1993), Ostendorf et al. (1995), and

* Corresponding author.
E-mail addresses: heyunhuang@qq.com (H. Huang), l.tenbosch@let.ru.nl (L. ten Bosch), b.cranen@let.ru.nl (B. Cranen), l.boves@let.ru.nl (L. Boves).

http://dx.doi.org/10.1016/j.specom.2015.10.005
0167-6393/

Yun and Oh (2002) attempted to model the temporal dynamics by using conditional probability distributions. All approaches mentioned above try to express the information about articulatory continuity explicitly. And most of these approaches, if not all, mainly or exclusively aim at improving the performance of some Automatic Speech Recognition (ASR) system.

Other research directions aim at using machine learning approaches to benefit from the fact that the intrinsic dimensionality of speech signals is limited, instead of directly attempting to obtain explicit parametric representations of the articulatory dynamics. These approaches take a (very) high-dimensional representation as a starting point, due to the fact that they capture temporal dynamics by stacking a number of 10 ms frames of spectral features (MFCCs, PLPs, Mel energy spectra, etc.) (e.g., De Wachter et al., 2007; Gemmeke et al., 2011; Tahir et al., 2011). In order to appropriately represent articulatory dynamics at the level of a syllable, feature representations must span at least 250 ms, i.e. 25 frames with a rate of 100 frames per second (Hermansky, 2010). Using 13-dimensional MFCCs, this yields a feature space of dimension $25 \times 13 = 325$. To exploit the fact that the intrinsic dimensionality of the speech signals is much lower than 325, and to avoid the 'curse of dimensionality' (Beyer et al., 1999), some form of dimensionality reduction is required. For example, in conventional ASR, Linear Discriminant Analysis (LDA) (Fisher, 1936) (also known as Fisher Discriminant Analysis, FDA) has often been used to map high-dimensional stacks of MFCC features to lower-dimensional feature vectors, while maximizing the information that discriminates between phone models (e.g., Haeb-Umbach and Ney, 1992; Erdogan, 2005; Pylkkönen, 2006). However, while most of the previous research into exploiting the effects of the low dimensionality of the articulatory system was aimed at improving ASR, recently an interest has emerged in harnessing the results of machine learning approaches to establish links with the large store of phonetic and phonological knowledge (Jansen and Niyogi, 2013). Interestingly, the authors of Jansen and Niyogi (2013) point out that the machine learning community has developed multiple algorithms that aim to discover the underlying low-dimensional structure in data, but that with the exception of ISOMAP (Tenenbaum et al., 2000; ten Bosch et al., 2011) none of these algorithms has been tested on a realistic speech task. While the authors in Jansen and Niyogi (2013) focus attention on the class of machine learning approaches based on the Graph Laplacian and the Laplace–Beltrami operator (see e.g. Singer, 2006), we here focus on extensions of LDA that allow for manifold learning in relation to the use of adjacency graphs. As in Jansen and Niyogi (2013), the goal of our research is to advance knowledge about the underlying structure in speech signals; a corollary goal is to understand the degree to which LDA algorithms that preserve the local neighborhood relations in the speech data can uncover and exploit structure. In other words, the main objective of our study is to investigate to what extent knowledge about the distributions of the acoustic representation of phones – expressed in the form of neighborhood structure or manifolds – might be exploited for speech signal processing. Our goal is to gain understanding, rather than developing a particular step in an ASR processing cascade with minimization of error rates as single aim. For that reason we focus on a task that is closely related to general classification problems, namely phone classification. We use the TIMIT corpus as the test platform (Garofolo, 1988). It is because of these goals that we decided not to pursue the extremely fruitful research line of using Deep Neural Networks (DNN), e.g. (Seide et al., 2011; Hinton et al., 2012). Although (Huang et al., 2014) showed that the relative phone error rate decreased by phone-dependent proportions between 15.6% and 39.8% when they replaced a GMM-based posterior probability estimator by a DNN-based system, the results do not provide insight in the phonetic structure. In this paper, our aim is to better understand the phonetic structure by investigating the local structure in the adjacency graph representation of extensions of LDA, which is difficult to achieve by using DNNs.

Classical LDA assumes that all classes that must be distinguished obey a single and homoscedastic normal distribution. In the phone classification task this assumption is highly unlikely to be true: the high degree of variation in the speech production process, in combination with the coarticulation with surrounding phones, will make the distributions within the phone classes much more complex (Jurafsky et al., 2001). Therefore, it appears useful to extend the traditional LDA by taking into account the resulting substructure in the acoustic space. Because a substantial part of the variation is systematic, rather than random, the acoustic space occupied by the speech signal is likely to be structured along (possibly several) lower-dimensional manifolds. This manifold structure in the acoustic space (the space defined by the feature representation) is likely to result from the locally different densities in the neighborhood of feature vectors that are representative of a specific phone in a specific context.

In Yan et al. (2007) it was demonstrated that the neighborhood structure can be expressed in terms of adjacency graphs, and that different extensions of classical LDA can be unified in a general graph embedding framework. In this paper we investigate whether and to what extent the LDA algorithms subsumed by the framework of adjacency graphs can harness the neighborhood structure to the benefit of the TIMIT phone classification task. In addition, we will propose a novel adaptive kernel (based on older, well-known kernels, see e.g. Abramson, 1982; Kim and Scott, 1992) to extend one of the most promising LDA algorithms, i.e. heteroscedastic linear discriminant analysis (HLDA) (Kumar and Andreou, 1998; Burget, 2004; Sakai et al., 2009). Feature frames are represented by a single high-dimensional vector created by stacking 23 consecutive 13-dimensional MFCC vectors,