Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 38 (2016) 28–45

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

# Locally learning heterogeneous manifolds for phonetic classification☆

Heyun Huang [a], Yang Liu [b], Louis ten Bosch [a,*], Bert Cranen [a], Lou Boves [a]

[a] *Centre for Language and Speech Technology, Radboud University Nijmegen, The Netherlands*
[b] *Department of Statistics, Yale University, 24 Hillhouse Avenue, New Haven, CT 06511, USA*

Received 23 August 2013; received in revised form 21 November 2015; accepted 23 December 2015
Available online 31 December 2015

**Abstract**

Most state-of-the-art phone classifiers use the same features and decision criteria for all phones, despite the fact that different broad classes are characterized by different manners and place of articulation that result in different acoustic features. This paper uses manifold learning to address structure in the acoustic space. Previous approaches to dimensionality reduction based on manifold learning assumed that the acoustic space can be characterized by a uniform manifold structure. In this paper we relax this assumption by learning different manifold structures for broad phonetic classes. Because all known classifiers make confusions between broad classes, we designed a two-level classifier in which the top level consists of a number of partially overlapping broad classes. Since the resulting classifiers are not statistically independent, we propose a new method for fusing the classifiers. Experimental results show that our two-level classifier obtained slightly better results when broad-class specific manifolds were learned, compared to a uniform manifold. However, the accuracy is still considerably lower than what could be obtained with *oracle* knowledge about broad class membership. From this we infer that phones do not form compact clusters in acoustic space.


*Keywords:* Manifold learning; Dimensionality reduction; Partial classification; Classifier fusion; Phone classification; TIMIT

,

## 1. Introduction

Phone classification (Singh-Miller and Collins, 2009; Huang et al., 2011a; Halberstadt and Glass, 1997; Dahl et al., 2010; Chang and Glass, 2007) is an attractive way to evaluate the power of acoustic modeling techniques for future use in automatic speech recognition (ASR) systems (Sainath et al., 2011; Lee and Hon, 1989) and for advancing our understanding of speech processing in general. Phone classification, even in a carefully read and manually labeled speech corpus such as TIMIT (Garofolo, 1988) appears to be surprisingly difficult, due to the wide range within which realizations of a phone can vary. As is well known, many sources contribute to the variation, the most important of which are differences between transmission channels and between speakers and differences due to phonetic context

☆ This paper has been recommended for acceptance by Katrin Kirchhoff.
* Corresponding author. Tel.: +31 243616069.
*E-mail address:* l.tenbosch@let.ru.nl (L. ten Bosch).

http://dx.doi.org/10.1016/j.csl.2015.12.002
0885-2308/

(which includes the preceding and following phones and the prosodic structure). Several effective techniques have been proposed for channel and speaker normalization, for example (de Veth and Boves, 1998; Claes et al., 1998). This paper focuses on the effects of the phonetic context. Speech is produced by semi-continuous ballistic movements of the articulators. As a result, there are seldom clear and unambiguous boundaries between successive phones, while a substantial part of the identity of a phone is encoded in the dynamic transitions out of the preceding and into the following phone, processes that are known as *co-articulation* in the phonetics literature.

Due to the impact of co-articulation it is extremely difficult to classify segments of a speech signal that are too short to capture the articulatory dynamics. $\Delta$ and $\Delta^2$ features cover part, but certainly not all, of the articulatory dynamics that is important for understanding speech and for classifying segments. There are several options for covering more of the dynamics, for example by using the modulation spectrum (e.g., Kanadera et al., 1998) or by using an autoregressive model of feature trajectories (e.g., Gish and Ng, 1996; Han et al., 2007). Another way for covering the articulatory dynamics, the approach investigated in this paper, is stacking short-time spectral features, e.g., Mel-Frequency Cepstral Coefficients (MFCC) or Perceptual Linear Prediction (PLP) coefficients (Sim and Gales, 2007; Han et al., 2007; Schuppler et al., 2009; Markov et al., 2006; Frankel et al., 2007). To fully capture the dynamics in the speech signals, for example at the syllable level, it appears necessary to stack up to about 25 frames of 10 ms (Hermansky and Jain, 2003; Jaitly et al., 2014). For languages such as English, which have complex syllable structures, such frame stacks comprise a very wide range of variation.

Variation induced by co-articulation is to a large extent systematic, and it reflects the intrinsic structure imposed by the limited number of degrees of freedom of the articulators (Sainath et al., 2010, 2011; Gemmeke et al., 2011; De Wachter et al., 2007). Thus, the distributions of the acoustic characteristics of phones may take the form of manifolds that reflect the phonetic contexts. Several recent papers, e.g., (Singh-Miller and Collins, 2009; Singh-Miller et al., 2007), have explored manifold learning for characterizing the underlying structure in the acoustic feature space. In Huang et al. (2011a,b) we showed that phone classification performance can be improved by means of manifold learning in reducing the dimensionality of the acoustic feature space. A similar approach was taken by Sakai et al. (2009) and Jafari and Almasganj (2010) to improve ASR performance. More often than not, research aiming at multiple goals, such as improving automatic speech recognition as well as advancing basic understanding of speech processing, must put more weight on one goal than on the other. In this paper we attach more importance to basic understanding than to performance per se. If we would have prioritized performance over understanding, we would probably have opted for an approach based on deep neural networks (Zhang et al., 2014; Weng et al., 2014; Kubo et al., 2014; Toth, 2014; Jaitly et al., 2014).

Stacking a number of speech frames is the simplest way for capturing dynamic changes in the signal, but is comes with a price, namely the high dimensionality of the resulting feature space, which calls for dimensionality reduction techniques. The graph-embedding framework (Yan et al., 2007) provides an elegant mathematical basis for research into dimensionality reduction based on manifold learning. In this unifying framework the relations between observations are represented in the form of a graph. In the graph-embedding framework it is easy to see that classical Linear Discriminant Analysis (Fisher, 1936) does not exploit the manifold structure, because the underlying graph is fully connected: each observation is connected to all other observations, and the weights on the edges do not depend on the distance between the observations (Sugiyama and Roweis, 2007). In our previous research we have shown that partially connected graphs, i.e., graphs in which observations are only connected to their closest neighbors (in the same class or in other classes) are more effective in capturing manifold structure than using distance-dependent weights in a fully connected graph (Huang et al., 2011a). In our previous research we aimed at a single transformation that maps the high-dimensional observations into a lower-dimensional space in such a manner that the manifold structure is kept, and the separability of the phones is maximized. However, no matter how precisely the manifolds are characterized, the use of a single transformation matrix yields a projection that treats all phones in the same manner (Kim and Kittler, 2005).

It has long been known that phones are best characterized by phone-specific features, and that separating specific phone pairs may require pair-specific features (Halberstadt and Glass, 1997). Scanlon et al. (2007) proposed using different "experts", each using a local set of features, for classifying the phones in different "Broad Phonetic Classes". The idea of exploiting different features for different phones is also explored in the proposal of Large-Margin Gaussian Mixture Models (Sha and Saul, 2006; Chang and Glass, 2007). Using class-dependent features is an active research topic in the field of pattern recognition (Kim and Kittler, 2005; Liu et al., 2011; Mahanta et al., 2012). Using different features for different broad classes requires a hierarchical classifier architecture, in which the broad classes are separated