Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 51 (2009) 151–166
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Broad phonetic classification using discriminative Bayesian networks

Franz Pernkopf [a,*], Tuan Van Pham [a,c], Jeff A. Bilmes [b]

[a] Signal Processing and Speech Communication Laboratory, Graz University of Technology, Inffeldgasse 12, A-8010 Graz, Austria
[b] Department of Electrical Engineering, University of Washington, Box 352500, Seattle, WA 98195-2500, USA
[c] Faculty of Electronics and Telecommunications, Danang University of Technology, 54 Nguyen Luong Bang, Danang, Vietnam

Received 30 July 2007; received in revised form 7 July 2008; accepted 21 July 2008

Abstract

We present an approach to broad phonetic classification, defined as mapping acoustic speech frames into broad (or clustered) phonetic categories. Our categories consist of silence, general voiced, general unvoiced, mixed sounds, voiced closure, and plosive release, and are sufficiently rich to allow accurate time-scaling of speech signals to improve their intelligibility in, e.g. voice-mail applications. There are three main aspects to this work. First, in addition to commonly used speech features, we employ acoustic time-scale features based on the intra-scale relationships of the energy from different wavelet subbands. Secondly, we use and compare against discriminatively learned Bayesian networks. By this, we mean Bayesian networks whose structure and/or parameters have been optimized using a discriminative objective function. We utilize a simple order-based greedy heuristic for learning discriminative structure based on mutual information. Given an ordering, we can find the discriminative classifier structure with $\mathcal{O}(N^q)$ score evaluations (where q is the maximum number of parents per node). Third, we provide a large assortment of empirical results, including gender dependent/independent experiments on the TIMIT corpus. We evaluate *both* discriminative *and* generative parameter learning on *both* discriminatively *and* generatively structured Bayesian networks and compare against generatively trained Gaussian mixture models (GMMs), and discriminatively trained neural networks (NNs) and support vector machines (SVMs). Results show that: (i) the combination of time-scale features and mel-frequency cepstral coefficients (MFCCs) provides the best performance; (ii) discriminative learning of Bayesian network classifiers is superior to the generative approaches; (iii) discriminative classifiers (NNs and SVMs) perform better than both discriminatively and generatively trained and structured Bayesian networks; and (iv) the advantages of generative yet discriminatively structured Bayesian network classifiers still hold in the case of missing features while the discriminatively trained NNs and SVMs are unable to deal with such a case. This last result is significant since it suggests that discriminative Bayesian networks are the most appropriate approach when missing features are common.

Keywords: Broad phonetic class recognition; Wavelet transform; Time-scale features; Bayesian networks; Discriminative learning

1. Introduction

Automatic broad speech unit classification is crucial for a number of different speech processing methods and various speech applications. We define *broad phonetic classification* as processing that maps a speech signal into a sequence of integers, where each integer represents a coarser-grained category than that of a phone. While mapping to a sequence of phones, or at least a distribution over such sequences, is a favored approach to automatic speech recognition (ASR), broad phonetic classification is useful for a number of distinct applications.

For example, some speech coding and compression systems use broad phonetic classification to determine the number of bits that should be allocated for each speech frame (Kubin et al., 1993). Such a source-controlled variable rate coder would for example allocate more bits to voiced and mixed frames than to unvoiced frames, and would assign only a few bits to silence frames (Zhang et al., 1997). In Internet telephony applications (Sanneck, 1998), for example, the adaptive loss concealment algorithm is based on a voiced/unvoiced detector at the sender.

* Corresponding author. Tel.: +43 316 873 4436; fax: +43 316 873 4432.
E-mail addresses: pernkopf@tugraz.at (F. Pernkopf), v.t.pham@tugraz.at (T. Van Pham), bilmes@ee.washington.edu (J.A. Bilmes).

0167-6393/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2008.07.003

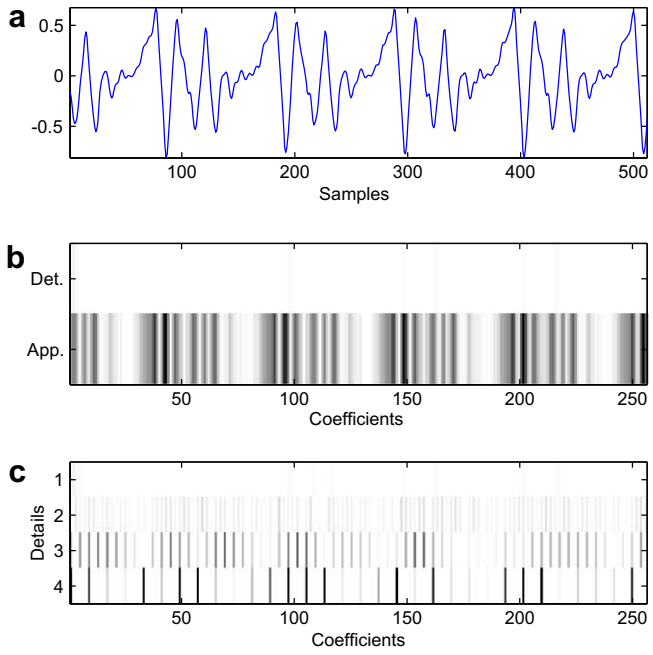

Fig. 1. (a) A voiced speech segment (phoneme /a/), (b) approximation (App.) and detail (Det.) coefficients derived at 1st scale DWT, (c) power variation of different detail subbands (the 2nd, 3rd, and 4th details were upsampled to have the same length as the 1st detail).

This helps the receiver to conceal the loss of information due to the similarity between the lost segments and the adjacent segments.

As another example, the utilization of information about broad phonetic classes can improve the perceptual quality of time-scaling algorithms for speech signals (Kubin and Kleijn, 1994) – a desirable capability in voice-mail and voice-storage applications as it allows the user to listen to messages in a fraction of the original recording time. A speech utterance can be efficiently time-scaled by applying different scaling factors to different speech segments, depending on the broad phonetic characteristics, without reducing its quality and naturalness (Donnellan et al., 2003). It was concluded in (Kuwabara and Nakamura, 2000) that voiced frames need to be more affected by time-scaling than mixed frames, and much more than unvoiced frames (Campbell and Isard, 1991). To maintain the characteristics of plosives or parts of plosives (a closure or release), time-scale modification should not be so applied. Silence frames, moreover, should be treated like voiced frames (Donnellan et al., 2003).

A broad phonetic classifier can also be used as a pre-classification step to support the phonetic transcription task of very large databases thereby making the transcriber's job much easier and less costly. Furthermore, it can be used as a step in addition to word labeling for preparing corpora for concatenative synthesis. Broad phonetic classification can also be fused into standard speech recognition systems at levels other than the acoustic feature vector (Subramanya et al., 2005; Bartels and Bilmes, 2007) and can also be used to facilitate out-of-vocabulary (OOV) detection (Lin et al., 2007). In order to improve robustness of automatic speech recognition, moreover, Kirchhoff et al. (2002) investigated the benefits of articulatory phonetics by using 28 articulatory features, both as an alternative to, and in combination with standard acoustic features for acoustic modeling. For a similar purpose, framewise phonetic classification of the TIMIT database has been performed using Gaussian mixture models (GMMs) for four manner classes (Halberstadt and Glass, 1997), and support vector machines (SVMs) (Salomon et al., 2002) and large margin GMMs (Fei and Saul, 2006) have been used for 39 phonetic classes. Recently, ratio semi-definite classifiers have been developed and applied to phoneme classification (Malkin and Bilmes, 2008).

In this article, several general-purpose broad phonetic classifiers have been developed for classifying speech frames into either four or six broad phonetic classes. Beside the silence class (S), we also consider a voiced class (V) which includes vowels, semivowels, diphthongs and nasals, an unvoiced class (U) which includes only unvoiced fricatives, and a mixed-excitation class (M) including voiced and glottal fricatives. Furthermore, we are interested in plosives that are formed by two parts, a closing and a release (R) of a vocal-tract articulator. Normally, plosives have a transient characteristic, whereas, voiced, unvoiced, and mixed sounds are continuant sounds. While the closed interval of unvoiced plosives is similar to silence, voiced plosives have a subtle voiced closure interval (VC) which has a periodic structure at very low power (Olive et al., 1993).

There are three main contributions of this work: (1) in tandem with more traditional acoustic features, we employ wavelet derived acoustic features that are useful to represent speech in, e.g. the aforementioned VC interval; (2) we use discriminatively learned Bayesian network classifiers and their comparison to standard discriminative models of various forms; and (3) we provide results that compare the various classifiers in particular in the case of missing acoustic features. These contributions are summarized in this section and then fully described within the article.

First, in order to improve the detection of subtle cues in our broad phonetic categories, we use wavelet derived features in addition to commonly used time domain (Kedem, 1986; Childers et al., 1989) and mel-frequency cepstral coefficients (MFCC) features. We extract time-scale features by applying the discrete wavelet transform (DWT) and then by performing additional processing thereafter (full details are given below). We show that the intra-scale relations of the energy from different wavelet subbands are beneficial to reflect the acoustic properties of our phonetic classes.

Numerous classification approaches have been proposed to classify speech units given a set of speech features in the past with one of the earliest being that of Atal and Rabiner (1976). In this work, by speech unit classification, we specifically mean frame-by-frame classification, where the speech signal has been segmented into overlapping fixed-length time windows, and where each window is then input to a classifier whose goal it is to decide what the correct cat-