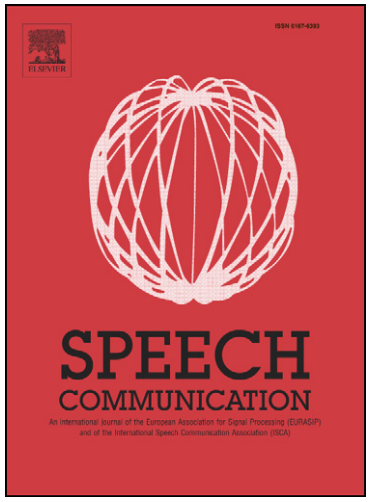
Broad Phonetic Classification Using Discriminative Bayesian Networks

Franz Pernkopf, Tuan van Pham, Jeff A. Bilmes



Broad Phonetic Classification Using Discriminative Bayesian Networks

Franz Pernkopf, Tuan Van Pham, Jeff A. Bilmes

PII: S0167-6393(08)00124-6
DOI: 10.1016/j.specom.2008.07.003
Reference: SPECOM 1740

To appear in: *Speech Communication*

Received Date: 30 July 2007
Revised Date: 7 July 2008
Accepted Date: 21 July 2008

Please cite this article as: Pernkopf, F., Van Pham, T., Bilmes, J.A., Broad Phonetic Classification Using Discriminative Bayesian Networks, *Speech Communication* (2008), doi: 10.1016/j.specom.2008.07.003

Broad Phonetic Classification Using Discriminative Bayesian Networks

Franz Pernkopf [1]

Signal Processing and Speech Communication Laboratory, Graz University of Technology, Inffeldgasse 12, A-8010 Graz, Austria

Tuan Van Pham

Signal Processing and Speech Communication Laboratory, Graz University of Technology, Inffeldgasse 12, A-8010 Graz, Austria

Jeff A. Bilmes

Dept. of Electrical Engineering, University of Washington, Box 352500, Seattle, Washington 98195-2500, USA

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.

Key words: 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. 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

Email addresses: pernkopf@tugraz.at (Franz Pernkopf), v.t.pham@tugraz.at (Tuan Van Pham), bilmes@ee.washington.edu (Jeff A. Bilmes).
[1] Corresponding author

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 and her colleagues (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 4 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 category is of the speech at the center of that window. This then becomes a standard pattern classification problem. Generally, there are two avenues for such classifiers, generative and discriminative (Jebara, 2001; Bilmes et al., 2001; Bahl et al., 1986; Ephraim et al., 1989; Ephraim and Rabiner, 1990; Juang and Katagiri, 1992; Juang et al., 1997; Bishop and Lasserre, 2007; Pernkopf and Bilmes, 2008). Let $\mathbf{X}_{1:N}$ be a set of N features and C be a class variable. Generative models in one way or another represent the joint distribution $p(\mathbf{X}_{1:N}, C)$ or at least $p(\mathbf{X}_{1:N}|C)$. Generative models can be trained either generatively (which means optimizing an objective function that is maximized when the joint distribution scores a data set highly, such as penalized maximum likelihood (ML)) or can also be trained discriminatively (which means to use a discriminative objective to train a generative model (Pernkopf and Bilmes, 2008)). Discriminative models are those that inherently represent either the conditional distribution $p(C|\mathbf{X}_{1:N})$ directly, or alternatively represent only the decision regions in $\mathbf{X}_{1:N}$ between classes, and are specified based on some discriminant function $f(\mathbf{X}_{1:N}, C)$ which have no normalization constraints (and thus are not guaranteed to provide a probabilistic interpretation, only the rank order is important). Discriminative models are trained using only discriminative objective functions, such as conditional likelihood or some form of exact or smoothed loss function (Bartlett et al., 2006).

Generative approaches (such as the Gaussian mixture model (Leung et al., 1993; Duda et al., 2001) or the hidden Markov Model (Levinson et al., 1989; Rabiner, 1989)) have in the past been used for phonetic classification as well as

speech recognition. Some of the most prominent discriminative models are neural networks (Bishop, 1995; Mitchell, 1997; Duda et al., 2001) (NNs) and support vector machines (Schölkopf and Smola, 2001; Burges, 1998) (SVMs) which have also been widely applied to the problem of speech classification Bourlard and Morgan (1994); Minghu et al. (1996); Salomon et al. (2002); Smith and Gales (2002); Pham and Kubin (2005); Borys and Hasegawa-Johnson (2005) although this limited set of references does not do the field justice.

Our second main contribution in this work is that we employ discriminatively learned Bayesian network classifiers. Specifically, we apply both discriminative parameter learning by optimizing conditional likelihood (CL) and generative maximum-likelihood (ML) parameter training on *both* discriminatively *and* generatively structured Bayesian networks. We use either CL or classification rate (CR) (equivalently, empirical risk) for producing discriminative structure. These classifiers are further restricted to be either naive Bayes (NB) classifiers (where all features are assumed independent given the class variable), and relaxations of such an approach (where the features are no longer presumed independent given the class, such as 1-tree or 2-tree augmented naive Bayes (TAN)). We use an algorithm for discriminative structure learning of Bayesian networks based on a computed variable order (Pernkopf and Bilmes, 2008). The proposed metric for establishing the ordering of the features is based on the conditional mutual information. Given a resulting ordering, we can find the discriminative network structure with $\mathcal{O}(N^q)$ score evaluations (constant q limits the number of parents per node). Hence, e.g. the TAN classifier can be discriminatively optimized in $\mathcal{O}(N^2)$ queries using either either CL or CR as a evaluative score function. We present results for framewise broad phonetic classification using the TIMIT database (Lamel et al., 1986). We provide classification results using Bayesian network classifiers on time-scale features and on MFCC features. Additionally, we compare our Bayesian network classifiers to GMMs, NNs, and SVMs on the joint time-scale and MFCC feature set. Gender dependent and gender independent experiments have been performed to assess the influence on the classification rate (CR).

A third contribution of our work is in the case of missing features. A primary advantage of our generative Bayesian networks over standard discriminative models (such as NNs and SVMs) is that they can be applied to cases where some of the features are at times missing (or known to be highly unreliable and thus useless). This is done essentially by marginalizing over the unknown (or unreliable) variables, something that is still possible since the model is inherently generative, even if it is discriminatively trained. Spectro-temporal regions of speech which are dominated by noise can, for example, be treated as missing or unreliable (Cooke et al., 2001; Raj and Stern, 2005). What is not known, however, is if discriminatively trained generative models still hold a performance advantage in the broad phonetic classification domain, something which we investigate and verify in this work. In particular, we

find that discriminatively trained Bayesian network classifiers still hold an advantage over generatively trained ones in the case of missing features.

The paper is organized as follows: Our DWT and multiresolution analysis is introduced in Section 2.1. Section 2.2 studies intra-scale relations of the energy from different wavelet subbands with respect to the phonetic classes. This section also introduces the time-scale features used for classification. Section 3 introduces Bayesian network classifiers and different network structures. The most commonly used approaches for generative and discriminative structure learning are summarized in Section 3.2. Section 3.3 describes our OMI heuristic for efficient discriminative structure learning. Experiments on the TIMIT database and the discussion are presented in Section 4. Section 5 concludes and gives perspectives for future research. The abbreviations are summarized in Appendix B.

2 Extraction of wavelet based time-scale features

2.1 Wavelet transform by multiresolution analysis

The potential advantage of a DWT in speech processing is its inherent multiresolution representation: namely, a DWT allows a multiscale representation of speech signals in the time-scale domain. In other words, various positions in the time-frequency plane are analyzed with different time-frequency resolutions. This allows e.g. higher frequencies to be granted the higher temporal resolution they naturally require, and lower frequencies to be granted the fine spectral resolution they require.

A discrete-time signal $x[k]$ can be represented as

$$x[k] = \sum_{m=1}^{M} \sum_{n=1}^{N_m} \langle \psi_{m,n}[k], x[k] \rangle \, \psi_{m,n}[k], \tag{1}$$

where $\langle \cdot \rangle$ denotes the inner product, M represents the number of scales, $N_m = \frac{N_f}{2^m}$ is the number of coefficients at the m^{th} scale, and N_f is the number of samples in one speech frame. The set of discrete-time wavelet basis functions $\psi_{m,n}[k] = a_0^{-m/2} \psi(a_0^{-m} k - n b_0)$ are generated by translating and dilating the mother wavelet $\psi(k)$ using iterated filters (Vetterli and Kovacevic, 1995). With $a_0 = 2$ and $b_0 = 1$ we obtain the dyadic-parameter wavelet basis functions. The discrete-time signal $x[k]$ can be further decomposed into the sum of one approximation plus M detail subbands at M resolution stages by a decimated

non-uniform filterbank as follows:

$$x[k] = \sum_{n=1}^{N_m} X^{(M)}[2n] \cdot g_0^{(M)}[k - 2^M n] + \sum_{m=1}^{M} \sum_{n=1}^{N_m} X^{(m)}[2n+1] \cdot g_1^{(m)}[k - 2^m n], \tag{2}$$

where $X^{(M)}[2n]$ and $X^{(m)}[2n+1]$ are the approximation coefficients (low-frequency part) and the detail coefficients (high-frequency parts) respectively. They are defined as:

$$\begin{aligned} X^{(M)}[2n] &= \left\langle h_0^{(M)}[2^M n - l], x[l] \right\rangle, \text{ and} \\ X^{(m)}[2n+1] &= \left\langle h_1^{(m)}[2^m n - l], x[l] \right\rangle, \end{aligned} \tag{3}$$

where $g_j^{(m)}[k]$ is an equivalent filter obtained through m stages of synthesis filters $g_j[k]$, each preceded by a factor of two upsampler, $h_j^m[k]$ is an equivalent analysis filter where $h_j^{(m)}[k] = g_j^{(m)}[-k]$, $j \in \{0,1\}$, $k, m, n \in \mathbb{Z}$. By applying the DWT at the scale $M = 4$ on each speech frame, we obtain one approximation subband and four detail subbands which form the sequence of wavelet coefficients $W_{4,i}[n] = \left\{ X^{(4)}[2n], \left(X^{(m)}[2n+1] \right)_{m \in \{1,2,3,4\}} \right\} = \left\{ X^{(4)}[2n], X^{(4)}[2n+1], X^{(3)}[2n+1], X^{(2)}[2n+1], X^{(1)}[2n+1] \right\}$. The number of coefficients in the 4^{th} approximation subband is $N_{a4} = \frac{N_f}{16}$, and the four following detail subbands are denoted as $\left\{ N_4 = \frac{N_f}{16}, N_3 = \frac{N_f}{8}, N_2 = \frac{N_f}{4}, N_1 = \frac{N_f}{2} \right\}$.

2.2 Intra-scale energy relations and feature extraction

The power distribution in different subbands varies and largely depends on the localized phonetic context. For our analysis, we apply our DWT at the 1^{st} decomposition scale on voiced, unvoiced, and mixed frames. We empirically observed that the power of wavelet coefficients derived from voiced frames is concentrated within the approximation part and not so much contained in the detail part as depicted in Figure 1b. The opposite is true for the unvoiced speech frames as shown in Figure 2b. For the mixed frames, the power difference between the approximation and detail coefficients is not as significant as it is for voiced and unvoiced frames as visualized in Figure 3b.

Additionally, an analysis of intra-scale relations is performed by considering the power change of the detail subbands at different scales. Figures 1c, 2c, and 3c show the power variation of the detail coefficients which are derived

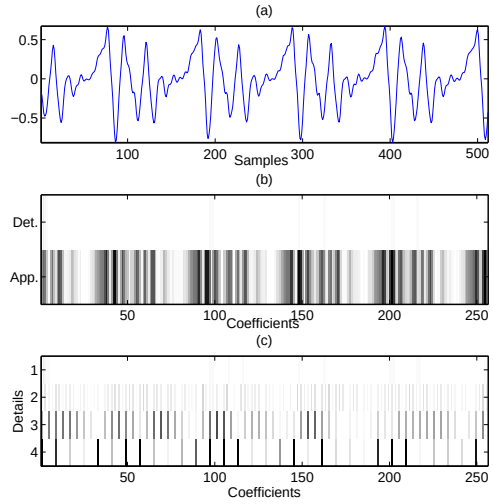


Fig. 1. (a) A voiced speech segment (phoneme /a/), (b) Approximation (App.) and detail (Det.) coefficients derived at 1^{st} scale DWT, (c) Power variation of different detail subbands (the 2^{nd}, 3^{rd}, and 4^{th} details were upsampled to have the same length as the 1^{st} detail).

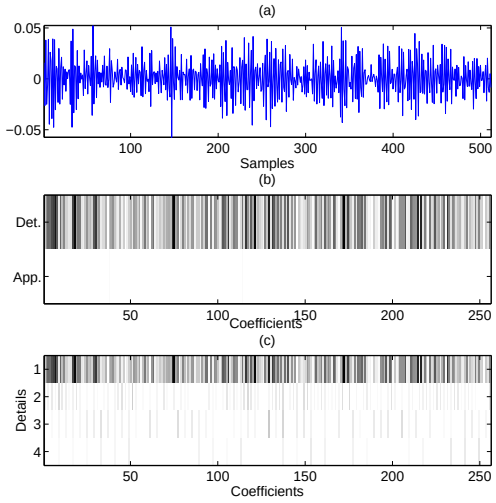


Fig. 2. (a) An unvoiced speech segment (phoneme /s/), (b) Approximation (App.) and detail (Det.) coefficients derived at 1^{st} scale DWT, (c) Power variation of different detail subbands (the 2^{nd}, 3^{rd}, and 4^{th} details were upsampled to have the same length as the 1^{st} detail).

at the 4^{th} decomposed scale of voiced, unvoiced, and mixed speech segments, respectively. The power of detail coefficients extracted from voiced frames increases from scale 1 to scale 4. However, the opposite is observed for unvoiced frames. There is less power change over various scales for mixed frames. All derived sequences of wavelet coefficients were normalized to their absolute maximum values.

Based on these observations, we extract several time-scale features (TSF) that

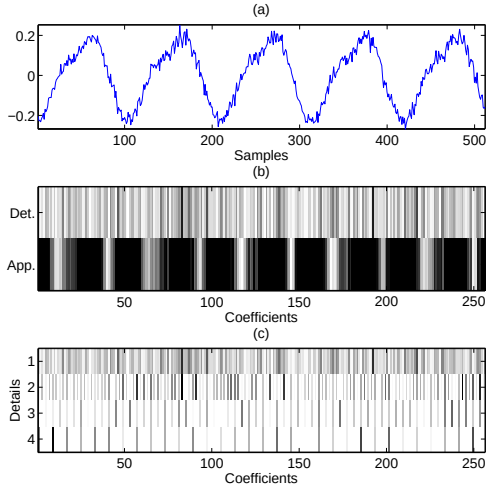


Fig. 3. (a) A mixed speech segment (phoneme /j/), (b) Approximation (App.) and detail (Det.) coefficients derived at 1^{st} scale DWT, (c) Power variation of different detail subbands (the 2^{nd}, 3^{rd}, and 4^{th} details were upsampled to have the same length as the 1^{st} detail).

should make it relatively easy to distinguish between the three different classes (V/U/M):

- ***Power delta*** *(PD)* is the power difference between the approximation and detail subbands at 4^{th} scale and the detail subband at 1^{st} scale:

$$PD(i) = \frac{1}{2N_4}\sum_{n=1}^{2N_4} W_{4,i}^2[n] - \frac{1}{N_1}\sum_{n=N_1+1}^{N_f} W_{4,i}^2[n]. \tag{4}$$

- ***First power ratio*** *(PR_1)* is the power ratio between the approximation subband at 4^{th} scale and the three detail subbands from 4^{th} scale to 2^{nd} scale:

$$PR_1(i) = \frac{N_4 + N_3 + N_2}{N_4}\frac{\sum_{n=1}^{N_4} W_{4,i}^2[n]}{\sum_{n=N_4+1}^{N_1} W_{4,i}^2[n]}. \tag{5}$$

- ***Second power ratio*** *(PR_2)* is the power ratio between the two detail subbands of the 2^{nd} and 1^{st} scale and the approximation and detail subbands at the 4^{th} scale:

$$PR_2(i) = \frac{2N_4}{N_2 + N_1}\frac{\sum_{n=N_2+1}^{N_f} W_{4,i}^2[n]}{\sum_{n=1}^{2N_4} W_{4,i}^2[n]}. \tag{6}$$

We also see that the VC interval of voiced plosives shows a periodic structure similar to voiced sounds and a slightly higher energy in the approximation subband derived at the 4^{th} scale compared to silence. Hence, we use the power of the approximation subband as a feature. Furthermore, to detect the weak periodic structure of VC and some voiced consonants (mixed sounds), we derive a peak delta feature from the autocorrelation function (estimated for each

speech frame). While the release of plosives has a similar energy distribution over the subbands compared to unvoiced sounds, it shows a lower standard deviation of the detail coefficients at the 1^{st} scale. These features are summarized in the following:

- ***Power of approximation*** *(PA)* subband at 4^{th} scale:

$$PA(i) = \frac{1}{N_4} \sum_{n=1}^{N_4} W_{4,i}^2[n]. \tag{7}$$

- ***Peak delta*** *(PeD)* is defined for every speech frame as follows:

$$PeD(i) = R_i(j_1) - R_i(j_2), \tag{8}$$

  where R_i is the autocorrelation function of speech frame i. A distance between the peak values of the central lobe (at lag j_1) and the first lobe (at lag j_2) is calculated. To select the peak value of the first lobe properly, we first smooth the normalized autocorrelation coefficients by using a first order recursive filter, then the smoothed coefficients are sorted and finally the second biggest coefficient is chosen from the set of ranked coefficients.
- ***Standard deviation*** *(SD)* of coarsest detail subband derived at 1^{st} scale:

$$SD(i) = \sqrt{\frac{1}{N_1} \sum_{n=N_1+1}^{N_f} (W_{4,i}[n] - \overline{W_{4,i}[n]})^2}. \tag{9}$$

Finally, some statistical measures from the time domain (Kedem, 1986; Childers et al., 1989) such as the short-term energy and the zero crossing rate are also used. The zero crossing rate of voiced sounds is low compared to unvoiced sounds.

- ***Logarithmic short-term energy*** *(LgSE)*:

$$LgSE = 0.5 + \frac{16}{\ln(2)} \ln\left(1 + \frac{\sum_{k=1}^{N_f} x[k]^2}{32}\right). \tag{10}$$

- ***Zero crossing rate*** *(ZCR)* is the number of sign changes of successive samples in a speech frame:

$$ZCR = \sum_{k=1}^{N_f} |\mathrm{sgn}(x[k]) - \mathrm{sgn}(x[k-1])|. \tag{11}$$

Figure 4 displays the trajectory of the extracted features for an utterance with all 6 phonetic classes from the TIMIT database.

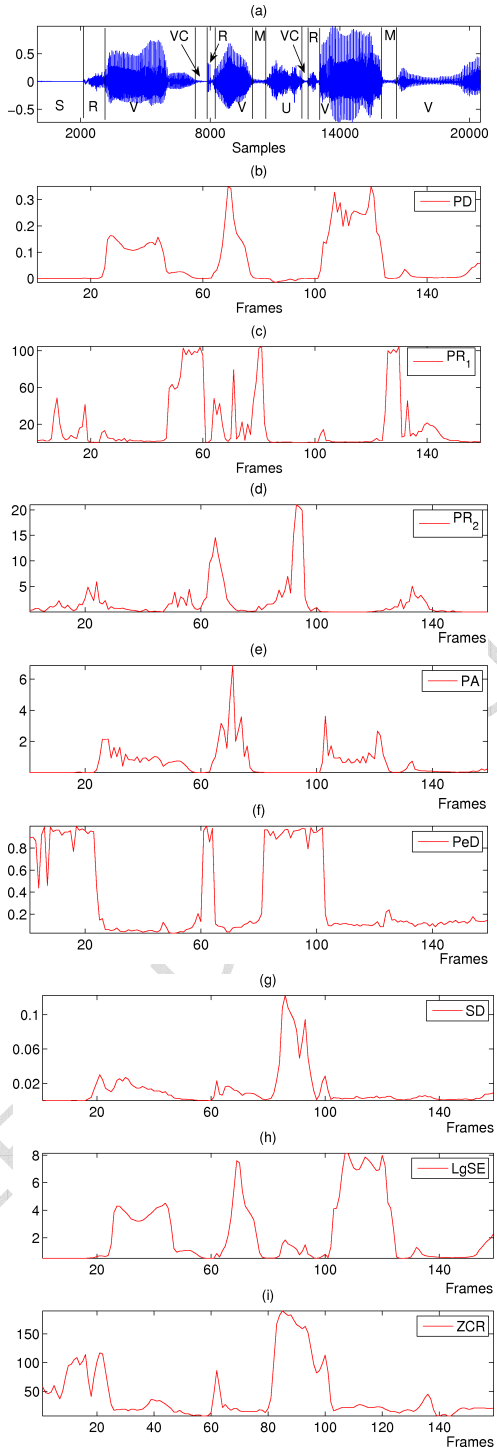


Fig. 4. (a) Waveform of an utterance with phonetic classes transcribed from the sequence of phonemes {/sil/}, {/p/}, {/ae/,/m/}, {/gcl/}, {/g/}, {/ih/}, {/v/}, {/s/}, {/dcl/}, {/d/}, {/r/,/ay/}, {/v/}, {/iy/}. The curly brackets of the phonemes mark the phonetic classes in (a). (b) Power delta (PD), (c) First power ratio (PR_1), (d) Second power ratio (PR_2), (e) Power of approximation (PA), (f) Peak delta (PeD), (g) Standard deviation (SD), (h) Logarithmic short-term energy ($LgSE$), (i) Zero crossing rate (ZCR).

3 Bayesian network classifier

A Bayesian network (Pearl, 1988; Cowell et al., 1999) $\mathcal{B} = \langle \mathcal{G}, \Theta \rangle$ is a directed acyclic graph $\mathcal{G} = (\mathbf{Z}, \mathbf{E})$ consisting of a set of nodes $\mathbf{Z}$ and a set of directed edges $\mathbf{E} = \left\{ E_{Z_i,Z_j}, E_{Z_i,Z_k}, \ldots \right\}$ connecting the nodes where E_{Z_i,Z_j} is an edge from Z_i to Z_j. This graph represents factorization properties of the distribution of a set of random variables $\mathbf{Z} = \{Z_1, \ldots, Z_{N+1}\}$. Each variable in $\mathbf{Z}$ has values denoted by lower case letters $\{z_1, z_2, \ldots, z_{N+1}\}$. We use boldface capital letters, e.g. $\mathbf{Z}$, to denote a set of random variables and correspondingly lower case boldface letters denote a set of instantiations (values). Without loss of generality, in Bayesian network classifiers the random variable Z_1 represents the class variable $C \in \{1, \ldots, |C|\}$, $|C|$ is the cardinality of C corresponding to the number of classes, $\mathbf{X}_{1:N} = \{X_1, \ldots, X_N\} = \{Z_2, \ldots, Z_{N+1}\}$ denote the set of random variables of the N attributes of the classifier. Each graph node represents a random variable, while the lack of edges specifies conditional independence properties. Specifically, in a Bayesian network each node is independent of its non-descendants given its parents. These conditional independence relationships reduce both number of parameters and required computation. The set of parameters which quantify the network are represented by Θ. Each node Z_j is represented as a local conditional probability distribution given its parents Z_{Π_j}. The joint probability distribution of the network is determined by the local conditional probability distributions as

$$P_\Theta(\mathbf{Z}) = \prod_{j=1}^{N+1} P_\Theta\left(Z_j | Z_{\Pi_j}\right). \tag{12}$$

Discriminative parameter learning by optimizing the CL and generative parameter learning, i.e. ML estimation and are summarized in Greiner and Zhou (2002); Pernkopf and Bilmes (2005) and in Pearl (1988), respectively. One of the key advantages of Bayesian networks over discriminative models (NN and SVM) is that it is easy to work with missing features by marginalizing over the unknown variables. Missing-feature approaches are useful in robust automatic speech recognition applications (Cooke et al., 2001; Raj and Stern, 2005; Parveen and Green, 2004).

3.1 Bayesian network structures

In this paper, we restrict the Bayesian network classifier to NB, TAN, and 2-tree structures, which we describe soon below. The NB network assumes that all the attributes are conditionally independent given the class label. As reported in the literature (Friedman et al., 1997), the performance of the NB classifier is surprisingly good even if the conditional independence assumption

between attributes is unrealistic in most of the data. The structure of the naive Bayes classifier represented as a Bayesian network is illustrated in Figure 5a.

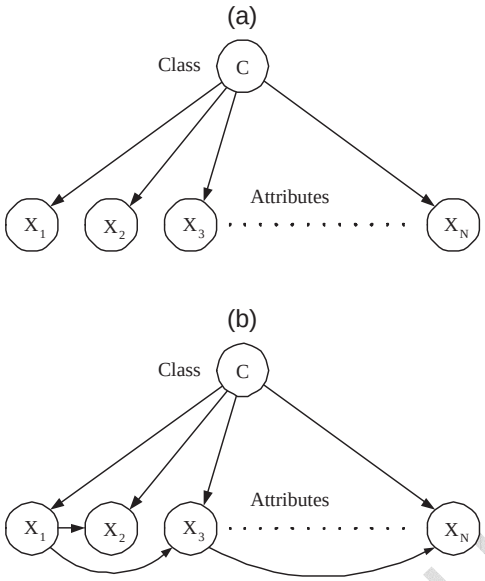


Fig. 5. Bayesian network: (a) NB, (b) TAN.

In order to correct some of the limitations of the NB classifier, Friedman et al. (1997) introduced the TAN classifier. A TAN is based on structural augmentations of the NB network, where additional edges are added between attributes in order to relax some of the most flagrant conditional independence properties of NB. Each attribute may have at most one other attribute as an additional parent which means that the tree-width of the attribute induced sub-graph is unity, i.e. we have to learn a 1-tree over the attributes. The maximum number of edges added to relax the independence assumption between the attributes is $N-1$. Thus, two attributes might not be conditionally independent given the class label in a TAN. An example of a TAN network is shown in Figure 5b. A TAN network is typically initialized as a NB network. Additional edges between attributes are determined through structure learning. An extension of the TAN network is to use a k-tree, i.e. each attribute can have a maximum of k attribute nodes as parents. TAN and 2-tree structures are restricted to a parentless class node $C_{\Pi} = \emptyset$. Many different network topologies have been suggested in the past. A good overview is provided in Acid et al. (2005).

3.2 Structure learning of TAN

Maximizing the CL for Bayesian networks is hard since CL is not decomposable, i.e. there is no efficient solution (Friedman et al., 1997) since it does not factorize. Recently, heuristic approaches have been suggested to learn the

structure and/or parameters discriminatively by maximizing the CL or the classification rate (CR). In Greiner and Zhou (2002); Greiner et al. (2005), logistic regression is extended to more general Bayesian networks – they optimize parameters with respect to the CL using a conjugate gradient method. Similarly, Wettig et al. (2003); Roos et al. (2005) provide conditions for general Bayesian networks under which correspondence to logistic regression holds. In Grossman and Domingos (2004) the CL function is used to learn a discriminative structure. The parameters are set using ML learning but they use a greedy hill climbing search with the CL function as scoring metric, where at each iteration one edge is added to the structure which conforms to the restrictions of the network topology (e.g. tree augmented naive Bayes) and the acyclicity property of Bayesian networks. In a similar algorithm, the CR has also been used for discriminative structure learning (Keogh and Pazzani, 1999; Pernkopf, 2005). These structure learning algorithms are computationally expensive. Many generative structure learning algorithms have been proposed and are overviewed in Heckerman (1995); Murphy (2002); Jordan (1999); de Campos (2006). An experimental comparison of discriminative and generative parameter training on both discriminatively and generatively structured Bayesian network classifiers has been performed in Pernkopf and Bilmes (2005, 2008). In the experiments we use the generative structure learning approach from Friedman et al. (1997) and the discriminative greedy hill climbing search using the CR introduced in Keogh and Pazzani (1999); Pernkopf (2005).

In the following, we describe a variety of both generative and discriminative structure and parameter learning algorithms that we later use to build Bayesian network classifiers for broad phonetic classification. One of the algorithms (the OMI algorithm) has been developed by the authors in previous work (Pernkopf and Bilmes, 2008) and successfully applied to entirely different data.

3.2.1 Generative structure learning

The conditional mutual information (CMI) (Cover and Thomas, 1991) between the attributes given the class variable is defined as

$$I\left(X_i; X_j|C\right) = E_{P(X_i,X_j,C)} \log \frac{P\left(X_i, X_j|C\right)}{P\left(X_i|C\right) P\left(X_j|C\right)}. \tag{13}$$

This measures the information between X_i and X_j in the context of C. Friedman et al. (1997) provide an algorithm for constructing a TAN network using this measure. This is an extension of the algorithm in (Chow and Liu, 1968), and is summarized herein:

(1) Compute the pairwise CMI $I\left(X_i; X_j|C\right) \quad \forall \quad 1 \leq i \leq N$ and $i < j \leq N$.

(2) Build a undirected 1-tree using the maximal weighted spanning tree algorithm (Kruskal, 1956) where each edge connecting X_i and X_j is weighted by $I(X_i; X_j|C)$.
(3) Transform the undirected 1-tree to a directed tree. In other words, select a root variable and direct all edges away from this root. Add to this tree the class node C and the edges from C to all attributes $X_1, \ldots, X_N$.

3.2.2 Greedy Discriminative structure learning

In the case of the TAN structure, the network is initialized to NB and with each iteration we add the edge which gives the largest improvement of the scoring function. The greedy hill climbing search is terminated when there is no edge which further improves the score. This means that we might get a partial 1-tree (forest) when learning the TAN structure.

As a scoring function, the CR (Keogh and Pazzani, 1999; Pernkopf, 2005)

$$CR(\mathcal{B}_{\mathcal{S}}|\mathcal{S}) = \frac{1}{M_{\mathcal{S}}} \sum_{m=1}^{M_{\mathcal{S}}} \delta\left(\mathcal{B}_{\mathcal{S}}\left(\mathbf{x}_{1:N}^m\right), c^m\right) \tag{14}$$

or the CL (Grossman and Domingos, 2004)

$$CL(\mathcal{B}|\mathcal{S}) = \prod_{m=1}^{M_{\mathcal{S}}} P_\Theta\left(C = c^m|\mathbf{X}_{1:N} = \mathbf{x}_{1:N}^m\right) \tag{15}$$

can be used for learning a discriminative network structure. The expression $\delta\left(\mathcal{B}_{\mathcal{S}}\left(\mathbf{x}_{1:N}^m\right), c^m\right) = 1$ if the Bayesian network classifier $B_{\mathcal{S}}\left(\mathbf{x}_{1:N}^m\right)$ trained with samples in $\mathcal{S}$ assigns the correct class label c^m to the attribute values $\mathbf{x}_{1:N}^m$ and 0 otherwise (this therefore corresponds to empirical risk based on 0/1 loss). The training data consists of $M_{\mathcal{S}}$ samples $\mathcal{S} = \{\mathbf{z}^m\}_{m=1}^{M_{\mathcal{S}}} = \{(c^m, \mathbf{x}_{1:N}^m)\}_{m=1}^{M_{\mathcal{S}}}$.

This approach is computationally the most expensive one we consider, as a complete re-evaluation of the training set is needed for each considered edge. The CR, however, is the discriminative criterion that is perhaps closest to the ideal criterion (true risk minimization), so we suspect that it may do well. There are two approaches for accelerating this algorithm that we utilize:

(1) The data samples are reordered during structure learning so that misclassified samples from previous evaluations are classified first. The classification is terminated as soon as the performance drops below the currently best network score (Pazzani, 1996).
(2) During structure learning the parameters are set to the ML values. When learning the structure we only have to update the parameters of those nodes where the set of parents Z_{Π_j} changes.

3.3 The OMI algorithm

In this section, we describe our order-based greedy search heuristic (Pernkopf and Bilmes, 2008) for efficient learning of the discriminative structure of a Bayesian network classifier. It was first noticed in Buntine (1991); Cooper and Herskovits (1992) that the best network consistent with a given variable ordering can be found with $\mathcal{O}(N^q)$ score evaluations where q is the upper bound of parents per node. Our procedure first looks for a total ordering $\prec$ of the variables $\mathbf{X}_{1:N}$ according to the conditional mutual information (Cover and Thomas, 1991). If the graph is consistent with the ordering $X_i \prec X_j$ then the parent $X_{\Pi_j} \in \mathbf{X}_{\Pi_j}$ is one of the variables which appears before X_j in the ordering, where $\mathbf{X}_{\Pi_j}$ is the set of possible parents for X_j. This constraint ensures that the network stays acyclic. In the second step of the algorithm, we select X_{Π_j} for X_j under constant k maximizing either CL or CR. A generative structure learning approach over the space of orderings using a decomposable score was presented in Teyssier and Koller (2005). Unlike this approach, we establish only one ordering of variables and the goal is to learn a discriminative structure. Our scoring cost is discriminative, and thus does not decompose (Friedman et al., 1997).

3.3.1 Step 1: Establishing an order $\prec$

Our simple heuristic provides an ordering $\prec$ of the nodes using conditional mutual information. The mutual information $I(C;\mathbf{X}_{1:N})$ measures the degree of dependence between the features $\mathbf{X}_{1:N}$ and the class, and we have that $I(C;\mathbf{X}_{1:N}) = H(C) - H(C|\mathbf{X}_{1:N})$ where the negative entropy $-H(C|\mathbf{X}_{1:N}) = E_{P(C,\mathbf{X}_{1:N})} \log P(C|\mathbf{X}_{1:N})$ is related to what ideally should be optimized.

Our approach of finding an order first chooses a feature that is most informative about C. The next node in the order is the node that is most informative about C conditioned on the first node. More specifically, our algorithm forms an ordered sequence of nodes $\mathbf{X}_{\prec}^{1:N} = \left\{X_{\prec}^1, X_{\prec}^2, \ldots, X_{\prec}^N\right\}$ according to

$$X_{\prec}^j \leftarrow \arg \max_{X \in \mathbf{X}_{1:N} \setminus \mathbf{X}_{\prec}^{1:j-1}} \left[I\left(C; X | \mathbf{X}_{\prec}^{1:j-1}\right) \right], \tag{16}$$

where $j \in \{1, \ldots, N\}$. The first node $X_{\prec}^1$ is the node with the largest information about C, i.e. it is most important for C. The next node $X_{\prec}^2$ is the node among the remaining nodes $\mathbf{X}_{1:N} \setminus \{X_{\prec}^1\}$ which leads to the largest $I(C; X_{\prec}^2 | X_{\prec}^1)$ and so forth. We also note that any mutual information query can be computed efficiently making use of the sparsity of the joint probability distribution (i.e. by essentially making one pass over the training data). Therefore, if we make a polynomial number of mutual-information queries, we have an algorithm that is polynomial in the number of training data samples.

3.3.2 Step 2: Select parent

Once we have the ordering $\mathbf{X}_{\prec}^{1:N}$, we select $X_{\Pi_j} \in \mathbf{X}_{\Pi_j} = \mathbf{X}_{\prec}^{1:j-1}$ for each $X_{\prec}^{j}$ ($j \in \{3, \ldots, N\}$). When the size of $\mathbf{X}_{\Pi_j}$ (i.e. N) and of k are small we can even use a computational costly scoring function to find X_{Π_j}. In case of a large N, we can restrict the size of the parent set $\mathbf{X}_{\Pi_j}$ similar to the *sparse candidate algorithm* (Friedman et al., 1999). Basically, either the CL or the CR can be used as a cost function to select the parents for learning a discriminative structure. We restrict our experiments to CR for parent selection and call our algorithm OMI-CR (empirical results show it performed better). We connect a parent to $X_{\prec}^{j}$ only when CR is improved, and otherwise leave $X_{\prec}^{j}$ parentless. This therefore might result in a partial 1-tree (forest) over the attributes. Our algorithm can be easily extended to learn k-trees ($k > 1$) by choosing more than one parent, leading to an $\mathcal{O}\left(N^{1+k}\right)$ algorithm (corresponds to $\mathcal{O}\left(N^{q}\right)$). The OMI-CR algorithm is summarized in Appendix A and more details are given in Pernkopf and Bilmes (2008).

4 Experiments

We present results for framewise broad phonetic classification using the TIMIT database. We provide classification results using Bayesian network classifiers on time-scale features (TSF) and on MFCC features (see Section 4.3). In Section 4.4, we compare our Bayesian network classifiers using NB, TAN, and 2-tree structures to GMMs, SVMs, and NNs on the joint TSF and MFCC feature set. Additionally, we give a comparison of OMI-CR to random orderings at the end of the current section.

Different combinations of the following parameter/structure learning approaches are evaluated for use to learn the Bayesian network classifiers:

- Generative (ML) (Pearl, 1988) and discriminative (CL) (Greiner et al., 2005) parameter learning.
- CMI: Generative structure learning using CMI as proposed in Friedman et al. (1997) (see Section 3.2.1).
- CR: Discriminative structure learning with the naive greedy heuristic using CR as scoring function (Keogh and Pazzani, 1999; Pernkopf, 2005) (see Section 3.2.2).
- OMI-CR: Discriminative structure learning using CMI for ordering the variables (step 1) and CR for parent selection in step 2 of the order-based heuristic (see Section 3.3).
- RO-CR: Discriminative structure learning using a random ordering (RO) in step 1 and CR for parent selection in step 2 of the order-based heuristic. This method is used as a comparison against our ordering heuristic described

above.

- For the order-based heuristic OMI-CR, we propose discriminative parameter learning by optimizing CL during the selection of the parent in step 2 (for which we utilize the abbreviation OMI-CRCL). All abbreviations are summarized in Appendix B. Discriminative parameter learning while optimizing the discriminative structure of the network is computationally feasible only on rather small data sets due to the computational costs of the conjugate gradient parameter optimization.

4.1 Experimental setup

For our Bayesian network classifiers, any continuous features were discretized using the recursive minimal entropy partitioning (Fayyad and Irani, 1993) where the codebook is produced using only the training data. This discretization method uses the class entropy of candidate partitions to determine the bin boundaries. The candidate partition with the minimal entropy is selected. This is applied recursively on the established partitions and the Minimum Description Length is used as stopping criteria for the recursive partitioning. In Dougherty et al. (1995), an empirical comparison of different discretization methods has been performed and the best results have been achieved with this entropy-based discretization.

Throughout our experiments, we use exactly the same data partitioning for each training procedure. We performed simple smoothing, where zero probabilities in the conditional probability tables are replaced with small values ($\varepsilon = 0.00001$). For discriminative parameter learning, the parameters are initialized to the values obtained by the ML approach. In Greiner et al. (2005), it has been empirically observed that this is a good strategy. The termination of gradient descent occurs after either the change in scores falls below a threshold (2%), or after a specified maximum number of iterations (currently 20) has been performed. Greiner et al. (2005) introduce a variant of cross validation on the *training* data to establish the optimal stopping point for parameter optimization

4.2 Data characteristics

Experiments have been performed on the data from the TIMIT speech corpus. The standard NIST sets of 462 speakers and 168 speakers have been used for training and testing, respectively. Framewise classification accuracies are reported for 1344 utterances from 168 speakers. In the experiments, we only use the *sx* and *si* sentences since the *sa* sentences introduce a bias for certain phonemes in a particular context. The speech is sampled at 16 kHz

and the DWT is applied at the 4^{th} scale on windowed speech frames of 16ms length and 8ms overlap (similarly for the MFCCs). We perform speaker independent experiments with only four classes V/U/S/M and all six classes V/U/S/M/VC/R using 1691462 and 1886792 samples, respectively. The class distribution of the four class experiment V/U/S/M is 58.63%, 14.69%, 23.36%, 3.32% and of the six class case V/U/S/M/VC/R is 52.55%, 13.17%, 20.94%, 2.97%, 3.54%, 6.81%. Additionally, we perform classification experiments on data of male speakers (Ma), female speakers (Fe), and both genders (Ma+Fe). Speakers in the training set do not appear in the test set and vice versa. The classification experiments have been performed with 8 TSF, 13 MFCC (log-energy included) features, and the combination of both feature sets.

4.3 Classification results on TSF and MFCC features

Table 1 presents the classification rate for different generative/discriminative Bayesian network classifiers for 4 and 6 phonetic classes.

For TAN structures, the proposed time-scale features perform slightly better on the Ma+Fe and the Ma data set for both 4 and 6 phonetic classes than the baseline MFCC features, whereas, for Fe data, MFCC mostly outperforms TSF. In contrast, the NB classifier performs slightly better on the Ma, Fe, and Ma+Fe data using MFCCs compared to TSF features. One reason for this might be that the MFCCs are features whose elements tend in practice to be fairly statistically independent, while this is not as much the case with TSF. Hence, the TSF seem to be more suitable for the TAN classifier since it can model the dependency between attributes.

By considering two more categories VC and R (i.e. 6 classes), the classification accuracy drops by $\sim 7\%$ on average. For TSF we have only 8 features compared to 13 MFCC features. This results in a lower complexity of the classifier and faster learning. The small differences of classification performance between Ma+Fe, Ma, and Fe open an approach for gender independent broad phonetic classification.

The CR objective function for structure learning produces the best performing network structures. However, the evaluation of the CR measure is computationally very expensive, since a complete re-evaluation of the training set is needed for each considered edge. However, due to the ordering of the variables in the order-based heuristics, we can reduce the number of CR evaluations from $\mathcal{O}(N^3)$ to $\mathcal{O}(N^2)$ for TAN structures (TAN-CR versus TAN-OMI-CR). The order-based heuristic TAN-OMI-CR achieves a similar performance at lower computational cost.

Discriminative parameter learning (CL) produces (most often) a slightly but

Table 1: Classification rate in [%] for 4 and 6 classes with standard deviation. Best results use bold font. The bottom line gives the average performance for each classification approach.

Classifier			NB		TAN		TAN		TAN		TAN	
Struct. Learn.			-		CMI		OMI-CR		OMI-CRCL		CR	
Param. Learn.			ML	CL	ML	CL	ML	CL	ML	CL	ML	CL
Data set	Features	Class										
Ma+Fe	TSF	4	87.78	87.91	90.66	90.69	91.05	91.07	91.05	91.07	**91.17**	**91.20**
			± 0.07	± 0.07	± 0.06	± 0.06	± 0.06	± 0.06	± 0.06	± 0.06	± 0.06	± 0.06
Ma	TSF	4	87.74	87.93	90.84	90.87	90.82	90.83	90.82	90.83	**91.28**	**91.29**
			± 0.09	± 0.09	± 0.08	± 0.08	± 0.08	± 0.08	± 0.08	± 0.08	± 0.07	± 0.07
Fe	TSF	4	87.82	87.92	90.04	90.07	90.28	90.31	90.28	90.31	**90.71**	**90.74**
± 0.12	± 0.12	± 0.11	± 0.11	± 0.11	± 0.11	± 0.11	± 0.11	±			0.11	± 0.11
Ma+Fe	MFCC	4	88.06	88.17	90.03	90.05	**90.74**	**90.75**	**90.74**	**90.75**	**90.69**	**90.71**
			± 0.07	± 0.07	± 0.06	± 0.06	± 0.06	± 0.06	± 0.06	± 0.06	± 0.06	± 0.06
Ma	MFCC	4	88.09	88.19	90.21	90.22	**90.78**	**90.78**	**90.78**	**90.78**	**90.83**	**90.83**
			± 0.09	± 0.09	± 0.08	± 0.08	± 0.08	± 0.08	± 0.08	± 0.08	± 0.08	± 0.08
Fe	MFCC	4	88.30	88.43	89.86	89.89	90.52	90.52	90.59	90.60	90.71	90.73
			± 0.12	± 0.12	± 0.11	± 0.11	± 0.11	± 0.11	± 0.11	± 0.11	± 0.11	± 0.11
Ma+Fe	TSF	6	80.45	80.53	82.47	82.50	**83.20**	**83.21**	**83.20**	**83.21**	**83.25**	**83.27**
			± 0.08	± 0.08	± 0.08	± 0.08	± 0.08	± 0.08	± 0.08	± 0.08	± 0.08	± 0.08
Ma	TSF	6	80.28	80.37	82.34	82.39	**83.26**	**83.27**	**83.26**	**83.27**	**83.24**	**83.25**
			± 0.10	± 0.10	± 0.10	± 0.10	± 0.09	± 0.09	± 0.09	± 0.09	± 0.09	± 0.09
Fe	TSF	6	80.97	81.02	82.55	82.60	82.86	82.88	82.86	82.88	**83.17**	**83.20**
			± 0.14	± 0.14	± 0.14	± 0.13	± 0.13	± 0.13	± 0.13	± 0.13	± 0.13	± 0.13
Ma+Fe	MFCC	6	80.91	80.95	82.54	82.57	**83.22**	**83.23**	**83.25**	**83.26**	**83.25**	**83.26**
			± 0.08	± 0.08	± 0.08	± 0.08	± 0.08	± 0.08	± 0.08	± 0.08	± 0.08	± 0.08
Ma	MFCC	6	80.82	80.88	82.45	82.47	**83.09**	**83.10**	**83.09**	**83.10**	**83.05**	**83.05**
			± 0.10	± 0.10	± 0.10	± 0.10	± 0.09	± 0.09	± 0.09	± 0.09	± 0.09	± 0.09
Fe	MFCC	6	81.18	81.23	82.47	82.49	82.88	82.88	82.90	82.91	**83.16**	**83.16**
			± 0.14	± 0.14	± 0.14	± 0.14	± 0.13	± 0.13	± 0.13	± 0.13	± 0.13	± 0.13
Average			84.37	84.46	86.37	86.40	86.89	86.90	86.90	86.91	87.04	87.06

not significantly better classification performance than ML parameter learning for all different classification approaches.

Discriminative parameter learning during discriminative structure learning using our order-based heuristics (TAN-OMI-CRCL) can slightly improve the performance. However, the average performance is similar to TAN-OMI-CR. This is possible only on small data sets due to the computational costs for the conjugate gradient parameter optimization.

Table 2 presents a summary of the classification results over all experiments from Table 1. We compare all pairs of classifiers (with ML parameter learning) using the one-sided paired t-test (Mitchell, 1997). The t-test determines whether the classifiers differ significantly under the assumption that the paired classification differences over the data sets are independent and identically normally distributed. In this table, each entry gives the significance of the difference in classification accuracy of two classification approaches. The arrow points to the superior learning algorithm and a double arrow indicates whether the difference is significant at a level of 0.005.

This table shows that discriminative structure learning, TAN-OMI-CR, TAN-OMI-CRCL, and TAN-CR, significantly outperform generative structure learning. However, TAN-CR does not significantly outperform our discriminative structure learning approaches TAN-OMI-CR and TAN-OMI-CRCL.

Table 2
Comparison of different classifiers (with ML parameter learning) using the one-sided paired t-test: Each entry of the table gives the significance of the difference of the classification accuracy of two classifiers over the data sets. The arrow points to the superior learning algorithm. We use a double arrow if the difference is significant at the level of 0.005.

Classifier Struct. Learn. Param. Learn.	TAN CMI ML	TAN OMI-CR ML	TAN OMI-CRCL ML	TAN CR ML
NB-ML	⇑<0.00001	⇑<0.00001	⇑<0.00001	⇑<0.00001
TAN-CMI-ML		⇑0.00001	⇑0.00001	⇑<0.00001
TAN-OMI-CR-ML			↑0.05215	↑0.00704
TAN-OMI-CRCL-ML				↑0.01023

4.4 Classification results on the joint TSF and MFCC feature set

We observed that the TSF and the MFCC features complement each other. Due to this fact, we report classification results on the joint feature space. In addition to the previous experiment, we compare our Bayesian network classifiers to state-of-the-art discriminative classifiers, i.e. NNs and SVMs, and to the generative GMM which is popular in many speech applications.

Since the log-energy is contained in both feature sets we removed it so that it only occurs once.

Table 3 summarizes the classification performance for different generative/discriminative Bayesian network classifiers for 4 and 6 phonetic classes on the joint TSF and MFCC features. Additionally, we compare the Bayesian network classifiers to the following generative and discriminative classification approaches:

- NB-Cont: Naive Bayes classifier on continuous features. We use a Gaussian distribution to model the features so this really becomes a single diagonal covariance Gaussian model.
- GMM-500: Gaussian mixture model with 500 components. We use diagonal covariance matrices for each Gaussian component. The main reasons for this are that they are computationally more efficient than full covariance Gaussians, and with a sufficiently large number of components they can represent a perhaps richer collection of different distributions than a smaller number of full-covariance components can with a similar number of total parameters.
- NN-2-100: Neural network (multi-layered perceptron) with 2 layers. The number of units in the input and output layer is set to the number of features and the number of classes, respectively. The number of units in the hidden layer is set to 100. We use standard Levenberg-Marquardt backpropagation for training, a hyperbolic tangent sigmoid transfer function for the units at the hidden layer, and a linear transfer function at the output layer.
- SVM-1-0.1: The support vector machine with the radial basis function (RBF) kernel uses two parameters C^* and σ, where C^* is the penalty parameter for the errors of the non-separable case and σ is the parameter for the RBF kernel. We set the values for these parameters to $C^* = 1$ and $\sigma = 0.1$.

The optimal choice of the parameters, kernel function, number of neurons in the hidden layer, and transfer functions of the above mentioned classifiers was optimized in each case by performing extensive experiments. The numbers given above were found to be the best for each classifier. In contrast, for the Bayesian network classifiers we have to select the model family (e.g. TAN). We also note that all these classifiers are applied exclusively on continuous features (which gives them a distinct advantage).

The structure of Bayesian networks is implicitly regularized when we fix the optimization a-priori over a given model family (e.g. 1-trees) assuming sufficient training data. We noticed for 2-trees that the data will over-fit without the use of regularization. Therefore, we introduce 5-fold cross validation on the *training* data to find the optimal classifier structure. A similar validation procedure has been also used for the training of the NN.

Table 3: Classification accuracy in [%] for 4 and 6 classes with standard deviation using the joint TSF and MFCC feature set. The bottom line gives the average performance for each classification approach.

Classifier		NB	TAN	TAN	2-tree	TAN	NB-Cont	GMM-500	NN-2-100	SVM-1-0.1
Struct. Learn.		-	CMI	OMI-CR	OMI-CR	CR	-	-	-	-
Param. Learn.		ML	ML	ML	ML	ML	-	-	-	-
Data set	Class									
Ma+Fe	4	87.13	90.37	90.79	91.35	91.10	88.89	90.12	92.58	92.78
		± 0.07	± 0.06	± 0.06	± 0.06	± 0.06	± 0.07	± 0.06	± 0.06	± 0.06
Ma	4	87.17	90.21	90.95	91.44	91.19	88.81	90.30	92.73	92.69
		± 0.09	± 0.08	± 0.08	± 0.07	± 0.08	± 0.08	± 0.08	± 0.07	± 0.07
Fe	4	87.50	90.27	90.79	91.47	91.11	89.17	90.76	92.91	92.97
		± 0.12	± 0.11	± 0.11	± 0.10	± 0.11	± 0.12	± 0.11	± 0.10	± 0.10
Ma+Fe	6	80.58	82.68	83.21	83.61	83.52	80.11	82.30	86.05	86.26
		± 0.08	± 0.08	± 0.08	± 0.08	± 0.08	± 0.08	± 0.08	± 0.07	± 0.07
Ma	6	80.43	81.60	83.29	83.64	83.30	80.13	82.08	86.04	86.16
		± 0.10	± 0.10	± 0.09	± 0.09	± 0.09	± 0.10	± 0.10	± 0.09	± 0.09
Fe	6	81.16	82.23	83.33	84.02	83.40	80.47	82.95	86.37	86.65
		± 0.14	± 0.14	± 0.13	± 0.13	± 0.13	± 0.14	± 0.13	± 0.12	± 0.12
Average		83.99	86.23	87.06	87.59	87.27	84.60	86.42	89.45	89.59

Table 4: Comparison of different classifiers using the one-sided paired t-test: Each entry of the table gives the significance of the difference of the classification accuracy of two classifiers over the data sets. The arrow points to the superior learning algorithm. We use a double arrow if the difference is significant at the level of 0.05.

Classifier	TAN	TAN	2-tree	TAN	NB-Cont	GMM-500	NN-2-100	SVM-1-0.1
Struct. Learn.	CMI	OMI-CR	OMI-CR	CR	-	-	-	-
Param. Learn.	ML	ML	ML	ML	-	-	-	-
NB-ML	⇑0.00041	⇑0.00001	⇑<0.00001	⇑<0.00001	↑0.08534	⇑0.00009	⇑<0.00001	⇑<0.00001
TAN-CMI-ML		⇑0.00203	⇑0.00011	⇑0.00013	⇐0.00008	↑0.10376	⇑0.00005	⇑0.00005
TAN-OMI-CR-ML			⇑0.00004	⇑0.00369	⇐0.00003	⇐0.00318	⇑0.00002	⇑0.00002
2-TREE-OMI-CR-ML				⇐0.00167	oogle⇐0.00001	⇐0.00002	⇑0.00009	⇑0.00012
TAN-CR-ML					⇐0.00001	⇐0.00052	⇑0.00006	⇑0.00008
NB-CONT						⇑0.00003	⇑0.00002	⇑0.00002
GMM-500							⇑0.00003	⇑0.00003
NN-2-100								⇑0.01007

The combination of both feature sets (i.e. TSF and MFCC) improves the absolute classification accuracy by $\sim$ 0.5% on average (for comparison see Table 1). The NB classifier on continuous features is slightly better than NB on the discretized feature space. The discriminative 2-tree Bayesian network classifier significantly outperforms all other Bayesian network classifiers and GMM-500. Whereas, also the discriminatively trained TAN structures (i.e. TAN-OMI-CR and TAN-CR) perform better that the generative GMM-500. However, the best classification performance is achieved with NNs and SVMs that use continuous features. In contrast to NN and SVM, however, a Bayesian network even when discriminatively structured is a generative model which can be easily applied to classification tasks with missing features.

The classification results of Table 3 for 4 classes are summarized graphically in Figure 6 and for 6 classes in Figure 7.

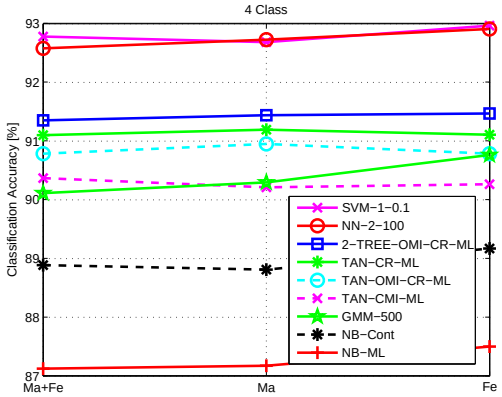


Fig. 6. Classification accuracy over the Ma+Fe, Ma, and Fe data sets for the 4 class data.

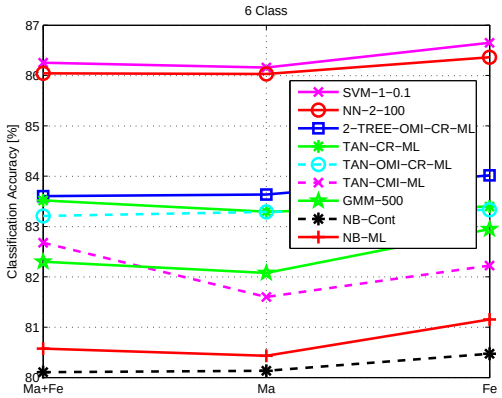


Fig. 7. Classification accuracy over the Ma+Fe, Ma, and Fe data sets for the 6 class data.

Table 4 presents a summary of the classification results over all experiments from Table 3. All pairs of classifiers are compared using the one-sided paired t-test (Mitchell, 1997). Each entry in this table depicts the significance of the difference in classification accuracy of two classification methods. The arrow points to the better classifier. If the arrow is doubled the difference is significant at a level of 0.05. Minghu Generative models can easily deal with missing features simply by marginalizing out from the model the missing feature. We are particularly interested in a testing context which has known, unanticipated at training time, and arbitrary sets of missing features for each classification sample. In such case, it is not possible to re-train the model for each potential set of missing features without also memorizing the training set. Due to the local-normalization property of Bayesian networks and the structure of any model with a parentless class node, marginalization is as easy as an $O(r^{k+1})$ operations on a k-tree , where r is the domain size of each feature.

In Figure 8, we present the classification accuracy of discriminative and generative structures assuming missing features using the Ma+Fe data for 4 and 6 phonetic classes. The x-axis denotes the number of missing features in each frame. The curves are the average over 100 classifications of the test data with uniformly at random selected missing features. Variance bars are omitted to improve readability. We note, however, that the variance numbers over the different test cases do indicate that the resulting differences are significant. We use exactly the same missing features for each classifier. We observe that discriminatively structured Bayesian network classifiers outperform TAN-CMI-ML even in the case of missing features. This demonstrates, at least empirically, that discriminative structured generative models do not loose their ability to impute missing features.

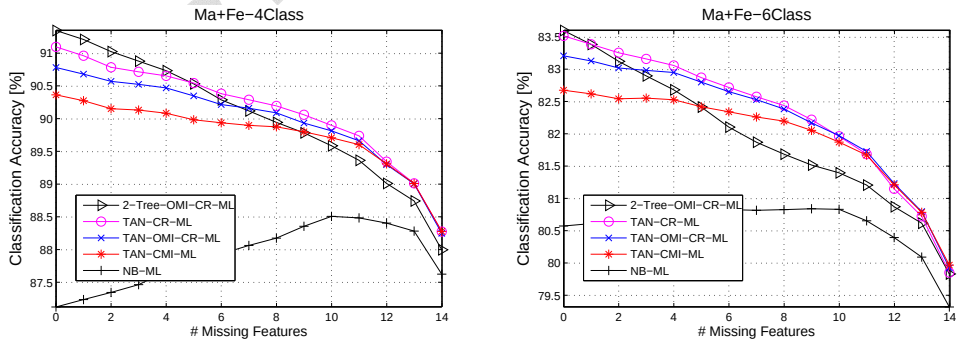


Fig. 8. Classification accuracy assuming missing features using the Ma+Fe data. The x-axis denotes the number of missing features.

Four our last set of results, we empirically show that the chosen approach (i.e. OMI) for ordering the variables improve the classification performance compared to simple random orderings. We compare 2-tree-OMI-CR to 2-tree-RO-CR using TSF+MFCC features in Table 5. We use 100 random orderings

Table 5
Classification accuracy in [%] with 2-tree-RO-CR compared to 2-tree-OMI-CR for 4 and 6 classes. For Max (Min), we take the structure which achieves the maximum (minimum) CR over the 100 random orderings on the training set and report the performance on the test set. Best results use bold font.

		Classifier	2-tree-OMI-CR	2-tree-RO-CR			
Data set	Features	#Class		Mean ± Std	Median	Min	Max
Ma+Fe	TSF+MFCC	4	**91.35**	91.26 ± 0.10	91.27	91.18	91.28
Ma	TSF+MFCC	4	**91.44**	91.31 ± 0.09	91.32	91.39	91.38
Fe	TSF+MFCC	4	**91.47**	91.43 ± 0.07	91.43	91.43	91.28
Ma+Fe	TSF+MFCC	6	**83.61**	83.49 ± 0.15	83.51	83.37	83.01
Ma	TSF+MFCC	6	**83.64**	83.49 ± 0.13	83.50	83.53	83.49
Fe	TSF+MFCC	6	**84.02**	83.79 ± 0.12	83.81	83.85	83.28
Average			**87.59**	87.46	87.47	87.46	87.29

for 2-tree-RO-CR and report the mean (Mean), minimum (Min), and maximum (Max) classification accuracy. For Max (Min), we take the structure which achieves the maximum (minimum) CR over the 100 random orderings on the training set and present the performance on the test set. In some cases, the structure with the best CR on the training set performs poorly on the test set, presumably due to overfitting. These results show that our OMI-CR heuristic improves over random orders.

Finally, the running time of the TAN-CMI, TAN-OMI-CR, and TAN-CR structure learning algorithms is summarized in Table 6. The numbers represent the percentage of time that is needed for a particular algorithm compared to TAN-CR. TAN-CMI is roughly 3 times faster than TAN-OMI-CR and TAN-CR takes about 10 times longer for establishing the structure than TAN-OMI-CR.

Table 6
Running time of structure learning algorithms relative to TAN-CR.

TAN-CMI	TAN-OMI-CR	TAN-CR
3.56%	11.47%	100.00%

5 Conclusion

Bayesian networks, Gaussian mixture models, neural networks, and support vector machines are used to classify speech frames into the broad phonetic classes of silence, voiced, unvoiced, mixed sounds, and two more categories voiced closure and release of plosives. The classification is based on time-scale features derived from the discrete Wavelet transform, on MFCCs, and on the combination of both. Gender dependent/independent experiments have been performed using the TIMIT database. Discriminative and generative param-

eter and/or structure learning approaches are used for learning the Bayesian network classifiers. We introduce a simple order-based greedy heuristic for learning a discriminative Bayesian network structure. We show that the proposed metric for establishing the ordering is performing better than simple random ordering.

We observed that the time-scale features and the MFCC features complement each other. The combination of both feature sets improves the (absolute) classification accuracy by $\sim$ 0.5%. Discriminative structure learning of Bayesian networks is superior to the generative approach. In particular, the discriminative 2-tree Bayesian network classifier significantly outperforms all other Bayesian network classifiers and the Gaussian mixture model. The best classification performances are achieved with neural networks and support vector machines. However, in contrast to neural network and support vector machines, a Bayesian network is a generative model. A generative model has the advantage that it is easy to work with missing features, and generative Bayesian network can still be trained and structured discriminatively without loosing its generative capability. We show that discriminatively structured Bayesian network classifiers are superior to generative approaches even in the case of missing features.

Future work will focus on the application of the broad phonetic classifier for speech modification such as time-scaling. Based on the phonetic information of every speech frame, the proper time-scaling factors are assigned to achieve a better quality and naturalness of scaled speech sound. Additionally, we intend to investigate the influence of the broad phonetic classification to the selection of proper smoothing strategies at concatenation points for preparing databases for concatenative synthesis.

Acknowledgments

We would like to acknowledge support for this project from the Austrian Science Fund (Project number P19737-N15). This work was also supported by an ONR MURI grant, No. N000140510388. This research was carried out in the context of COAST (http://www.coast.at). We gratefully acknowledge funding by the Austrian KNet Program, ZID Zentrum fuer Innovation und Technology, Vienna, the Steirische WirtschaftsfoerderungsGmbh, and the Land Steiermark.

A Appendix: OMI-CR algorithm

OMI-CR for learning a discriminative TAN structure is summarized in Algorithm 1. We merge both steps, establish an ordering and parent selection, into one loop. This is equivalent to considering both steps separately.

Algorithm 1 OMI-CR

Input: $\mathbf{X}_{1:N}, C, \mathcal{S}$
Output: set of edges $\mathbf{E}$ for TAN network
$X^1_\prec \leftarrow \arg\max_{X \in \mathbf{X}_{1:N}} \left[I\left(C; X\right) \right]$
$X^2_\prec \leftarrow \arg\max_{X \in \mathbf{X}_{1:N} \setminus X^1_\prec} \left[I\left(C; X | X^1_\prec\right) \right]$
$\mathbf{E} \leftarrow \left\{ \mathbf{E}_{\text{Naive Bayes}} \cup E_{X^1_\prec, X^2_\prec} \right\}$
$j \leftarrow 2$
$CR_{old} \leftarrow 0$
repeat
  $j \leftarrow j + 1$
  $X^j_\prec \leftarrow \arg\max_{X \in \mathbf{X}_{1:N} \setminus \mathbf{X}^{1:j-1}_\prec} \left[I\left(C; X | \mathbf{X}^{1:j-1}_\prec\right) \right]$
  $X^*_\prec \leftarrow \arg\max_{X \in \mathbf{X}^{1:j-1}_\prec} CR\left(\mathcal{B}_\mathcal{S} | \mathcal{S}\right)$ where
    edges of $\mathcal{B}_\mathcal{S}$ are $\mathbf{E} \leftarrow \left\{ \mathbf{E} \cup E_{X, X^j_\prec} \right\}$
  $CR_{new} \leftarrow CR\left(\mathcal{B}_\mathcal{S} | \mathcal{S}\right)$ where
    edges of $\mathcal{B}_\mathcal{S}$ are $\mathbf{E} \leftarrow \left\{ \mathbf{E} \cup E_{X^*_\prec, X^j_\prec} \right\}$
  if $CR_{new} > CR_{old}$ **then**
    $CR_{old} \leftarrow CR_{new}$
    $\mathbf{E} \leftarrow \left\{ \mathbf{E} \cup E_{X^*_\prec, X^j_\prec} \right\}$
  end if
until $j = N$

B Appendix: Abbreviations

CL	Conditional likelihood
CR	Classification rate
CMI	Conditional mutual information
DWT	Discrete Wavelet transform
Fe	Female speakers
GMM	Gaussian mixture model
$LgSE$	Logarithmic short-term energy (feature)
M	Mixed-excitation (class)
Ma	Male speakers
MFCC	Mel-frequency cepstral coefficient
ML	Maximum likelihood
NB	Naive Bayes
NN	Neuronal network
OMI	Order mutual information
PA	Power of approximation (feature)
PD	Power delta (feature)
PeD	Peak delta (feature)
PR_1	First power ratio (feature)
PR_2	Second power ratio (feature)
R	Release of plosive (class)
RBF	Radial basis function
RO	Random ordering
S	Silence (class)
SD	Standard deviation (feature)
SVM	Support vector machine
TAN	Tree augmented naive Bayes
TSF	Time-scale features
U	Unvoiced (class)
V	Voiced (class)
VC	Voiced closure (class)
ZCR	Zero crossing rate (feature)

References

Acid, S., de Campos, L., Castellano, J., 2005. Learning Bayesian network classifiers: Searching in a space of partially directed acyclic graphs. Machine Learning 59, 213–235.

Atal, B., Rabiner, L. R., 1976. A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition. IEEE Transactions on Acoustic, Speech and Signal Processing 24 (3), 201–212.

Bahl, L., Brown, P., de Souza, P., Mercer, R., 1986. Maximum Mutual Information estimation of HMM parameters for speech recognition. In: 11^{th} IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). pp. 49–52.

Bartels, C., Bilmes, J., 2007. Use of syllable nuclei locations to improve ASR.

In: IEEE Automatic Speech Recognition and Understanding (ASRU). pp. 335–340.

Bartlett, P., Jordan, M., J.D., M., 2006. Convexity, classification, and risk bounds. Journal of the American Statistical Association 101 (473), 138–156.

Bilmes, J., Zweig, G., Richardson, T., Filali, K., Livescu, K., Xu, P., Jackson, K., Brandman, Y., Sandness, E., Holtz, E., Torres, J., Byrne, B., 2001. Discriminatively structured graphical models for speech recognition: JHU-WS-2001 final workshop report. Tech. rep., CLSP, Johns Hopkins University, Baltimore MD.

Bishop, C., Lasserre, J., 2007. Generative or discriminative? Getting the best of both worlds. Bayesian Statistics 3, 3–24.

Bishop, C., 1995. Neural networks for pattern recognition. Oxford University Press.

Borys, S., Hasegawa-Johnson, M., 2005. Distinctive feature based SVM discriminant features for improvements to phone recognition on telephone band speech. In: 9^{th} European Conference on Speech Communication and Technology (Interspeech). pp. 697–700.

Bourlard, H., Morgan, N., 1994. Connectionist Speech Recognition: A Hybrid Approach. Kluwer Academic Publishers.

Buntine, W., 1991. Theory refinement on Bayesian networks. In: 7^{th} International Conference of Uncertainty in Artificial Intelligence (UAI). pp. 52–60.

Burges, C., 1998. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery 2 (2), 121–167.

Campbell, W., Isard, S., 1991. Segment durations in a syllable frame. Journal of Phonetics 19, 37–47.

Childers, D. G., Hahn, M., Larar, J. N., 1989. Silence and voiced/unvoiced/mixed excitation classification of speech. IEEE Transactions on Acoustic, Speech and Signal Processing 37 (11), 1771–1774.

Chow, C., Liu, C., 1968. Approximating discrete probability distributions with dependence trees. IEEE Transaction on Information Theory 14, 462–467.

Cooke, M., Green, P., Josifovski, L., Vizinho, A., 2001. Robust automatic speech recognition with missing and unrealiable acoustic data. Speech Communication 34, 267–285.

Cooper, G., Herskovits, E., 1992. A Bayesian method for the induction of probabilistic networks from data. Machine Learning 9, 309–347.

Cover, T., Thomas, J., 1991. Elements of information theory. John Wiley & Sons.

Cowell, R., Dawid, A., Lauritzen, S., Spiegelhalter, D., 1999. Probabilistic networks and expert systems. Springer Verlag.

de Campos, L., 2006. A scoring function for learning Bayesian networks based on mutual information and conditional independence tests. Journal of Machine Learning Research 7, 2149–2187.

Donnellan, O., Jung, E., Coyle, E., 2003. Speech-adaptive time-scale modification for computer assisted language-learning. In: 3^{rd} International Con-

ference on Advanced Learning Technologies. pp. 165–169.

Dougherty, J., Kohavi, R., Sahami, M., 1995. Supervised and unsupervised discretization of continuous features. In: 12^{th} International Conference on Machine Learning. pp. 194–202.

Duda, R., Hart, P., Stork, D., 2001. Pattern Classification. John Wiley & Sons.

Ephraim, Y., Dembo, A., Rabiner, L., 1989. A minimum discrimination information approach for Hidden Markov Models. IEEE Transactions on Information Theory 35 (5), 1001–1013.

Ephraim, Y., Rabiner, L., 1990. On the releations between modeling approaches for speech recognition. IEEE Transactions on Information Theory 36 (2), 372–380.

Fayyad, U., Irani, K., 1993. Multi-interval discretizaton of continuous-valued attributes for classification learning. In: 13^{th} International Joint Conference on Artificial Intelligence. pp. 1022–1027.

Fei, S., Saul, L., 2006. Large margin Gaussian mixture modeling for phonetic classification and recognition. In: 31^{st} IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). pp. 265 – 268.

Friedman, N., Geiger, D., Goldszmidt, M., 1997. Bayesian network classifiers. Machine Learning 29, 131–163.

Friedman, N., Nachman, I., Peer, D., 1999. Learning Bayesian network structure form massive datasets: The sparse candidate algorithm. In: 15^{th} International Conference of Uncertainty in Artificial Intelligence (UAI). pp. 196–205.

Greiner, R., Su, X., Shen, S., Zhou, W., 2005. Structural extension to logistic regression: Discriminative parameter learning of belief net classifiers. Machine Learning 59, 297–322.

Greiner, R., Zhou, W., 2002. Structural extension to logistic regression: Discriminative parameter learning of belief net classifiers. In: 18th Conference of the AAAI. pp. 167–173.

Grossman, D., Domingos, P., 2004. Learning bayesian network classifiers by maximizing conditional likelihood. In: 21^{st} International Conference on Machine Learning (ICML). pp. 361–368.

Halberstadt, A., Glass, J., 1997. Heterogeneous acoustic measurements for phonetic classification. In: 5^{th} European Conference on Speech Communication and Technology (Eurospeech). pp. 401 – 404.

Heckerman, D., 1995. A tutorial on learning Bayesian networks. Tech. Rep. MSR-TR-95-06, Microsoft Research.

Jebara, T., 2001. Discriminative, generative and imitative learning. Ph.D. thesis, Media Laboratory, MIT.

Jordan, M., 1999. Learning in graphical models. MIT Press.

Juang, B.-H., Chou, W., Lee, C.-H., 1997. Minimum classification error rate methods for speech recognition. IEEE Transactions on Speech and Ausio Processing 5 (3), 257–265.

Juang, B.-H., Katagiri, S., 1992. Discriminative learning for minimum error classification. IEEE Transactions on Signal Processing 40 (12), 3043–3054.

Kedem, B., 1986. Spectral analysis and discrimination by zero-crossings. Proceedings of the IEEE 74, 1477–1493.

Keogh, E., Pazzani, M., 1999. Learning augmented Bayesian classifiers: A comparison of distribution-based and classification-based approaches. In: 7^{th} International Workshop on Artificial Intelligence and Statistics. pp. 225–230.

Kirchhoff, K., Fing, G., Sagerer, G., 2002. Combining acoustic and articulatory feature information for robust speech recognition. Speech Communication 37, 303–319.

Kruskal, J., 1956. On the shortest spanning subtree and the traveling salesman problem. In: Proceedings of the American Mathematical Society. Vol. 7. pp. 48–50.

Kubin, G., Atal, B., Kleijn, W., 1993. Performance of noise excitation for unvoiced speech. In: IEEE Workshop on Speech Coding for Telecommunications. pp. 35–36.

Kubin, G., Kleijn, W., 1994. Time-scale modification of speech based on a nonlinear oscillator model. In: 19^{th} IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Vol. 1. pp. 453–456.

Kuwabara, H., Nakamura, M., 2000. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, Ch. Acoustic and Perceptual Properties of Syllables in Continuous Speech as a Function of Speaking Rate, pp. 163–245.

Lamel, L., Kassel, R., Seneff, S., 1986. Speech database development: Design and analysis of the acoustic-phonetic corpus. In: Proceedings of the DARPA Speech Recognition Workshop, Report No. SAIC-86/1546.

Leung, H., Chigier, B., Glass, J., 1993. A comparative study of signal representations and classification techniques for speech recognition. In: 18^{th} IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). pp. 657 – 664.

Levinson, S., Liberman, M., Ljolje, A., Miller, L., 1989. Speaker independent phonetic transcription of fluent speech for large vocabulary speech recognition. In: Human Language Technology Conference. pp. 75 – 80.

Lin, H., Bilmes, J., Vergyri, D., Kirchhoff, K., 2007. OOV detection by joint word/phone lattice alignment. In: IEEE Automatic Speech Recognition and Understanding (ASRU). pp. 478–483.

Malkin, J., Bilmes, 2008. Ratio semi-definite classifiers. In: 33^{rd} IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). pp. 4113–4116.

Minghu, J., Baozong, Y., Biquin, L., 1996. The consonant/vowel (C/V) speech classification using high-rank function neural network (HRFNN). In: 3^{rd} Interantional Conference on Signal Processing. Vol. 2. pp. 1469–1472.

Mitchell, T., 1997. Machine Learning. McGraw Hill.

Murphy, K., 2002. Dynamic Bayesian networks: Representation, inference and learning. PhD Thesis, University of California, Berkeley.

Olive, J. P., Greenwood, A., Coleman, J., 1993. Acoustic of American English.

Springer.

Parveen, S., Green, P., 2004. Speech enhancement with missing data techniques using recurrent neural networks. In: 29^{th} IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). pp. 733–736.

Pazzani, M., 1996. Searching for dependencies in Bayesian classifiers. In: Learning from data: Artificial intelligence and statistics V. pp. 239–248.

Pearl, J., 1988. Probabilistic reasoning in intelligent systems: Networks of plausible inference. Morgan Kaufmann.

Pernkopf, F., Bilmes, J., 2005. Discriminative versus generative parameter and structure learning of Bayesian network classifiers. In: 22^{nd} International Conference on Machine Learning (ICML). pp. 657 – 664.

Pernkopf, F., Bilmes, J., 2008. Ordering-based discriminative structure learning for Bayesian network classifiers. In: 10^{th} International Symposium on Artificial Intelligence and Mathematics. p. accepted.

Pernkopf, F., 2005. Bayesian network classifiers versus selective k-NN classifier. Pattern Recognition 38 (3), 1–10.

Pham, T. V., Kubin, G., 2005. DWT-based phonetic groups classification using neural network. In: 30^{th} IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). pp. 401–404.

Rabiner, L., 1989. A tutorial on Hidden Markov Models and selected applications in speech recognition. Proceedings of the IEEE 77 (2), 257–286.

Raj, B., Stern, R., 2005. Missing-feature approaches in speech recognition. IEEE Signal Processing Magazine 22 (5), 101–116.

Roos, T., Wettig, H., Grünwald, P., Myllymäki, P., Tirri, H., 2005. On discriminative Bayesian network classifiers and logistic regression. Machine Learning 59, 267–296.

Salomon, J., King, S., Osborne, M., 2002. Framewise phone classification using support vector machines. In: International Conference on Spoken Language Processing. pp. 2645 – 2648.

Sanneck, H., 1998. Concealment of lost speech packets using adaptive packetization. In: IEEE International Conference on Multimedia Computing and Systems. pp. 140–149.

Schölkopf, B., Smola, A., 2001. Learning with kernels: Support Vector Machines, regularization, optimization, and beyond. MIT Press.

Smith, N., Gales, M., 2002. Using SVMs and discriminative models for speech recognition. In: 27^{th} IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). pp. 77–80.

Subramanya, A., Bilmes, J., Chen, C.-P., 2005. Focused word segmentation for ASR. In: 9^{th} European Conference on Speech Communication and Technology (Interspeech). pp. 393–396.

Teyssier, M., Koller, D., 2005. Ordering-based search: A simple and effective algorithm for learning Bayesian networks. In: 21^{th} International Conference of Uncertainty in Artificial Intelligence (UAI). pp. 584 – 590.

Vetterli, M., Kovacevic, J., 1995. Wavelets and subband coding.

Wettig, H., Grünwald, P., Roos, T., Myllymäki, P., Tirri, H., 2003. When

discriminative learning of Bayesian network parameters is easy. In: International Joint Conference on Artificial Intelligence (IJCAI). pp. 491 – 496.

Zhang, L., Wang, T., Cuperman, V., 1997. A CELP variable rate speech codec with low average rate. In: 22^{nd} IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Vol. 2. pp. 735–738.