

Partial-tied-mixture Auxiliary Chain Models for Speech Recognition Based on Dynamic Bayesian Networks

Hui Lin and Zhijian Ou, *Member, IEEE*

Abstract—It is observed that the cepstral-based features used for speech recognition are sensitive to some auxiliary information (e.g. pitch). Encoding the auxiliary information in discrete auxiliary variables based on dynamic Bayesian networks (DBNs) typically results in an increased number of parameters. There are tradeoffs to be studied between parameter reduction and dependency modeling. In this paper, we propose a method using state-specific partial tying with information-theoretic dependency selection. This method is essentially to relax the conditional independence assumptions imposed by the full-tied-mixture model, by adding strong dependencies (i.e. those with large mutual information computed from training data). Experiments were carried out on the OGI Numbers database, considering pitch as the auxiliary information. The results show that the partial-tied-mixture auxiliary chain models can efficiently improve recognition performances with an economical way of increasing parameters.

I. INTRODUCTION

In state-of-the-art automatic speech recognition (ASR) systems, the cepstral-based features (e.g. MFCCs) are used as the standard acoustic features to discriminate between different phonetic states. However, it is observed that these standard features are sensitive to some auxiliary information such as pitch, rate-of-speech (ROS), gender and etc. Various methods have been proposed to incorporate such auxiliary information to improve ASR robustness. Bayesian networks [1][2], in particular, dynamic Bayesian networks (DBN) [3], in which HMMs can be considered as one small instance, has been used for these studies [4][5][6][7][8].

One method is to encode the auxiliary information in continuous observed variables. It is shown in [6][9] that simply appending the auxiliary feature to the standard feature vector degrades the recognition performance. It is beneficial to use the auxiliary feature serving as a conditional variable to model the distribution of the standard acoustic feature. To have tractable exact inference in using hidden continuous variables, only the dependencies within a given time frame is considered [6].

On the other hand, the auxiliary information can also be incorporated in the form of discrete variables [4][5][7][8], which can be temporally linked to account for contextual information. The works in [4][5] show the advantage to include a discrete context variable, which forms an auxiliary chain along time. The context variable is always hidden during both training and recognition, and therefore it is not clear what auxiliary information it may represent. In [7],

pitch information is explicitly related to a discrete variable by quantization, and it is found that the performances is degraded when having the auxiliary variable observed during decoding. In [8], ROS information is used by introducing an additional discrete mode variable.

An important issue in incorporating discrete auxiliary variables is that for each value q of the phonetic state variable Q_t , the new conditional probabilistic distribution (CPD) of the acoustic feature O_t , $p(O_t|Q_t = q, A_t)$, requires a separate distribution for each value a of the auxiliary variable A_t . Each individual distribution $p(O_t|Q_t = q, A_t = a)$ is usually implemented as a Gaussian mixture model (GMM). This typically results in an increased number of parameters. To reduce the number of parameters for robust parameter estimation, an approach often used by ASR systems is parameter tying, where certain parameters are shared among a number of different models. This idea is used in [5][8] for modeling with auxiliary information, where for each phonetic state q , the Gaussian components of the GMMs for different values of A_t are tied. Only the mixture weights are different. While the number of parameters is greatly reduced, such parameter tying implies an overly constrained modeling of the influence of the auxiliary variable A_t on the acoustic feature O_t .

Usually, using a constrained implementation of the CPDs will lead to a sparse model structure, which represents fewer dependencies. An ideal parameter reduction scheme should be able to reduce the number of parameters to a number that can be robustly estimated, whilst retaining sufficient ability to model the necessary dependencies in the data. There are successful attempts, where the model structure (and the number of parameters) is adjusted by adding dependencies in an order of ranked mutual information [10][11]. These procedures can be viewed as structure learning of Bayesian networks [12].

In this paper, in order to better balance between parameter reduction and dependency modeling, we propose to use state-specific partial tying¹ with information-theoretic dependency selection. Specifically, we implement a partial-tied-mixture auxiliary chain model based on DBNs for exploiting pitch information. The quantized pitch variable A_t for each time frame are temporally linked, forming an auxiliary chain. For each phonetic state, the Gaussian components of the GMMs for different values of A_t are partial-tied.

We start from a full-tied-mixture auxiliary chain model. For each phonetic state, the mutual information between O_t

This work was supported by NSFC (No. 60402029)

The authors are with the Dept. of Electronic Engineering, Tsinghua Univ., Beijing 100084, China. (E-mail: linhui99@mails.tsinghua.edu.cn, ozj@tsinghua.edu.cn).

¹Note that the tied-mixture model discussed in our paper occurs only within each state, and is different from the tied-mixture/semi-continuous HMM [13] [14] which employs parameter tying across the phonetic states.

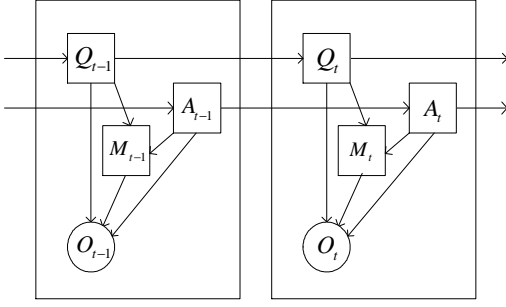


Fig. 1. Basic auxiliary chain model. Round nodes represent continuous variables, while square nodes represent discrete variables.

and A_t conditional upon each individual mixture component is computed using training data. Accordingly, a portion of mixture components is selected to be un-tied and becomes uniquely associated with each value of A_t . This procedure is essentially to relax the conditional independence assumptions imposed by the full-tied-mixture model, by adding strong dependencies (i.e. those with large mutual information computed from training data). Experiments were carried out on the OGI Numbers database [15], which is an English telephone speech corpus consisting of continuously spoken numbers. The results show that the partial-tied-mixture auxiliary chain models can efficiently improve recognition performances with an economical way of increasing parameters.

The paper is organized as follows. In section II, we begin by describing the basic auxiliary chain model, in which no parameters are tied. Then after discussing the full-tied-mixture auxiliary chain model, the partial-tied-mixture auxiliary chain model is introduced, including the information-theoretic dependency selection procedure to achieve state-specific partial tying. Section III presents experimental results, followed by conclusions in the last section.

II. AUXILIARY CHAIN MODEL FORMULATION BASED ON DBNS

Dynamic Bayesian networks (DBNs) are a flexible framework for modeling sequential data, equipped with a graphical way of model representation, and a set of general algorithms for inference and learning. An advantage of using DBNs is that they can model complex probabilistic dependencies and allow novel models to be easily developed.

A. Basic Auxiliary Chain Model

Fig. 1 shows the DBN representation of the basic auxiliary chain model as in [4][5][7]. The discrete variables are used to encode auxiliary information (e.g. pitch, ROS, or conceptual context). Q_t , O_t , A_t are respectively the discrete phonetic state variable, the continuous standard feature variable and the discrete auxiliary variable at time t . Their joint probability distribution over time is

$$p(Q_{1:T}, O_{1:T}, A_{1:T}) = \prod_{t=1}^T p(Q_t | Q_{t-1}) p(O_t | Q_t, A_t) p(A_t | A_{t-1}) \quad (1)$$

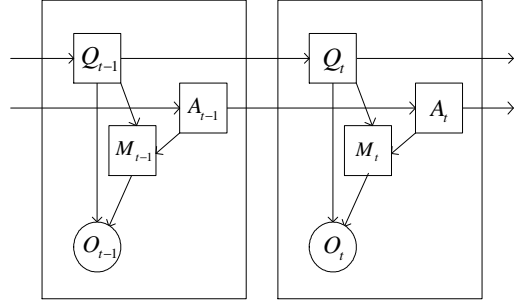


Fig. 2. Full-tied-mixture auxiliary chain model.

where $p(O_t | Q_t, A_t)$ is often implemented as a set of GMMs (i.e. one GMM for each possible combination of the values of the variables Q_t and A_t):

$$p(O_t | Q_t = q, A_t = a) = \sum_{m=1}^M p(M_t = m | Q_t = q, A_t = a) p(O_t | Q_t = q, A_t = a, M_t = m) \quad (2)$$

Here M_t denotes the hidden mixture component variable, which is explicitly shown in Fig. 1. In the general case, incorporating the auxiliary variable A_t will increase the number of Gaussian components by a factor of the cardinality of A_t . To reduce the number of parameters, the full-tied-mixture model is used in [5][8].

B. Full-tied-mixture Auxiliary Chain Model

The full-tied-mixture auxiliary chain model is shown in Fig. 2. This corresponds to deleting the directed edge from A_t to O_t in Fig. 1, and (2) is simplified to:

$$p(O_t | Q_t = q, A_t = a) = \sum_{m=1}^M p(M_t = m | Q_t = q, A_t = a) p(O_t | Q_t = q, M_t = m) \quad (3)$$

For each phonetic state q , there is a pool of Gaussians $\{p(O_t | Q_t = q, M_t = m) | m = 1, \dots, M\}$. The GMMs for different values of A_t share this pool, and differ only in their mixture weights. This implies an overly constrained modeling of the influence of the auxiliary variable A_t on the acoustic feature O_t . The underlying conditional independence assumptions are:

$$O_t \perp A_t | Q_t = q, M_t = m, \forall q, m \quad (4)$$

which means that for all possible combinations of the values of Q_t and M_t , the acoustic feature O_t is conditional independent of the auxiliary variable A_t . In the following, we propose to relax these assumptions to use state-specific partial tying, by adding strong dependencies discovered from training data.

C. Partial-tied-mixture Auxiliary Chain Model

Consider the dependency between O_t and A_t given specific phonetic state q and mixture component m . Its strength can

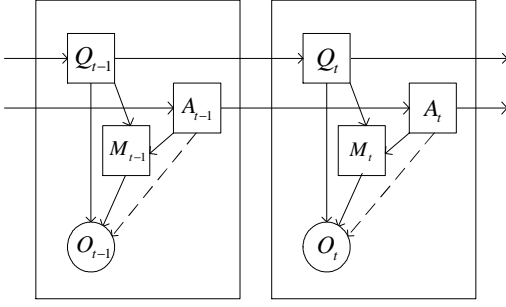


Fig. 3. Partial-tied-mixture auxiliary chain model.

be naturally measured by the (conditional) mutual information $I(O_t, A_t | Q_t = q, M_t = m)$. An equivalent condition to the conditional independence assumption (4) is

$$I(O_t, A_t | Q_t = q, M_t = m) = 0 \quad (5)$$

Intuitively, the conditional independence assumption should be relaxed (i.e. adding the dependency), if the mutual information $I(O_t, A_t | Q_t = q, M_t = m)$ is large enough (above a threshold). The set of strong dependencies that are to be added can therefore be defined for each state q by selecting the mixture components:

$$D_q = \{m | I(O_t, A_t | Q_t = q, M_t = m) > \theta\} \quad (6)$$

After adding these strong dependencies, (3) becomes:

$$\begin{aligned} & p(O_t | Q_t = q, A_t = a) \\ &= \sum_{m \notin D_q} p(M_t = m | Q_t = q, A_t = a) p(O_t | Q_t = q, M_t = m) \\ &+ \sum_{m \in D_q} p(M_t = m | Q_t = q, A_t = a) p(O_t | Q_t = q, A_t = a, M_t = m) \end{aligned} \quad (7)$$

where the set $\{m \in D_q\}$ represents un-tied mixture components, and the set $\{m \notin D_q\}$ contains the mixture components that remain tied for different values of A_t . The resulting partial-tied-mixture model essentially become a Bayesian multinet [16], where the variable A_t is switched on and off to be the parent of O_t , according to different instantiations of the other parents of O_t (i.e. Q_t and M_t). The DBN representation of the partial-tied-mixture model is shown in Fig. 3, where the dashed edge from A_t to O_t represents such context-sensitive independence [17].

The basic auxiliary chain model corresponds to the case of $D_q = \{1, \dots, M\}$, where there is a large number of parameters with the problem for reliable parameter estimation. The full-tied-mixture auxiliary chain model corresponds to the case of $D_q = \emptyset$, where, as discussed above, the modeling power is limited. By selectively adding strong dependencies, the resulting partial-tied-mixture model can potentially lead to a better balance between parameter reduction and dependency modeling.

D. Information-theoretic Dependency Selection

What remains is to obtain the conditional mutual information $I(O_t, A_t | Q_t = q, M_t = m)$, which can be computed as follows:

$$\begin{aligned} & I(O_t; A_t | Q_t = q, M_t = m) \\ &= H(O_t | Q_t = q, M_t = m) - H(O_t | Q_t = q, M_t = m, A_t) \\ &= H(O_t | Q_t = q, M_t = m) \\ &- \sum_a p(A_t = a | Q_t = q, M_t = m) H(O_t | Q_t = q, M_t = m, A_t = a) \end{aligned} \quad (8)$$

As described in the above formulation of the partial-tied-mixture auxiliary model, we start from a full-tied-mixture model, and additional dependencies are added in the order of decreasing strength until below a threshold. Therefore, we first train a full-tied-mixture model, containing relatively few but reliably estimated parameters.

Using this model, the data associated with specific $Q_t = q$ and $M_t = m$ are obtained via hard alignment. Then, for each state q and mixture component m , $p(O_t | Q_t = q, M_t = m)$ is estimated from the aligned data as a K -dimensional diagonal Gaussian density with the variances $(\sigma_{1,(q,m)}^2, \sigma_{2,(q,m)}^2, \dots, \sigma_{K,(q,m)}^2)$ as follows:

$$\sigma_{j,(q,m)}^2 = \frac{\sum_{t \in \{t | Q_t = q, M_t = m\}} (O_{j,t} - \mu_{j,(q,m)})^2}{\sum_{t \in \{t | Q_t = q, M_t = m\}} 1} \quad (9)$$

$$\mu_{j,(q,m)} = \frac{\sum_{t \in \{t | Q_t = q, M_t = m\}} O_{j,t}}{\sum_{t \in \{t | Q_t = q, M_t = m\}} 1} \quad (10)$$

where $O_{j,t}$ is the j th element of the observed vector O_t .

The aligned data can be further divided into several groups according to different values of A_t . For each $A_t = a$, $p(O_t | Q_t = q, M_t = m, A_t = a)$ is also estimated as a K -dimensional diagonal Gaussian density with the variances $(\sigma_{1,(q,m,a)}^2, \sigma_{2,(q,m,a)}^2, \dots, \sigma_{K,(q,m,a)}^2)$ as follows:

$$\sigma_{j,(q,m,a)}^2 = \frac{\sum_{t \in \{t | Q_t = q, M_t = m, A_t = a\}} (O_{j,t} - \mu_{j,(q,m,a)})^2}{\sum_{t \in \{t | Q_t = q, M_t = m, A_t = a\}} 1} \quad (11)$$

$$\mu_{j,(q,m,a)} = \frac{\sum_{t \in \{t | Q_t = q, M_t = m, A_t = a\}} O_{j,t}}{\sum_{t \in \{t | Q_t = q, M_t = m, A_t = a\}} 1} \quad (12)$$

Then (8) can be written as

$$\begin{aligned} & I(O_t; A_t | Q_t = q, M_t = m) \\ &= \sum_a p(A_t = a | Q_t = q, M_t = m) \sum_{j=1}^K \ln \frac{\sigma_{j,(q,m)}}{\sigma_{j,(q,m,a)}} \end{aligned} \quad (13)$$

where $p(A_t = a | Q_t = q, M_t = m)$ is estimated via

$$p(A_t = a | Q_t = q, M_t = m) = \frac{\sum_{t \in \{t | Q_t = q, M_t = m, A_t = a\}} 1}{\sum_a \sum_{t \in \{t | Q_t = q, M_t = m, A_t = a\}} 1} \quad (14)$$

For each state q and mixture component m , the conditional mutual information $I(O_t, A_t | Q_t = q, M_t = m)$ is computed

from training data as above. The set of strong dependencies D_q is then selected as in (6) and added, which sets up the partial-tied-mixture structure.

III. EXPERIMENTS

A. Experimental Setup

Experiments were carried out on the OGI Numbers database [15], which is an English telephone speech corpus consisting of naturally spoken numbers with 30-word vocabulary. We used 6049 utterances from the corpus for training and 2061 utterances for testing, as configured by MONC [18]. All utterances were framed with 25ms length and 10ms shift. From each frame, 12 mel-frequency cepstral coefficients (MFCCs) plus normalized log-energy were extracted along with their first and second derivatives, giving a feature vector of 39 dimension. Cepstral mean subtraction was then applied to the feature vector. The Graphical Model Toolkit (GMTK) [19] was utilized for DBN implementation. There were 26 monophone models, a silence model, and a short-pause model. The silence and all monophones were modeled with three emitting states each, and the short-pause had only one state which was tied to the middle state of the silence model.

B. Auxiliary Information

For current work, we mainly consider pitch (the fundamental frequency f_0) as the auxiliary information. The Entropic Signal Processing System (ESPS) [20] tool *get_f0* was used in the experiments to estimate the pitch. *get_f0* is a program to perform fundamental frequency estimation using the normalized cross correlation function and dynamic programming [21].

Two kinds of quantization were used here to obtain the discrete auxiliary information. One (referred to as *pitch_2c*) is to quantize the estimated f_0 to binary to reflect high-low pitch (low: below 140Hz including unvoiced frames). The other (referred to as *pitch_d_3c*) is to quantize the first-order derivative of the estimated pitch to trinary, which reflected steadiness, rising and falling of pitch.

During the training, the auxiliary variables were always observed to explicitly represent the auxiliary information.

C. Baseline HMM

A baseline DBN was built to emulate the standard HMM. There is an upper layer including position, transition variables as introduced in [4]. The various DBNs replace the lower layer with the different new structures from Fig. 1,2,3. Gaussian mixtures were trained for each phonetic state using GMTK, which employs a splitting process that doubles the number of Gaussian components after each split [19].

As shown in Table I, the word error rate (WER) of the baseline HMM with 16 Gaussian components per state is 9.80%. For comparison, we examined the effect of appending the estimated pitch (without quantization) to the standard feature vector, and forming a 40-dimension feature vector. The performance was degraded (from 9.80% to 11.74%), as found in [6][9].

TABLE I
WORD ERROR RATES FOR DIFFERENT MODELS

Model Type	Param.	WER(%)	
HMM	101k	9.80	
HMM(+pitch)	102k	11.74	
		O	H
Full-tied-mix. Aux. Chain (<i>pitch_2c</i>)	102k	9.34	9.34
Full-tied-mix. Aux. Chain (<i>pitch_d_3c</i>)	103k	9.35	9.26
Partial-tied-mix. Aux. Chain (<i>pitch_2c</i>)	117k	9.21	9.18
	123k	9.16	9.35
	141k	9.06	9.16
	177k	9.31	9.20
Partial-tied-mix. Aux. Chain (<i>pitch_2c,random</i>)	123k	9.57	9.57
Partial-tied-mix. Aux. Chain (<i>pitch_d_3c</i>)	114k	9.18	9.13
	131k	9.10	9.09
	158k	8.86	8.95
	227k	9.03	8.95
Partial-tied-mix. Aux. Chain (<i>pitch_d_3c,random</i>)	141k	9.21	9.16
Basic Aux. Chain (<i>pitch_2c</i>)	202k	10.14	9.63
Basic Aux. Chain (<i>pitch_d_3c</i>)	303k	9.62	9.26

D. Partial-tied-mixture Auxiliary Chain Models

First, the full-tied-mixture auxiliary chain model was trained. The initialization procedure is as follows:

- 1) Using the baseline HMM, the state-level viterbi path was obtained. Then each frame was hard aligned to the most probable mixture component.
- 2) For each state, the Gaussian components for different values of A_t were duplicated from the baseline HMM. The mixture weights were initialized from the aligned data $\{O_t, A_t, Q_t = q, M_t = m\}$.
- 3) The state transition matrix was copied from the baseline HMM, and the transition distribution for the auxiliary variable was initialized with uniform distributions.

The initialized model was then trained using several EM iterations with no splitting or vanishing [19] until the relative difference in the global log-likelihood is less than 0.1%.

By changing θ in (6) to generate different sizes of D_q , various partial-tied-mixture models with varying numbers of parameters were created and initialized from the previously trained full-tied-mixture auxiliary chain model. As a special case, the basic auxiliary chain model was initialized by setting $D_q = \{1, \dots, M\}$. In addition, we also experimented with adding a random set of dependencies, instead of those strong dependencies determined by (6). All the initialized models were then trained using several EM iterations with no splitting or vanishing until the relative difference in the log-likelihood is less than 0.1%.

E. Results

First, all types of DBN chain models were tested under the condition where the auxiliary variables were observed (O). The results are shown in the 'O' column of Table I and also plotted in Fig. 4. It can be seen that there are no efficient performance improvements with the basic auxiliary chain models. Using *pitch_d_3c* achieved 1.8% WER reduction

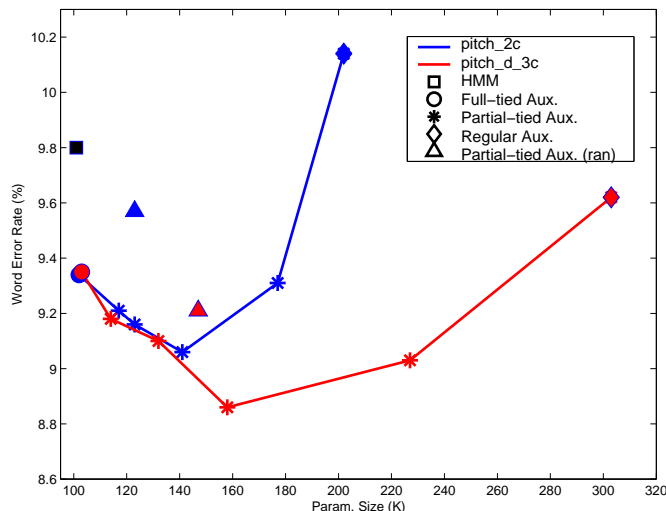


Fig. 4. Performance of different types of DBN chain models, tested under the condition where the auxiliary variable was observed (O). Blue indicates using high-low quantized pitch as the auxiliary information. Red indicates using trinary quantized first-order derivative of pitch.

from the baseline HMM, while increasing the number of parameters by a factor of 3. Using *pitch_2c* even degraded the performance. These results show that we need to reduce the number of parameters for reliable parameter estimation.

Small WER reductions (4.7% at best) from the baseline HMM were obtained using the full-tied-mixture models, with comparable numbers of parameters. Due to the overly constrained modeling of the influence of the auxiliary information on the acoustic features, the full-tied-mixture models cannot achieve further improvements.

The results show that the partial-tied-mixture auxiliary chain models are able to achieve a better balance between parameter reduction and dependency modeling. Without a large parameter increase, the WER was reduced from the baseline HMM by 9.6% when using *pitch_d_3c*. Increasing parameters by adding strong dependencies proves to be useful for building compact yet powerful statistical models. This is further confirmed by the worse results of randomly adding dependencies.

In addition, the models were also tested under the condition where the auxiliary variables were left hidden (H). The results are given in the 'H' column of Table I. In [7], it is found that during recognition, the chain models perform significantly better when using hidden pitch variables than when using observed pitch variables, presumably due to that the pitch estimates were noisy. However, the results in Table I show that for the tied-mixture models, using observed pitch variables during recognition achieved comparable performance with using hidden pitch variables. This is encouraging, since inference using hidden auxiliary variables typically requires much greater complexities (in both computation and memory) than using hidden ones.

IV. CONCLUSIONS

In this paper, in order to better balance between parameter reduction and dependency modeling for incorporating auxiliary information into state-of-the-art ASR systems, we propose to use state-specific partial tying with information-theoretic dependency selection. Various chain models based on DBNs were evaluated on the OGI Numbers database, considering pitch as the auxiliary information. The results show that the proposed partial-tied-mixture auxiliary chain models can efficiently improve recognition performances with an economical way of increasing parameters. In future, we plan to apply the proposed method to exploit more auxiliary information (e.g. ROS, the state of articulators, noise condition, etc).

REFERENCES

- [1] F. V. Jensen, *Bayesian Networks and Decision Graphs*. Springer-Verlag, 2001.
- [2] R. G. Cowell, A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter, *Probabilistic Networks and Expert Systems*. Springer-Verlag, 1999.
- [3] T. Dean and K. Kanazawa, "Probabilistic temporal reasoning," *Proc. AAAI*, pp. 524–538, 1988.
- [4] G. Zweig, "Speech recognition with dynamic bayesian networks," *Ph.D. dissertation, Univ. California, Berkeley*, 1998.
- [5] G. Zweig and M. Padmanabhan, "Dependency modeling with bayesian networks in a voicemail transcription system," *Proc. EUROSPEECH*, 1999.
- [6] T. Stephenson, M. Mathew, and H. Bourlard, "Speech recognition with auxiliary information," *IEEE Trans. on Speech and Audio Processing*, vol. 12, no. 3, 2004.
- [7] M. M. T.A. Stephenson and H. Bourlard, "Modeling auxiliary information in bayesian network based asr," *Proc. EUROSPEECH*, 2001.
- [8] T. Shinozaki and S. Furui, "Hidden mode hmm using bayesian networks for modeling speaking rate fluctuation," *Proc. Automatic Speech Recognition and Understanding*, pp. 417–422, 2003.
- [9] K. Fujinaga, M. Nakai, H. Shimodaira, and S. Sagayama, "Multiple-regression hidden markov model," *Proc. ICASSP*, 2001.
- [10] J. Bilmes, "Factored sparse inverse covariance matrices," *Proc. of ICASSP*, 2000.
- [11] J. Frankel, M. Wester, and S. King, "Articulatory feature recognition using dynamic bayesian networks," *Proc. of ICSLP*, 2004.
- [12] D. Heckerman, "A tutorial on learning with bayesian networks," *Learning in Graphical models, M. Jordan (Ed.)*, MIT press, 1999.
- [13] J. R. Bellegarda and D. Nahamoo, "Tied mixture continuous parameter modeling for speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, pp. 2033–2045, 1990.
- [14] X. D. Huang, "Phoneme classification using semicontinuous hidden markov models," *IEEE Trans. Signal Processing*, vol. 40, pp. 1062–1067, 1992.
- [15] R. Cole, M. Fanty, M. Noel, and T. Lander, "Telephone speech corpus development at cslu," *Proc. ICSLP*, 1994.
- [16] D. Geiger and D. Heckerman, "Knowledge representation and inference in similarity networks and bayesian multinets," *Artificial Intelligence*, 1996.
- [17] C. Boutilier, N. Friedman, M. Goldszmidt, and D. Koller, "Context-specific independence in bayesian networks," *Proc. UAI*, pp. 115–123, 1996.
- [18] "Multi channel overlapping numbers corpus (monc) distribution," <http://cslu.cse.ogi.edu/corpora/>.
- [19] J. Bilmes and G. Zweig, "The graphic models toolkit: An open source software system for speech and time-series processing," *Proc. ICASSP*, 2002.
- [20] "Espes with waves," *Entropic Research Laboratory, Inc. AT&T Bell Laboratories*, 1993.
- [21] B. Secrest and G. Doddington, "An integrated pitch tracking algorithm for speech systems," *Proc. ICASSP*, 1983.