In: **International Conference on Pattern Recognition (ICPR)**, HongKong, 2006,8.

Switching Auxiliary Chains for Speech Recognition based on Dynamic Bayesian Networks[1]

Hui Lin, Zhijian Ou

Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
linhui99@mails.tsinghua.edu.cn, ozj@tsinghua.edu.cn

Abstract

This paper investigates the problem of incorporating auxiliary information (e.g. pitch) for speech recognition using dynamic Bayesian networks (DBNs). Previous works usually model acoustic features conditional on the pitch auxiliary variable for both voiced and unvoiced phonetic states, and therefore ignore the fact that pitch (frequency) information is meaningful only for voiced states.. In this paper we propose a switching two auxiliary chain model tailored to voiced/unvoiced states for exploiting pitch information, which is essentially built on the switching parent functionality of Bayesian multinets. Experiments on the OGI Numbers database show that significant performance improvements are achieved from switching auxiliary chain modeling, compared with regular auxiliary chain modeling and the standard HMM.

1. Introduction

For automatic speech recognition, HMMs consist of two sets of random variables, the hidden phonetic state variable and the acoustic feature variable at each time. One important deficiency is that the single phonetic state variable is burdened to contain all relevant contextual information. There are clearly some contextual cues that are not explicitly represented by the phonetic states (e.g. pitch, rate of speech, the state of articulators, noise condition, etc.), which we could call auxiliary information.

Various methods have been proposed to incorporate auxiliary information to increase the representational capacity of the standard HMM. One method is to encode the auxiliary information in continuous observable variables (for both training and recognition) [1][2]. To have tractable (exact) inference in using hidden continuous variables, this method only considers dependencies within a given time frame as done in [1][2].

On the other hand, the auxiliary information could also be incorporated in the form of discrete variables [3][4][5], which can be temporally linked and directly complement the phonetic state variables to model long-term acoustic context. The works in [3][4] 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, thus it is not clear what auxiliary information it may represent. In [5], pitch information is explicitly related to a discrete auxiliary variable, which takes the quantized values of the pitch estimates. However, this work ignores the fact that pitch information is meaningful only for some particular phonetic states. While it is reasonable to augment a voice phonetic state like /a/ with an auxiliary variable representing pitch, it is less appropriate to associate such auxiliary variable to an unvoiced phonetic state like /s/. Given this observation, in this paper we propose switching auxiliary chains for modeling different auxiliary information tailored to different phonetic states. In addition, we investigate the broader case of using continuous acoustic features, instead of the discrete ones as in [3][4][5].

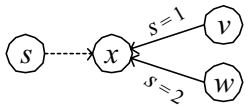

Fig.1: When s=1, v is x's parent, when s=2, w is x's parent.

The switching auxiliary chains are fully implemented in the framework of dynamic Bayesian networks (DBNs), and essentially built on the switching parent functionality of Bayesian multinets [6][7]. Normally in Bayesian networks, a variable has only one set of parents. However, Bayesian multinets allow a variable's parents to change (or switch) depending on the current values of other parents. The parents that may change are called conditional parents, and the parents which control the switching are called switching parents. Fig. 1 shows the case where variable s switches the parents of x between v and w, corresponding to the probability distribution:

$$p(x \mid v, w) = p(x \mid v, s=1)\, p(s=1) + p(x \mid w, s=2)\, p(s=2)$$

The use of switching auxiliary chains are motivated in two ways. First, we may need multiple auxiliary chains for representing different possible auxiliary information, and each chain may be meaningful only for some particular phonetic states. For example, exploiting pitch information is reasonable only for voiced phones. Lip rounding is more relevant for vowel phones, but not for consonantal phones [8], while manner of articulation (lateral, nasal, fricative, approximant) is more relevant for consonantal phones, but not for vowel phones. Switching chain representation can specify dependencies only when necessary, and thus reduce computation and parameter size. Second, although we could increase the cardinality of an auxiliary variable (e.g. by forced including the value of 'nil' or other less relevant values [8]) to make it play the role of conditional parents for all phonetic states, this will potentially increase confusion and hurt performance. For switching chain representation, each chain only

[1] This work was supported by NSFC (No. 60402029).

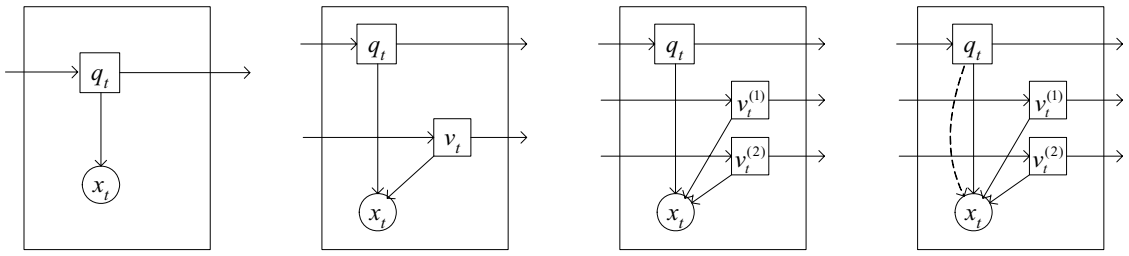

Fig. 2: DBN representation of different models. Round/square nodes show continuous/discrete variables. The arcs entering/leaving the box represent links to the previous/next time frame. (a) The standard HMM; (b) Regular one auxiliary chain model; (c) Regular two auxiliary chain model; (d) Switching two auxiliary chain model, where q_t is both a switching parent (dotted arc) and a conditional parent (solid arc).

needs to distinguish between a small number of relevant phonetic states, thus could be trained with more discriminative (conditional) distributions.

We intend to take advantage of the switching chain representation to exploit as much auxiliary information as possible. For preliminary study in this paper, we focus on implementing a switching two auxiliary chain model tailored to voiced/unvoiced states for exploiting pitch information. The Graphical Model Toolkit (GMTK) [9] is utilized for such implementation.

Experiments were carried out on the OGI Numbers database [10], which is an English telephone speech corpus consisting of continuously spoken numbers. We found significant improvements resulting from switching auxiliary chain modeling, compared with the regular auxiliary chain modeling which does not distinguish between voiced/unvoiced phones regarding the pitch information.

This paper is organized as follows. In Section 2, the regular and switching auxiliary chain models are described in the framework of DBNs. Section 3 presents experimental results, followed by discussion in the last section.

2. Model formulation based on DBNs

2.1. The Standard HMM

Fig. 2(a) shows DBN representation of the standard HMM. q_t, x_t are respectively the (discrete) phonetic state variable and the (continuous) acoustic feature variable at time t. The HMM is parameterized by the state transition probabilities $p(q_t \mid q_{t-1})$ and the state output distributions $p(x_t \mid q_t)$, which is often implemented as Gaussian mixture density:

$$p(x_t \mid q_t) = \sum_{i=1}^{M} p(m_t = i \mid q_t)\, p(x_t \mid q_t, m_t = i)$$

Here m_t denotes the hidden Gaussian component variable, which is not explicitly plotted in Fig. 2 for simplicity. The conditional independent assumptions implied by HMM burden the single q_t to contain all relevant contextual information. One method is to augment q_t with additional (auxiliary) variables that represent contextual auxiliary information (e.g. pitch, rate of speech, etc.).

2.2. Regular Auxiliary Chain Model

In [3][4], an auxiliary chain is formed by linking a conceptual 'context' variable at each time, which is always hidden during both training and recognition. An auxiliary chain directly related to pitch information is studied in [5], where during training, the auxiliary variable is observed as the quantized pitch values for learning pitch-related temporal correlation. Regardless of whether the auxiliary chain is hidden or observable during training/recognition, the regular auxiliary chain model is essentially the DBN shown in Fig. 2(b). v_t denotes the discrete auxiliary variable at time t. The (time-independent) local conditional probability distribution (CPD) associated with node q_t, x_t and v_t are respectively $p(q_t \mid q_{t-1})$, $p(x_t \mid q_t, v_t)$, and $p(v_t \mid v_{t-1})$, which are used to define the joint probability distribution:

$$p(q_{1:T}, x_{1:T}, v_{1:T}) = \prod_{t=1}^{T} p(q_t \mid q_{t-1})\, p(x_t \mid q_t, v_t)\, p(v_t \mid v_{t-1})$$

Instead of assuming a discrete x_t as in [3][4][5], here we use a continuous x_t and implement Gaussian mixture densities for the local CPD $p(x_t \mid q_t, v_t)$. That is, a Gaussian mixture density is used for each combination of the values of q_t and v_t.

Multiple auxiliary chains could be used to represent different possible auxiliary information. Fig. 2(c) shows the DBN using two auxiliary chains. And if the auxiliary information (e.g. the state of articulators) is correlated with phones, we can make the auxiliary variable dependent on the phonetic states. For current work, we consider pitch information, which does not have a direct dependency on the phones. From this viewpoint, it is appropriate to use a state-independent auxiliary chain, as done in [5].

However, to have the pitch auxiliary variable condition the acoustic feature sequence across the whole utterance ignores the fact that pitch information is meaningful only for voiced regions. While it is reasonable to augment a voiced phonetic state like /a/ with an auxiliary variable representing pitch, it is less appropriate to associate such auxiliary variable to an unvoiced phonetic state like /s/. To parsimoniously account for such selective effect of auxiliary information on different phonetic states, we propose the following switching auxiliary chain model.

2.3. Switching Auxiliary Chain Model

The switching auxiliary chains are essentially built on the switching parent functionality of Bayesian multinets. Suppose that there are L auxiliary variables $v_t^{(1)}, \cdots, v_t^{(L)}$, representing different

auxiliary information, then we could have the probability distribution:

$$p\left(q_{1:T}, x_{1:T}, \left\{v_{1:T}^{(l)}\right\}_{l=1,\cdots,L}\right)$$
$$= \prod_{t=1}^{T} p\left(q_t \mid q_{t-1}\right) p\left(x_t \mid q_t, v_t\left(q_t\right)\right) \prod_{l=1}^{L} p\left(v_t^{(l)} \mid v_{t-1}^{(l)}\right)$$

where $v_t\left(q_t\right) \subseteq \left\{v_t^{(1)}, \cdots, v_t^{(L)}\right\}$ is the selective parents of x_t according to different values of q_t. The switching function $v_t\left(q_t\right)$ is intended to be a (deterministically) mapping from a classification of the possible values of q_t to the set of auxiliary variables, where each class is suited to some particular auxiliary information. The classification may be determined using a data-driven approach, or as adopted below, could be specified by a-priori knowledge.

Fig. 2(d) shows the switching two auxiliary chain model (L=2) for exploiting pitch information. For now, we use an observed $v_t^{(1)}$ trained with the quantized pitch values for voiced states, and a hidden $v_t^{(2)}$ trained to encode contextual information other than pitch for unvoiced states. The parents of x_t switches to include either $v_t^{(1)}$ or $v_t^{(2)}$ according to a classification of q_t to voiced/unvoiced states. That is, $v_t\left(q_t\right)$ equals $v_t^{(1)}$ / $v_t^{(2)}$ if q_t takes on a voiced/unvoiced state. This effect is illustrated in Fig. 3 with an instantiation example of the phonetic state variables.

A multinet occurs since $v_t\left(q_t\right)$ is a function of q_t. If the underlying phonetic state chain changes, so will the set of dependencies. In general, the statistical dependencies in a multinet could be represented by a regular Bayesian network via specific values of the parameters [6]. However in practice, switching parent functionality could reduce computation and parameter size, and may improve discrimination through implementing dependencies only when necessary and relevant. For example, if we want to exploit one more auxiliary information complementary to an existing one, using two (binary) auxiliary chains in a regular way will double the model complexity, while using switching two chains will remain almost the same complexity.

3. Experimental Results

Experiments were carried out on the OGI Numbers database [10], 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 [11]. 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) [9] 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

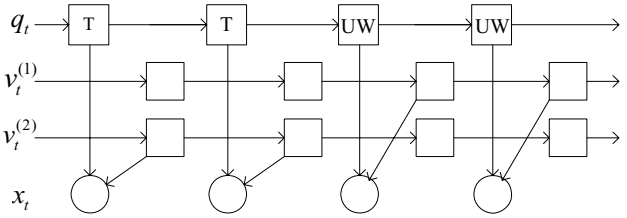

Fig. 3: A switching two auxiliary chain model. An instantiation example of the phonetic state variables for the word 'two' /T/-/UW/ is shown.

with three emitting states each, and the short-pause had only one state which was tied to the middle state of the silence model.

A baseline DBN was first built to emulate the standard HMM. There is an upper layer including position, transition variables as introduced in [3]. The various DBNs replace the lower layer with different structures from Fig. 2.

For current work, we consider pitch as auxiliary information. The ESPS tool 'get_f0' was used for pitch extraction. The pitch estimated were then quantized to binary in two ways: one is to reflect high-low pitch (low: below 140hz including unvoiced frame), and the other is to reflect voiced/unvoiced. Two types of DBN based auxiliary chain models were trained:

1) A regular one auxiliary chain model in which the single binary variable v_t was trained with the quantized high-low pitch values.

2) A switching two auxiliary chain model. The 27*3=81 states were classified into two classes of 60 voiced states and 21 unvoiced states according to phonetic knowledge. As described in Section 2.3, we used a binary observed $v_t^{(1)}$ trained with the quantized pitch values for voiced states, and a binary hidden $v_t^{(2)}$ trained to encode contextual information other than pitch for unvoiced states.

Both types of DBN chain models were then tested under two conditions. For the regular auxiliary chain model, the variable v_t was observed (O) or hidden (H). For the switching chain model, the $v_t^{(1)}$ was observed (O) or hidden (H), and the hidden $v_t^{(2)}$ during training was still hidden during recognition. WER results are shown in Table 1. For all the models, two series of experiments were taken, using 8 and 16 Gaussian mixtures respectively for the CPD with the acoustic feature variable x_t.

The results in Table 1 indicate that exploiting pitch information with a binary auxiliary variable could reduce the WER from the baseline HMM. The performance improvements are more evident and consistent when using switching auxiliary chain models. Significant error rate reductions of 6% (from 10.99% to 10.32%) and 7% (from 10.16% to 9.41%) are obtained for 8 and 16 Gaussian mixture system respectively.

For both chain models, it is also found, as reported in [2][5], that during recognition, it is better to hide the auxiliary variable (H) than to make it observed as the externally measured pitch values. Due to their blind treatment of pitch information, the regular auxiliary chain models performed worse than the switching ones under all conditions. The pitch (frequency) information is relevant only to voiced states; to have the pitch auxiliary variable condition

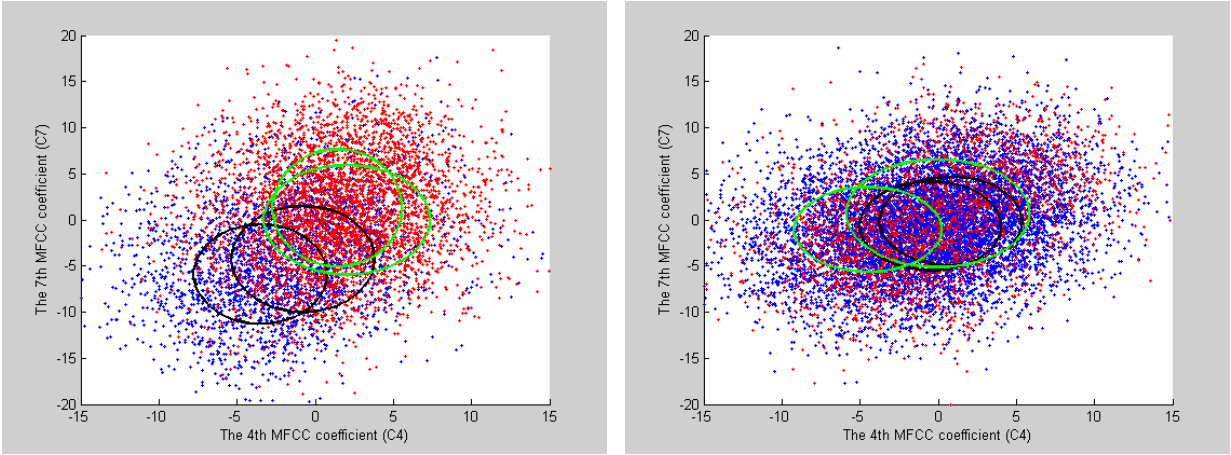

Fig. 4: The left and right picture show the Gaussian models along with the sample data for the second state of 'OW' (voiced) and the second state of 'F' (unvoiced) respectively. The Gaussian models, represented by their contours with 2 standard deviation, are taken from the trained regular one auxiliary chain model, which has 2 Gaussian mixtures for each possible value of the binary auxiliary variable v_t. The black and green correspond to v_t=0 (low pitch including unvoiced) and v_t=1 (high pitch) respectively. The sample data were obtained via force alignment, where blue and red points correspond to low pitch (including unvoiced) and high pitch frames respectively.

Models	Mix.	# of param.	H	O
Regular one aux. chain	8	102K	11.20	11.25
	16	204K	9.76	10.32
Switching two aux. chains	8	102K	10.32	10.87
	16	204K	9.41	9.66
HMM	8	51K	10.99	
	16	102K	10.16	

Table 1: Word Error Rates (%) for different models

the acoustic feature for the unvoiced states is superfluous and sometimes detrimental, as the results in Table 1 show. It is important and beneficial to selectively model the effects of different auxiliary information on different phonetic states, as successfully realized by the switching auxiliary chain representation.

Fig. 4 illustrate why such a selective modeling is necessary. For ease of observation, the 2 Gaussian mixtures are shown here, which were obtained during the process of increasing mixture number in training. It is clear that the conditional Gaussian mixture models for voiced state 'OW' changes systematically according to the values (low/high) of its conditioning pitch auxiliary variable, and cover the corresponding feature data properly. The variation due to pitch is well modeled for voiced states. However, for unvoiced state 'F', the feature data for different values of the pitch auxiliary variable are highly overlapped, and so are the learned conditional Gaussian mixture models. These suggest that pitch (frequency) information is of little use if any to account for the variation for unvoiced states. It is better to use a hidden auxiliary variable than to use an irrelevant pitch variable for unvoiced states.

4. Conclusions and Future

In this paper, we propose switching auxiliary chains for modeling different auxiliary information tailored to different phonetic states. In particular, we implement a switching two auxiliary chain model tailored to voiced/unvoiced states for exploiting pitch (frequency) information, and achieve significant performance improvements. In the future, we intend to take advantage of the switching chain representation to exploit as much auxiliary information as possible, and experiment with increasing cardinality of the auxiliary chains.

5. References

[1] K. Fujinaga, M. Nakai, H. Shimodaira, and S. Sagayama, "Multiple-regression hidden Markov model," in *Proc. ICASSP* 2001.

[2] T.A. Stephenson, M. Mathew, and H. Bourlard, "Speech Recognition with Auxiliary Information," *IEEE trans. on Speech and Audio Processing*, vol.12, No.3, 2004.

[3] G. Zweig, "Speech Recognition with Dynamic Bayesian Networks," Ph.D. dissertation, Univ. California, Berkeley, 1998.

[4] G. Zweig and M. Padmanabhan, "Dependency modeling with Bayesian networks in a voicemail transcription system," in *Proc. EUROSPEECH* 1999.

[5] T.A. Stephenson, M. Mathew, and H. Bourlard, "Modeling auxiliary information in Bayesian network based ASR," in *Proc. EUROSPEECH* 2001.

[6] D. Geiger, D. Heckerman, "Knowledge representation and inference in similarity networks and Bayesian multinets," Artificial Intelligence, Vol.82, 1996.

[7] J. Bilmes, "Dynamic Bayesian Multinets," in *Proc. UAI* 2000.

[8] K. Kirchhoff, G. Fink and G. Sagerer. "Combining acoustic and articulatory feature information for robust speech recognition," *Speech Communication*, Vol.37, 2002.

[9] J. Bilmes, and G. Zweig, "The Graphic Models Toolkit: An open source software system for speech and time-series processing", in *Proc. ICASSP* 2002.

[10] R.A. Cole M. Fanty, M. Noel, and T. Lander, "Telephone speech Corpus development at CSLU," in *Proc.ICSLP* 1994.

[11] Multi Channel Overlapping Numbers Corpus (MONC) distribution. http://cslu.cse.ogi.edu/corpora/