Switching Auxiliary Chains for Speech Recognition

Hui Lin, *Student Member, IEEE*, and Zhijian Ou, *Member, IEEE*

***Abstract*—This letter investigates the problem of incorporating auxiliary information, e.g., pitch, zero crossing rate (ZCR), and rate-of-speech (ROS), for speech recognition using dynamic Bayesian networks. In this letter, we propose switching auxiliary chains for exploiting different auxiliary information tailored to different phonetic states. The switching function can be specified by *a priori* knowledge or, more flexibly, be learned from data with information-theoretic dependency selection. Experiments on the OGI Numbers database show that the new model achieves 7% word-error-rate relative reduction by jointly exploiting pitch, ZCR, and ROS, while keeping almost the same parameter size as the standard HMM.**

***Index Terms*—Auxiliary features, dynamic Bayesian networks (DBNs), speech recognition.**

I. Introduction

FOR automatic speech recognition (ASR), 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, zero crossing rate, rate of speech, etc.), which we call auxiliary information. Various methods have been proposed to incorporate such auxiliary information to improve ASR robustness. Bayesian networks [1], in particular, dynamic Bayesian networks (DBN) [2], in which HMMs can be considered as one small instance, has been used for these studies [3]–[7].

One method is to encode the auxiliary information in continuous observed variables. It is shown in [5] and [8] that simply appending the auxiliary feature to the standard feature vector (MFCCs) degrades the recognition performance. It is beneficial, however, to use the auxiliary feature as a conditional variable to model the distribution of the cepstral-based features. To have tractable exact inference when using hidden continuous variables, only the dependencies within a given time frame are considered [5].

On the other hand, the auxiliary information can also be incorporated in the form of discrete variables [3], [4], [6], [7], which can be temporally linked to form an auxiliary chain along time to account for contextual information. The works in [3] and [4] show the advantage of including a discrete context variable that is always hidden during both training and recognition; it is not clear what auxiliary information it may represent. In [6], pitch information is explicitly related to a discrete variable by quantization. In [7], rate-of-speech (ROS) information is used by introducing an additional discrete mode variable. There are other possible sources for auxiliary information, e.g., zero crossing rate (ZCR), short-term energy. Each auxiliary feature has its own merit to aid the modeling of the standard feature O_t by conditioning O_t's distribution. Previous works mainly investigate the use of each auxiliary feature separately [6], [7]. If we consider jointly exploiting the different auxiliary features, say $A_t^{(1)}, \ldots, A_t^{(L)}$, via L auxiliary chains, a problem is that the model complexity will be increased by a factor equal to the product of the cardinalities of the L auxiliary variables.

Note that, for different phonetic states, the effects of an auxiliary feature may be different. For example, pitch information is meaningful only for voiced phones. 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/. It is also observed that vowels receive greater influence than consonants by speaking rate [9]. ZCR is well known to be useful for voiced/unvoiced detection, and additionally, it also helps to distinguish voiceless plosives (including affricates) from voiceless fricatives [10]. With these observations, in this letter, we propose switching auxiliary chains for exploiting different auxiliary information tailored to different phonetic states.

The new model is essentially built on the switching parent functionality of Bayesian multinets [12], [13]. Normally in Bayesian networks, a variable has only one set of parents. However, in Bayesian multinets, a variable's parents are allowed to change (or switch) depending on the current values of other parents (as illustrated in Fig. 1). This switching functionality enables us to jointly use multiple auxiliary features in a parsimonious way, by selectively exploiting the auxiliary features for different phonetic states. For each phonetic state, only the most effective auxiliary feature is switched to be the parent of the standard feature. The switching function can be specified by *a priori* knowledge (e.g., in our previous work [11], we implemented a knowledge-driven switching voiced/unvoiced auxiliary chain model for exploiting pitch information), or more flexibly, be learned from data with information-theoretic dependency selection [13].

In the data-driven approach, for each value q of the phonetic state Q_t, the conditional mutual information $I(O_t, A_t^{(l)} \,|\, Q_t = q)$ between the standard feature O_t and each individual auxiliary feature $A_t^{(l)}$, is computed using training data. Then the auxiliary feature which corresponds to the maximum mutual information is selected as the conditional parent of the standard feature.

Manuscript received September 1, 2006. This work was supported by the National Natural Science Foundation of China under Grant 60402029. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Steve Renals.

H. Lin is with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China, and also with the Electrical Engineering Department, University of Washington, Seattle, WA 98195-2500 USA (e-mail: linhui99@mails.tsinghua.edu.cn; linhui@u.washington.edu).

Z. Ou is with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (e-mail: ozj@tsinghua.edu.cn).

Digital Object Identifier 10.1109/LSP.2006.891314

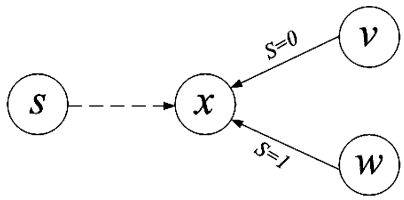

Fig. 1. Illustration of Bayesian multinets. Solid lines represent conditional dependency. Dashed lines represent switching dependency. When $s = 0$, v is x's parent, when $s = 1$, w is x's parent. s is called the switching parent and v, w are called the conditional parents. s switches the parents of x between v and w, corresponding to the probability distribution: $p(x \mid v, w) = p(x \mid v, s = 0)p(s = 0) + p(x \mid w, s = 1)p(s = 1)$.

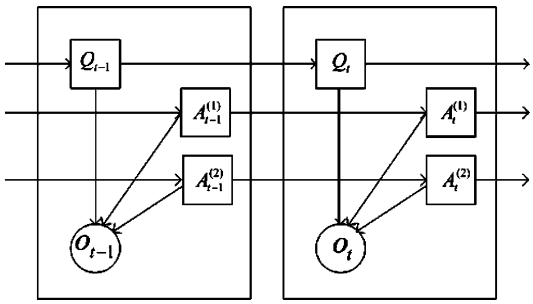

Fig. 2. Basic auxiliary chain model. Round nodes represent continuous variables, while square nodes represent discrete variables.

Experiments are carried out on the OGI Numbers database [14], which is an English telephone speech corpus consisting of continuously spoken numbers. The results show that the new model achieves 7% word-error-rate (WER) relative reduction by jointly exploiting pitch, ZCR, and ROS as the auxiliary features, while remaining almost the same parameter size as the standard HMM.

II. Switching Auxiliary Chains: Model Formulation

The switching auxiliary chains are implemented in the framework of dynamic Bayesian networks (DBNs) [2]. Fig. 2 shows the DBN representation of the basic auxiliary chain model as in [3], [4], and [6]. Q_t, O_t are, respectively, the discrete phonetic state variable and the continuous standard feature variable at time t. In Fig. 2, there are two discrete auxiliary variables $A_t^{(1)}$, $A_t^{(2)}$, which can be used to encode two kinds of auxiliary information. Suppose that there are L auxiliary variables $A_t^{(1)}, A_t^{(2)}, \ldots, A_t^{(L)}$, representing L different types of auxiliary information, then we have the following joint probability distribution:

$$p\left(Q_{1:T}, O_{1:T}, \left\{A_{1:T}^{(l)}\right\}_{l=1,2,\ldots,L}\right) = \prod_{t=1}^{T} p(Q_t \mid Q_{t-1}) p\left(O_t \mid Q_t, \left\{A_t^{(l)}\right\}_{l=1,2,\ldots,L}\right) \cdot \prod_{l=1}^{L} p\left(A_t^{(l)} \mid A_{t-1}^{(l)}\right). \quad (1)$$

Traditionally, the new conditional probabilistic distribution (CPD) of O_t, $p(O_t \mid Q_t, \{A_t^{(l)}\}_{l=1,\ldots,L})$, requires a model

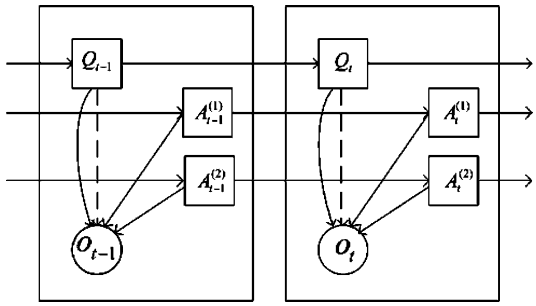

Fig. 3. Switching auxiliary chain model. Q_t is both the conditional parent and switching parent of O_t, and $A_t^{(1)}, A_t^{(2)}$ are O_t's conditional parents.

for each possible combination of the auxiliary variables $\{A_t^{(l)}\}_{l=1,\ldots,L}$. Note that, for different phonetic states, the effects of an auxiliary feature are different, maybe strong, weak or even irrelevant. A good example is that pitch information is meaningful only for voiced states. To account for such selective effect of auxiliary information for different phonetic states, we propose the following switching auxiliary chain model, which also reduces model complexity:

$$p\left(Q_{1:T}, O_{1:T}, \left\{A_{1:T}^{(l)}\right\}_{l=1,2,\ldots,L}\right) = \prod_{t=1}^{T} p(Q_t \mid Q_{t-1}) p(O_t \mid Q_t, A_t(Q_t)) \times \prod_{l=1}^{L} p\left(A_t^{(l)} \mid A_{t-1}^{(l)}\right) \quad (2)$$

where $A_t(Q_t) \subseteq \{A_t^{(1)}, A_t^{(2)}, \ldots, A_t^{(L)}\}$ is selected, by Q_t, to be a parent of O_t. Here, Q_t is both the conditional parent and switching parent of O_t, and $A_t^{(1)}, A_t^{(2)}, \ldots, A_t^{(L)}$ are O_t's conditional parents, as illustrated in Fig. 3

In the traditional multiple auxiliary chain model, all auxiliary features are used without selection for different values of Q_t. In contrast, the new model selects the most effective auxiliary features to condition the modeling of O_t, as specified by the switching function $A_t(Q_t)$. A multinet occurs here since $A_t(Q_t)$ is a function of Q_t. If the underlying phonetic state chain changes, so will the set of dependencies.

The switching function $A_t(Q_t)$ is defined to be a deterministic 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 mapping can be specified by *a priori* knowledge, or more flexibly, be learned from data with information-theoretic dependency selection, as described below.

III. Data-Driven Switching

For specific phonetic state $Q_t = q$, we define the effectiveness of an auxiliary feature $A_t^{(l)}$ by the strength of the (conditional) dependency between O_t and $A_t^{(l)}$. This can be naturally measured by the conditional mutual information $I(O_t; A_t^{(l)} \mid Q_t = q)$. If the conditional mutual information is large, the auxiliary feature is viewed as being effective for the modeling of the standard feature for state q.

Furthermore, if we constrain that only one auxiliary feature is selected to take effect for each phonetic state, that is

$$|A_t(Q_t = q)| = 1, \forall q \tag{3}$$

then the model size of using L auxiliary features remains almost the same as that of using only one auxiliary feature. It can be easily seen that the switching function under this constraint should be

$$A_t(Q_t = q) = \arg\max_{A_t^{(l)}} I(O_t; A_t^{(l)} \,|\, Q_t = q). \tag{4}$$

What remains is to obtain the conditional mutual information $I(O_t; A_t^{(l)} \,|\, Q_t = q)$, which can be computed as follows:

$$\begin{aligned} & I\left(O_t, A_t^{(l)} \,\middle|\, Q_t = q\right) \\ & \quad = H(O_t \,|\, Q_t = q) - H\left(O_t \,|\, Q_t = q, A_t^{(l)}\right) \\ & \quad = H(O_t \,|\, Q_t = q) \\ & \qquad - \sum_a p\left(A_t^{(l)} = a \,\middle|\, Q_t = q\right) \\ & \qquad \cdot H\left(O_t \,\middle|\, Q_t = q, A_t^{(l)} = a\right). \end{aligned} \tag{5}$$

Here, for a random variable X with the distribution $p(x)$,, $H(X)$ denotes its entropy [15], which is defined as

$$H(X) = -\int p(x) \log p(x) dx. \tag{6}$$

Therefore, we first need to estimate the distributions $p(O_t \,|\, Q_t = q)$ and $p(O_t \,|\, Q_t = q, A_t^{(l)} = a), l = 1, \ldots, L$. These can be achieved by fitting Gaussian mixture distributions to the training data. Specifically, the training data associated with specific $Q_t = q$ can be obtained via forced alignment using the standard HMM. For $l = 1, \ldots, L$, the aligned data can be further divided into several groups according to different values of $A_t^{(l)}$. Analytical formulae exist for the entropy of a Gaussian, but not for the entropy of a Gaussian mixture density (GMD); therefore, we use a Monte Carlo sampling method to compute the entropy in (6). In the experiments, GMDs with diagonal covariance matrices are used. It is easy to draw a set of samples $\{x^{(i)} \,|\, i = 1, \ldots, N\}$ from such diagonal GMDs, say $p(x)$ [16]. Then the entropy can be approximated by a finite sum

$$H(X) \simeq -\frac{1}{N} \sum_{i=1}^{N} \log p\left(x^{(i)}\right). \tag{7}$$

Finally, from the aligned training data, we have

$$p\left(A_t^{(l)} = a \,\middle|\, Q_t = q\right) = \frac{\sum_{t \in \left\{t \,\middle|\, Q_t = q, A_t^{(l)} = a\right\}} 1}{\sum_a \sum_{t \in \left\{t \,\middle|\, Q_t = q, A_t^{(l)} = a\right\}} 1}. \tag{8}$$

Combining (5), (7), and (8), we can obtain the required conditional mutual information.

TABLE I
WERs FOR DIFFERENT MODELS

Model	Aux.	Param.	WER(%)
HMM	–	101k	9.80
Single Aux. Chain	pitch	102k	9.39
	ZCR	102k	9.36
	ROS	102k	9.59
Switching Two Aux. Chains (Knowledge-Driven)	pitch+ZCR	102k	9.25
	pitch+ROS	102k	9.40
Switching Two Aux. Chains (Data-Driven)	pitch+ZCR	102k	9.22
	pitch+ROS	102k	**9.18**
	ZCR+ROS	102k	9.35
Switching Three Aux. Chains (Data-Driven)	pitch+ZCR+ROS	102k	**9.14**

IV. EXPERIMENTS

Experiments are carried out on the OGI Numbers database [14], which is an English telephone speech corpus consisting of naturally spoken numbers with 30-word vocabulary. We use 6049 utterances from the corpus for training and 2061 utterances for testing. All utterances are framed with 25-ms length and 10-ms shift. From each frame, 12 MFCCs plus normalized log-energy are extracted along with their first and second derivatives, giving a feature vector of 39 dimensions. Cepstral mean subtraction is then applied to the feature vector. The Graphical Model Toolkit (GMTK) [17] is utilized for DBN implementation. There are 26 monophone models, a silence model, and a short-pause model. The silence and all monophones are modeled with three emitting states each, and the short-pause has only one state which is tied to the middle state of the silence model.

Three kinds of auxiliary information are used: pitch (the fundamental frequency f_0), ZCR, and ROS. ROS is estimated by the *mrate* program [18] and the Entropic Signal Processing System (ESPS) [19] tool $\mathrm{get\_}f_0$ is used to estimate the pitch. All the extracted auxiliary information is then quantized into binary auxiliary features.

A baseline DBN is built to emulate the standard HMM. There is an upper layer including position and transition variables as introduced in [3]. The various DBNs replace the lower layer with the different structures from Figs. 2 and 3. 16 Gaussian components per state are used for the CPD of the acoustic feature O_t.

In the auxiliary chain models, for each state, the Gaussian components of the GMDs for different values of the quantized auxiliary variable are tied for robust parameter estimation, as done in [4], [7], and [20]. Only the mixture weights depend on the auxiliary chain variable. The WER results for various models are shown in Table I, along with the model parameter size.

First, we implement three single auxiliary chain models, using quantized pitch, ZCR, and ROS as the auxiliary variable, respectively. It can be seen that incorporating auxiliary information via a single chain is beneficial. Every single auxiliary chain model outperforms the baseline HMM. As we use tied Gaussian mixtures on the state level, introducing the auxiliary variable keeps almost the same number of parameters as the baseline HMM. If no parameter tying scheme was used, using a single binary auxiliary chain would double the parameter size,

which may cause unreliable parameter estimation. As reported in [20], using discrete auxiliary chain to carry pitch information without parameter tying may even degrade the recognition performance.

Next, we consider jointly exploiting multiple auxiliary features via switching auxiliary chains. Since we constrain the switching function to select only one auxiliary feature to take effect as in (3), the number of parameters of all the new models remain almost unchanged. For comparison, both the knowledge-driven and data-driven methods to determine the switching function are implemented.

Consider the case of exploiting two different auxiliary features. For the knowledge-driven method, we divide the phonetic states into voiced/unvoiced classes by *a priori* knowledge, and assign relevant auxiliary features to them, respectively. Two possible switching two-chain models are tested, which can be expressed as in the following, respectively, using the switching function notation:

$$A_t(Q_t = q) = \begin{cases} \text{quantized pitch}, & \text{if } q \text{ is voiced} \\ \text{quantized ZCR}, & \text{otherwise} \end{cases}$$

$$A_t(Q_t = q) = \begin{cases} \text{quantized pitch}, & \text{if } q \text{ is voiced} \\ \text{quantized ROS}, & \text{otherwise.} \end{cases}$$

For the data-driven method, all the three possible combinations of the three auxiliary features are considered. The switching function is determined by first computing conditional mutual information, and then selecting the auxiliary feature with the maximum effectiveness. In the sampling method used for entropy computation, diagonal GMDs with eight components are estimated, and one million samples are drawn. Using more samples or more Gaussian components gives the identical switching function in our experiment.

It can be seen from Table I that the switching two auxiliary chain models further improve the performance over the corresponding single auxiliary chain models, while keeping almost the same model size. The results also show that the data-driven approach is more effective and flexible than the knowledge-driven approach. The performance of the knowledge-driven switching two chain models is worse than corresponding data-driven models, especially in the case of using pitch and ROS. In this case, the knowledge-driven switching two chain model performs similarly to using pitch alone. In contrast, the corresponding data-driven model gives the best result among all switching two chain models.

Finally, the data-driven switching three auxiliary chain model is implemented, exploiting pitch, ZCR and ROS together. It further improves the performance, achieves 7% WER relative reduction over the baseline HMM, and again, keeps almost the same model size. Jointly exploiting multiple auxiliary features by information-theoretic dependency selection via switching function proves to be useful for building compact yet powerful acoustic models.

V. Conclusion

In this letter, we propose switching auxiliary chains for exploiting different auxiliary information tailored to different phonetic states. The new model is essentially built on the switching parent functionality of Bayesian multinets. The switching function can be determined by *a priori* knowledge, or more flexibly, be learned from data with information-theorectic dependency selection. Experiments on the OGI Numbers database show that the new model achieves 7% WER relative reduction by jointly exploiting pitch, ZCR, and ROS, while keeping almost the same parameter size as the standard HMM. In the future, we plan to use the switching chain representation to exploit as much auxiliary information as possible.

References

[1] R. G. Cowell, A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter, *Probabilistic Networks and Expert Systems*. New York: Springer-Verlag, 1999.

[2] T. Dean and K. Kanazawa, "Probabilistic temporal reasoning," in *Proc. AAAI*, 1988, pp. 524–538.

[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," presented at the EuroSpeech Conf., 1999.

[5] T. Stephenson, M. Mathew, and H. Bourlard, "Speech recognition with auxiliary information," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 3, pp. 189–203, 2004.

[6] M. M. T.A. Stephenson and H. Bourlard, "Modeling auxiliary information in Bayesian network based ASR," presented at the EuroSpeech Conf., 2001.

[7] T. Shinozaki and S. Furui, "Hidden mode hmm using Bayesian networks for modeling speaking rate fluctuation," in *Proc. IEEE Automatic Speech Recognition and Understanding*, 2003, pp. 417–422.

[8] K. Fujinaga, M. Nakai, H. Shimodaira, and S. Sagayama, "Multiple-regression hidden Markov model," presented at the Int. Conf. Acoustics, Speech, Signal Processing, 2001.

[9] H. Kuwabara, "Acoustic properties of phonemes in continuous speech for different speaking rate," presented at the Int. Conf. Speech Language Processing, 1996.

[10] L. Weigelt, S. Sadoff, and J. Miller, "Plosive/fricative distinction: The voiceless case," *J. Acoust. Soc. Amer.*, pp. 2729–37, Jun. 1990.

[11] H. Lin and Z. Ou, "Switching auxiliary chains for speech recognition based on dynamic Bayesian networks," presented at the Int. Conf. Pattern Recognition, 2006.

[12] D. Geiger and D. Heckerman, "Knowledge representation and inference in similarity networks and Bayesian multinets," *Artif. Intell.*, 1996.

[13] J. Bilmes, "Dynamic Bayesian multinets," presented at the UAI Conf., 2000.

[14] R. Cole, M. Fanty, M. Noel, and T. Lander, "Telephone speech corpus development at CSLU," presented at the Int. Conf. Speech Language Processing, 1994.

[15] T. Cover and J. Thomas, *Elements of Information Thoery*. New York: Wiley, 1991.

[16] D. MacKay, *Introduction to Monte Carlo Methods*, M. Jordan, Ed. Cambridge, MA: MIT Press, 1998.

[17] J. Bilmes and G. Zweig, "The Graphical Models Toolkit: An open source software system for speech and time-series processing," presented at the Int. Conf. Acoustics, Speech, Signal Processing, 2002.

[18] N. Morgan, E. Fosler, and N. Mirghafori, "Speech recognition using on-line estimation of speaking rate," presented at the EuroSpeech Conf., 1997.

[19] *ESPS with waves*, Entropic Research Laboratory, Inc., AT&T Bell Laboratories, 1993.

[20] H. Lin and Z. Ou, "Partial-tied-mixture auxiliary chain models for speech recognition based on dynamic Bayesian networks," presented at the IEEE Int. Conf. Systems, Man, Cybernetics, 2006.