The Use of Bayesian Network for Incorporating Accent, Gender and Wide-Context Dependency Information

Sakriani Sakti[1,2], *Konstantin Markov*[1,2], *and Satoshi Nakamura*[1,2]

[1]National Institute of Information and Communications Technology, Japan
[2]ATR Spoken Language Communication Research Laboratories, Japan
{sakriani.sakti,satoshi.nakamura,konstantin.markov}@atr.jp

Abstract

We propose a new method of incorporating the additional knowledge of accent, gender, and wide-context dependency information into ASR systems by utilizing the advantages of Bayesian networks. First, we only incorporate pentaphone-context dependency information. After that, accent and gender information are also integrated. In this method, we can easily extend conventional triphone HMMs to cover various sources of knowledge. The probabilistic dependencies between a triphone context unit and additional knowledge are learned through a BN. Another advantage is that during recognition, additional knowledge variables are assumed to be hidden, so that the existing standard triphone-based decoding system can be used without modification. The performance of the proposed model was evaluated on an LVCSR task using two different types of accented English speech data. Experimental results show that this proposed method improves word accuracy with respect to standard triphone models.

Index Terms: acoustic modeling, bayesian network, knowledge incorporation, wide-context dependency.

1. Introduction

Most current automatic speech recognition (ASR) systems usually use statistical data-driven approaches based on hidden Markov models (HMMs). A triphone acoustic unit is commonly used that includes the immediately preceding and following phonetic contexts. Although such statistical models have proven to be efficient choices, they are still deemed insufficient to handle the sources of variability that exist in everyday conversational speech. By completely relying on statistical models, only a limited level of success can be achieved.

Many researchers have tried to improve acoustic models by incorporating coarticulation effects of longer spans, such as tetraphone, quinphone/pentaphone, or etc. To date, the IBM and AT&T large-vocabulary continuous speech recognition (LVCSR) systems have quite successfully used pentaphone models [1, 2]. Various attempts also exist that integrate more explicitly knowledge-based and statistical approaches. As an example, research work in [3] proposed to incorporate acoustic phonetic knowledge sources using neural networks for rescoring frameworks. Recently, Bayesian Networks (BN) have also attracted the attention of speech recognition researchers. A BN can model complex joint probability distributions of many different (discrete and/or continuous) random variables in well structured and easy to represent ways [4]. Another advantage of BNs is that additional features which are difficult to estimate reliably during recognition may be left hidden, i.e., unobservable. In some of the first reports on Dynamic BNs (DBN) in speech recognition [5, 6], they were regarded as a generalization of HMM, which in addition to speech spectral information can easily incorporate additional knowledge, such as articulatory features, sub-band correlation, or speaking styles.

The approach we propose in this paper incorporates such additional knowledge as accent, gender, and wide-context dependency information by utilizing the BN advantages, while allowing us to retain the existing: (1) HMM-based triphone acoustic model topology and (2) standard triphone-based decoding system. This method is based on a scheme proposed in [7, 8], the so-called hybrid HMM/BN modeling framework, since temporal speech characteristics are still governed by standard HMM state transitions, but BN is used underneath to infer the state output likelihood. With this method, we can easily extend conventional triphone HMM to cover a wider context where probabilistic dependencies between the triphone context unit and various knowledge sources are learned through BNs. Our standard triphone-based decoding system can still be used without modification, since additional knowledge variables are assumed hidden during recognition. In our previous study [9], we have shown that by only incorporating the pentaphone context at the left and right states of the triphone HMM, our system achieved up to 10% relative word error rate (WER) reduction on an LVCSR task using the Wall Street Journal (American English) speech corpus [10]. In this paper, we explore ways to extend pentaphone HMM/BN models and investigate their performance on more challenging accented English speech data.

In the next section, we briefly describe the acoustic modeling structure, including the HMM/BN background, the proposed model topology, and the knowledge-based phoneme classes. In Section 3, we describe training and recognition issues. Details of experiments are presented in Section 4, including results and discussion. A conclusion is drawn in Section 5.

2. Acoustic Model Structure

2.1. HMM/BN Background

Figure 1 shows block diagrams of the conventional mixture of Gaussian HMM and the HMM/BN models. In both cases, temporal speech characteristics are governed by HMM state transitions. But, in contrast to the conventional mixture of Gaussians, the HMM/BN model uses a BN underneath to model HMM state probability distribution, which allows for very flexible and consistent models of state probability distribution that can easily integrate different speech parameterizations.

This HMM/BN model is described by two sets of probabilities: HMM transition probabilities $P(q_i|q_j)$ and the joint probability distribution of BN $P(Z_1, ..., Z_K)$, where Z_k, $k = 1, ..., K$

are BN variables. The BN joint probability density function (PDF) can be factorized as:

$$P(Z_1, Z_2, ..., Z_K) = \prod_{k=1}^{K} P(Z_k | Pa(Z_k)), \quad (1)$$

where $Pa(Z_k)$ denotes the parents of variable Z_k.

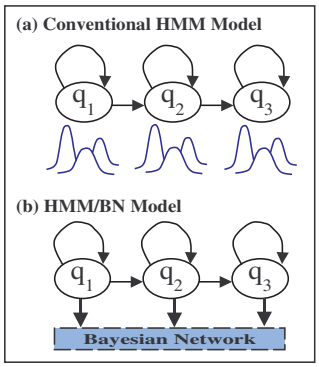

Figure 1: *(a) Conventional mixture of Gaussian HMM model, (b) HMM/BN model.*

Figure 2 shows several different examples of simple BN structures where variable Q represents the HMM state, X represents the spectrum observation variable, and both W and Y represent other additional information, such as pitch, articulatory positions, speaker gender, context information, etc. Here, Q, W, and Y are discrete variables denoted by square nodes, and X is a continuous variable denoted by a circle node. The dependency between two variables (parent and child nodes) is denoted by an arc and described by a conditional probability function. Since it is usually difficult to automatically learn BN structure, it is designed manually based on our knowledge of the data. More details about the HMM/BN approach can be found in [7, 8].

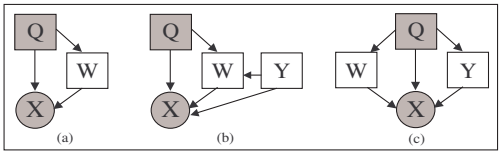

Figure 2: Three simple examples of different BN structures with variables Q,W,Y, and X.

2.2. Proposed HMM/BN Model

In our proposed HMM/BN, the HMM at the top level corresponds to triphone-context acoustic unit $/a^-, a, a^+/$. The BN at the bottom level is used to model the probabilistic dependencies between triphone-context units and various knowledge sources.

First, we only incorporate the pentaphone-context dependency information. If we extend the conventional triphone HMM with additional second preceding and succeeding contexts, we have a pentaphone context like $/a^{--}, a^-, a, a^+, a^{++}/$. Each HMM state output probability distribution can be represented by a BN topology, as shown in Figs. 3(a), which has two additional variables C_L for second preceding context $/a^{--}/$ and C_R for second succeeding context $/a^{++}/$. More details about other possibilities of pentaphone HMM/BN models can be found in [9, 11].

Next, we attempt to extend the pentaphone HMM/BN models by integrating the accent and gender information. By extending the pentaphone BN with an additional variable of gender G, BN topology becomes as shown in Fig. 3(b) and called BN-b. But if we extend with an additional variable of accent A, BN topology becomes as shown in Fig. 3(c) and called BN-c. The BN topology shown in Fig. 3(d) was extended with both additional variables of accent A and gender G, called BN-d.

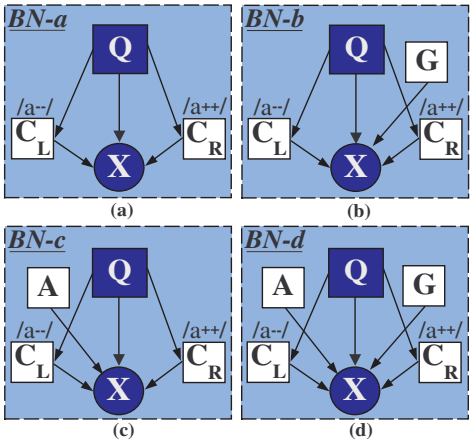

Figure 3: (a) is the BN-a topologies with additional variables C_L and C_R, (b) is the BN-b topologies with additional variables G, C_L and C_R, (c) is the BN-c topologies with additional variables A, C_L and C_R, and (d) is the BN-d topologies with additional variables A, G, C_L and C_R.

2.3. Knowledge-Based Phoneme Classes

The number of parameters for additional variables C_L and C_R equal the number of phonemes that appear in second preceding and second succeeding contexts, respectively. If we use a 44-phoneme set, it denotes that each C_L and C_R has 44 possible values ($C = c_1, c_2, ..., c_{44}$). Here, we attempt to classify the phoneme contexts based on major distinctions in the manner of articulation, in order to reduced the parameter size. Table 1 shows an example of knowledge-based phoneme classes adapted from [12]:

Table 1: *Knowledge-based phoneme classes based on manner of articulation.*

Classes	Phonemes
Plosives	b, d, g, k, p, t
Nasal	m, n, ng
Fricatives	ch, dh, f, jh, s, sh, th, v, z, zh
Liquid	hh, l, r, w, y
Vowels	ih, ix, iy, eh, ey, aa, ae, aw, axr, ay, er, ao, ow, oy, uh, ah, ax, uw

Vowels classes can be further reduced by classification using long-short vowels, front-central-back vowels, or i-e-a-o-u vowels.

3. Training and Recognition Issues

Parameter learning of the proposed model can be adopted from the general training of the HMM/BN model [7]. It is based on the forward-backward algorithm where each training consists of BN training and HMM transition probabilities updates. BN training is done using standard statistical methods. Since all variables, including triphone state Q, accent A, gender G, second preceding (C_L) context, second following (C_R) context, and probability distribution X are observable during training, simple ML parameter estimation can be applied. More details can be found in [7, 8]

Recognition in a conventional HMM is obtained by calculating the state output probability, where state PDF is usually repre-

sented by Gaussian mixture density:

$$P(x_t|q_i) = \sum_{m=1}^{M} b_m \mathcal{N}(x_t; \mu_m, \Sigma_m), \quad (2)$$

where b_m is the mixture weight for the m_{th} mixture in state q_i, and $\mathcal{N}(.)$ is a Gaussian function with mean vector μ_m and covariance matrix Σ_m.

In the case of pentaphone HMM/BN using BN-a (see Fig. 3(a)), state PDF is the BN joint probability model expressed as:

$$\begin{aligned} & P(X, C_L, C_R, Q) \\ & = P(X|C_L, C_R, Q)P(C_L|Q)P(C_R|Q)P(Q), \end{aligned} \quad (3)$$

where it depends on both second preceding context C_L and second following context C_R. $P(X|C_L, C_R, Q)$ is modeled by Gaussian density, and each $P(C_L|Q)$ and $P(C_R|Q)$ is represented by CPT. During recognition, state output probability is obtained from the BN assuming also that both additional variables C_L and C_R are hidden during recognition and take N_L and N_R values:

$$P(x_t|q_i) = \sum_{c_l=1}^{N_L} \sum_{c_r=1}^{N_R} P(c_l|q_i)P(c_r|q_i)P(x_t|c_l, c_r, q_i), \quad (4)$$

where for simplicity, we use x_t, q_i, c_l, and c_r instead of $\langle X = x_t \rangle$, $\langle Q = q_i \rangle$, $\langle C_L = c_l \rangle$, and $\langle C_R = c_r \rangle$, respectively. Here, we can see that Eq. (4) is equivalent to the state output probability of the conventional HMM of Eq. (2) if we treat term $P(c_l|q_i)P(c_r|q_i)$ as a mixture weight coefficient for Gaussian component $P(X|c_l, c_r, q_i)$.

For extended pentaphone HMM/BN models, state output probability is obtained using the same consideration. For example, the BN-d joint probability model is expressed as $P(X, C_L, C_R, Q, A, G)$ which depends on accent A, gender G, the second preceding context C_L and the second following context C_R. The additional variables A and G can also be represented by CPT. During recognition, state output probability is obtained from BN assuming also all additional variables A, G, C_L and C_R are hidden during recognition and take N_A, N_G, N_L and N_R values:

$$\begin{aligned} P(x_t|q_i) \quad = \quad & \sum_{a=1}^{N_A} \sum_{g=1}^{N_G} \sum_{c_l=1}^{N_L} \sum_{c_r=1}^{N_R} P(a)P(g)P(c_l|q_i)P(c_r|q_i) \\ & P(x_t|c_l, c_r, q_i, a, g) \end{aligned} \quad (5)$$

where for simplicity, we use x_t, q_i, a, g, c_l, and c_r instead of $\langle X = x_t \rangle$, $\langle Q = q_i \rangle$, $\langle A = a \rangle$, $\langle G = g \rangle$, $\langle C_L = c_l \rangle$, and $\langle C_R = c_r \rangle$, respectively. Here, we can see that Eq. (5) is also equivalent to the state output probability of the conventional HMM of Eq. (2) if we treat term $P(a)P(g)P(c_l|q_i)P(c_r|q_i)$ as a mixture weight coefficient for the Gaussian component $P(X|c_l, c_r, q_i, a, g)$.

Using these expressions (Eqs. (4) and (5)), we can perform recognition using existing triphone HMM based decoders without modification.

4. Experimental Results and Discussion

The ATR accented English speech corpus, was used in this experiments. The text material was based on basic travel expression domain. The speech corpus we used consisted of American (US) and Australian (AUS) English accents, with about 45k utterances (44 speech hours) spoken by 100 speakers (50 Males, 50 Females) for each accent. As training data, we use 90% of the data or about 40k utterances (20k utterances by 40 speakers for each male and female). Then, we randomly selected 200 utterances spoken by 20 different speakers (10 Males, 10 Females) from the 10% of each accented test data. We use both bi-gram and tri-gram language models which were trained on about 150,000 travel-related sentences. The available pronunciation dictionary consists of about 37k words which is based on US accent pronunciation.

A sampling frequency of 16 kHz, a frame length of a 20-ms Hamming window, a frame shift of 10 ms, and 25 dimensional feature parameters consisting of 12-order MFCC, Δ MFCC and Δ log power are used as feature parameters. Three states were used as initial HMM for each phoneme. Then, a shared state HMnet topology was obtained using a successive state splitting (SSS) training algorithm. Since the SSS algorithm used here is based on the minimum description length (MDL) optimization criterion, the number of shared HMM states is determined automatically by the algorithm. Details about MDL-SSS can be found in [13]. For topology training, we combined all training data (US+AUS) to get the same topology structure for all accent models. Then, an embedded training procedure was done for each accent to get US and AUS triphone HMM acoustic models. For each model, the total number of states is 2,126 with four different versions of Gaussian mixture component number per state: 5, 10, 15, and 20.

Using the same amount of training data, a pentaphone HMM/BN model was trained on each accent data labeled with phoneme class context variables as described in Section 2. The HMM/BN state topology, the total number of states, and the transition probabilities are all identical to the HMM baseline. So in terms of parameter number, they have similar complexity. The main difference is only the probability distribution of states where each Gaussian was explicitly conditioned on C_L or C_R. In contrast, each Gaussian component in HMM state is learned implicitly by the EM algorithm, without any "meaningful" interpretation of its mixture index. During training, there were some phoneme classes context of C_L or C_R which did not exist due to grammatical rules or were unseen in the training data, which after training resulted in about 50 Gaussians per states on average. To avoid unreliable estimated parameters and to compare their performances with the baseline having exactly the same total number of Gaussians, we used data-driven clustering technique and reduced the size of the pentaphone HMM/BN model to correspond to a 5, 10, 15, and 20 mixture component baseline.

Table 2: *Accuracy rates (%) for pentaphone model using BN-a (see Fig. 3(a)) on accented matched test set with different number of mixture components*

Mixture Number	US Accent		AUS Accent	
	Triphn Baseline	**Pentaphn HMM/BN**	**Triphn Baseline**	**Pentaphn HMM/BN**
5 Mix	84.30	85.19	82.33	84.24
10 Mix	84.66	85.91	82.21	84.12
15 Mix	84.78	85.55	83.46	84.18
20 Mix	85.25	85.67	82.63	84.60

Table 3: *Accuracy rates (%) for pentaphone model using BN-a (see Fig. 3(a)) on different accented test set with 15 mixture components*

Accented Test Set	US Accent		AUS Accent	
	Triphn Baseline	**Pentaphn HMM/BN**	**Triphn Baseline**	**Pentaphn HMM/BN**
US Test	84.78	85.55	75.22	76.96
AUS Test	64.78	65.43	83.46	84.18

First, the performance of pentaphone HMM/BN using BN-a

(see Fig. 3(a)) was evaluated on accented matched test set, e.g. the US trained model was only tested on the US test data. The results obtained by the different mixture component numbers are summarized in Table 2. It can be seen that within the same number of parameters, the performance of pentaphone HMM/BN models always performed better than the baseline. The best performance of the US pentaphone HMM/BN is obtained with 10 Gaussian mixtures, which gives about a 8% relative WER reduction, and the best performance of the AUS pentaphone HMM/BN is obtained with 20 Gaussian mixtures, which gives about a 11% WER reduction. We also evaluated the performance of this pentaphone HMM/BN models on different accented test set, e.g. the US trained model was tested on the AUS test data. The results obtained by 15 mixture components are summarized in Table 3. For easy comparison, the accented matched evaluation with the same number of mixture components were also included. The results show that the pentaphone HMM/BN model on accented mismatch condition still consistently improved performance over the standard HMM based triphone model.

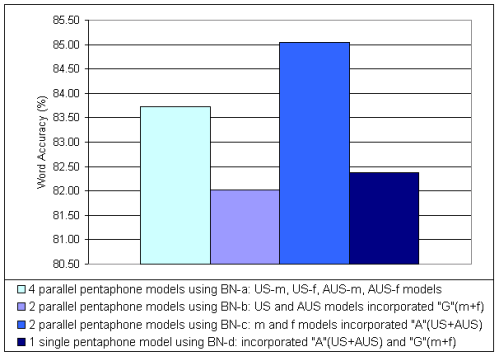

Figure 4: *Comparing recognition word accuracy rates of pentaphone HMM/BN model using different BN topologies (BN-a, BN-b, BN-c, BN-d as in Fig. 3), but having the same 5 mixture components per state on average.*

Next, we evaluated the performance among the pentaphone HMM/BN models using BN-a, BN-b, BN-c or BN-d as described in section 2.2. Having the same 5 mixture components per state on average, the results are shown in Fig. 4. The four bar results in the figure were obtained from: (1) the pentaphone HMM/BN models with BN-a which only incorporated C_L and C_R, so four pentaphone models of US-male, US-female, AUS-male and AUS-female, are used in parallel, (2) the pentaphone HMM/BN models with BN-b which incorporated C_L, C_R, and gender G (male, female), so both US and AUS gender-independent pentaphone models are used in parallel, (3) the pentaphone HMM/BN with BN-c incorporated C_L, C_R and accent A (US, AUS), so both male and female accent-independent pentaphone models are used in parallel, and (4) the pentaphone HMM/BN with BN-d incorporated C_L, C_R, accent A (US,AUS) and gender G (male,female), so only one single accented-gender-independent pentaphone model is used. The performance of gender-dependent model is better than gender-independent model where gender G is incorporated into state PDF. This might be due to high variability with respect to speaker gender, so gender dependent models can learn other variabilities better, and thus resulting a better performance. The best performance achieved 85.05% word accuracy with gender-dependent pentaphone models using BN-c topologies which incorporated additional knowledge of accent A, second preceding context C_L and succeeding context C_R.

5. Conclusion

We presented the possibility of utilizing the HMM/BN modeling framework to incorporate various knowledge sources. This method allows for easy integration of additional information into existing HMM-based triphone acoustic models, where additional knowledge sources are incorporated into the triphone state PDF by means of the BN. Beneficially, we can impose a kind of knowledge-based structure so that the state PDF can be learned more specifically and precisely. For issues of recognition, if we lack appropriate decoding for pentaphone HMM/BN models, we can still use the standard decoding system without modification, while the additional knowledge sources are then assumed hidden, and the state PDF can be calculated by marginalization over those BN joint PDFs. The recognition results indicate that ASR system performance can be improved with the proposed hybrid pentaphone HMM/BN model, even when it has the same number of Gaussians as the baseline triphone HMM. The best performance among pentaphone HMM/BN models was obtained by the model that incorporated additional knowledge of accent A, second preceding context C_L and succeeding context C_R. In future plan, we would like to implement the similar schemes of incorporating additional knowledge using DBN.

6. References

[1] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou, "Audio-visual speech recognition," Tech. Rep., CSLP John Hopkins University, Baltimore, USA, 2000.

[2] A. Ljolje, D. Hindle, M. Riley, and R. Sproat, "The AT&T LVCSR-2000 system," in *Speech Transcription Workshop*, University of Maryland, USA, 2000.

[3] J. Li, Y. Tsao, and C.-H. Lee, "A study on knowledge source integration for candidate rescoring in automatic speech recognition," in *Proc. ICASSP*, Philadelphia, USA, 2005, pp. 837–840.

[4] T. Dean and K. Kanazawa, "Probabilistic temporal reasoning," in *Proc. AAAI*, Minnesota, USA, 1988, pp. 524–528.

[5] G. Zweig and S. Russell, "Probabilistic modeling with Bayesian networks for automatic speech recognition," in *Proc. ICSLP*, Sydney, Australia, 1998, pp. 3010–3013.

[6] K. Daoudi, D. Fohr, and C. Antoine, "A new approach for multi-band speech recognition based on probabilistic graphical models," in *Proc. ICSLP*, Beijing, China, 2000, pp. 329–332.

[7] K. Markov and S. Nakamura, "A hybrid HMM/BN acoustic modeling for automatic speech recognition," *IEICE Trans. Inf. & Syst.*, vol. E86-D, no. 3, pp. 438–445, 2003.

[8] K. Markov and S. Nakamura, "Modeling successive frame dependencies with hybrid HMM/BN acoustic model," in *Proc. ICASSP*, Philadelphia, USA, 2005, pp. 701–704.

[9] S. Sakti, S. Nakamura, and K. Markov, "A hybrid HMM/BN acoustic model utilizing pentaphone-context dependency," *IEICE Trans. Inf. & Syst.*, vol. E89-D, no. 3, pp. 953–961, 2006.

[10] D.B. Paul and J.M. Baker, "The design for the Wall Street journal based CSR corpus," in *Proc. DARPA Workshop*, Pacific Groove, California, USA, 1992, pp. 357–361.

[11] S. Sakti, S. Nakamura, and K. Markov, "Incorporation of pentaphone-context dependency based on hybrid HMM/BN acoustic modeling framework," in *Proc. ICCASP*, Toulouse, France, 2006, p. pp. to appear.

[12] J.J. Odell, *The Use of Context in Large Vocabulary Speech Recognition*, Ph.D. thesis, Cambridge University, Cambridge, UK, 1995.

[13] T. Jitsuhiro, T. Matsui, and S. Nakamura, "Automatic generation of non-uniform HMM topologies based on the MDL criterion," *IEICE Trans. Inf. & Syst.*, vol. E87-D, no. 8, pp. 2121–2129, 2004.