# INCORPORATION OF PENTAPHONE-CONTEXT DEPENDENCY BASED ON HYBRID HMM/BN ACOUSTIC MODELING FRAMEWORK

*Sakriani Sakti, Konstantin Markov, Satoshi Nakamura*

ATR Spoken Language Communication Research Laboratories,
Keihanna Science City, Kyoto, Japan
{sakriani.sakti,konstantin.markov,satoshi.nakamura}@atr.jp

## ABSTRACT

This paper presents a new method of modeling pentaphone-context units using the hybrid HMM/BN acoustic modeling. Rather than modeling pentaphones explicitly, in this approach we extend the modeled phonetic context within the triphone framework, since the probabilistic dependencies between the triphone context unit and the second preceding/following contexts are incorporated into the triphone state output distributions by means of the BN. Another advantage is that we can use a standard decoding system by assuming the next preceding/following context variables hidden during recognition. In this study, the performance of pentaphone HMM/BN model was evaluated with our LVCSR system by phoneme recognition and by large-vocabulary continuous word recognition tasks. In both cases, we observed consistently improved performance over the standard HMM based triphone model with the same number of parameters.

## 1. INTRODUCTION

A triphone, which includes the immediate preceding and following phonetic contexts, is the most widely used acoustic unit in current hidden Markov model (HMM) based large-vocabulary continuous speech recognition (LVCSR) systems. Although such triphones have proven to be an efficient choice, they are considered insufficient for capturing all coarticulation effects. These effects may come not only from the first preceding/following contexts but also from further neighboring contexts. Thus, by incorporating something wider than the triphone context, such as a pentaphone (or more), more than just one preceding and one following phonetic context dependencies are taken into account, which is expected to improve the performance of such acoustic models.

Actually, the idea of using wider-than-triphone units is not novel in automatic speech recognition (ASR) systems. To date, the IBM and AT&T LVCSR systems have quite successfully used pentaphone models [1, 2]. Some researches also have tried to use wide-context models, such syllables or multi-phone units, that give better overall recognition rates [3, 4]. However, there is no common, flexible enough framework that allows integration of additional information of wide-context dependency into existing HMM-based triphone acoustic models. To train wide-context models from scratch and use them properly in cross-word decoding, we may encounter many difficulties due to the increased complexity of the model and computational costs; training data and memory space are also limited. Further difficulties may arise if the available decoding system adheres to a fixed model structure. In [5], it was proposed to compile wide-context-dependent models into a network of Weighted Finite State Transducers (WFST), so the decoding process is completely decoupled from dealing with the wide context. However, when higher order models are used, difficulties lie in the compilation itself. The work in [6] was thus conducted to simplify the compilation method. Another much simpler procedure in LVCSR systems is to apply wide context models in rescoring passes only.

Over the last decade, Bayesian Networks (BN) have become a popular method for encoding uncertainty in artificial intelligence, and recently they have also attracted from attention 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 [7]. In some of the first reports on Dynamic BNs (DBN) in speech recognition [8, 9], 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. Another advantage of BNs is that additional features which are difficult to estimate reliably during recognition may be left hidden, i.e., unobservable.

The approach we propose in this paper is to incorporate wide-context-dependency by utilizing the advantages of BNs, while allowing us to keep the existing: (1) HMM-based triphone acoustic model topology and (2) standard triphone-based decoding system. It is based on the hybrid HMM/BN model [10]. With this method, we can easily extend the conventional triphone HMM to cover a wider context where the probabilistic dependencies between the triphone context unit and the next preceding/following contexts are learned through a BN and the pentaphone state output probability distribution can be modeled. Our standard triphone-based decoding system can still be used without modification, since the next preceding/following context variables are assumed hidden during recognition.

In the next section, we briefly describe the hybrid HMM/BN background, followed by the structure of the hybrid pentaphone HMM/BN model in Section 3 including model topology, training procedure, and recognition issues. Details of the experiments are presented in Section 4, including results and discussions. A conclusion is drawn in Section 5.

## 2. HYBRID HMM/BN BACKGROUND

The HMM/BN model is a combination of an HMM and a BN. Temporal speech characteristics are still governed by HMM state transitions, but HMM state probability distributions are inferred from the BN. This allows for very flexible and consistent models of state probability distribution that can easily integrate different speech parameterizations. A block diagram of the HMM/BN is shown in Fig. 1, with HMM on top and BN underneath.

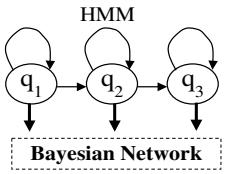



**Fig. 1**. HMM/BN model structure, where HMM transitions model speech temporal characteristics and BN represents state probability distributions.

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

It is also possible to use different BN structures for different sets of HMM states. Fig. 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 the BN structure, it is designed manually based on our knowledge about the data. More details about the HMM/BN approach can be found in [10, 11].

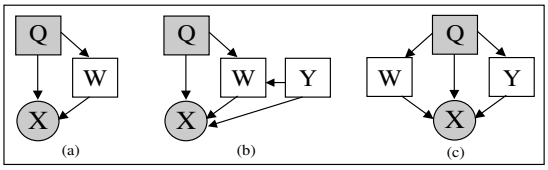



**Fig. 2**. Three simple examples of different BN structures with variables Q,W,Y, and X.

## 3. HYBRID PENTAPHONE HMM/BN ACOUSTIC MODEL

### 3.1. Topology of Pentaphone HMM/BN Model

In our pentaphone HMM/BN model, the HMM at the top level corresponds to the triphone-context acoustic unit and has three states. The BN at the bottom level is used to model the probabilistic dependencies between triphone-context units and the second preceding/following contexts represented by different BN variables. Let $/a^-, a, a^+/$ be a triphone context, then the corresponding pentaphone three-states left-to-right HMM/BN structure becomes the one shown in Fig. 3.

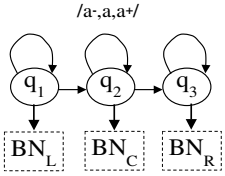



**Fig. 3**. Hybrid pentaphone HMM/BN structure.

If we extend the conventional triphone HMM with additional second preceding and following contexts, we have a pentaphone context like $/a^{--}, a^-, a, a^+, a^{++}/$. The left, center, and right state output probability distributions can be represented by three different BN topologies, as shown in Figs. 4(a), (b), and (c), respectively. $BN_L$ and $BN_R$ have only one additional discrete variable, as in Fig. 2(a), which is the second preceding context $C_L$ (for $BN_L$) and the second following $C_R$ (for $BN_R$). $BN_C$, however, has two additional context variables $C_L$ and $C_R$, which is similar to Fig. 2(c).

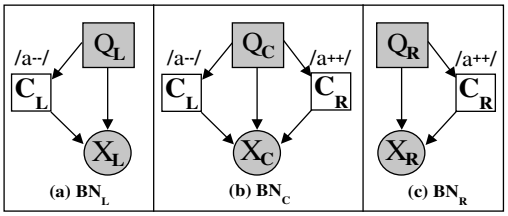



**Fig. 4**. BN topologies of the left state (a), center state (b), and right state (c) of pentaphone-HMM/BN, for modeling a pentaphone context $/a^{--}, a^-, a, a^+, a^{++}/$.

### 3.2. Pentaphone HMM/BN Model Training

The training procedure for the hybrid pentaphone HMM/BN model can be adopted from the general training of the HMM/BN model [10]. It is based on the Viterbi algorithm and consists of the following steps:

1. Initialization: HMM/BN parameter initialization using the bootstrap conventional HMM model.
2. Viterbi alignment: Obtain time aligned state segmentation of the training data.
3. BN training: Train the BN using state labelled training data.
4. Transition probability updating.
5. Embedded BN/HMM training.
6. Convergence check: Stop if convergence criterion is met, otherwise go to step 2.

The training of the state BN at step 3 above is done using standard statistical methods. Since all variables, including triphone state $Q$, second preceding ($C_L$) context, second following ($C_R$) context, and probability distribution $X$ are observable during training, only simple ML parameter estimation can be applied on the training of the state BN at step 3 of the algorithm.

### 3.3. Recognition with a Pentaphone HMM/BN Model

In a conventional HMM, the state PDF is usually represented by Gaussian mixture density, and the state output probability is obtained as:

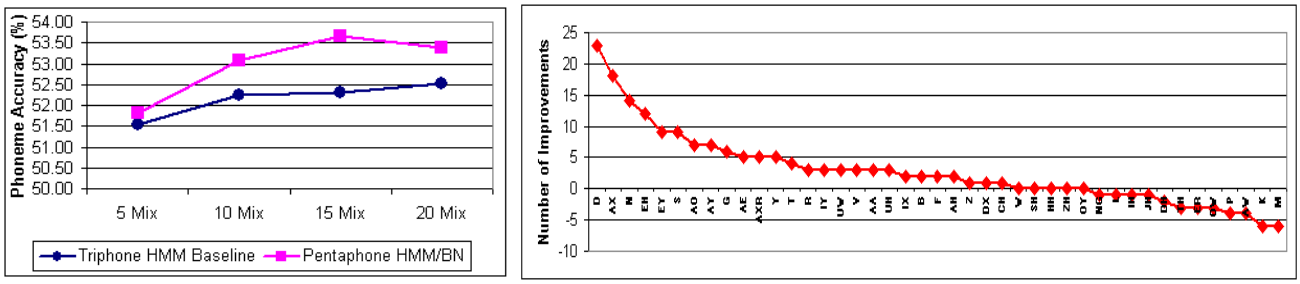
$$P(x_t|q_i) = \sum_{m=1}^{M} b_m \mathcal{N}(x_t; \mu_m, \Sigma_m), \quad (2)$$

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 $\mu_m$ and covariance matrix $\Sigma_m$.

In the case of pentaphone HMM/BN, the state PDF is the BN joint probability model. For the left and right state PDF, the BN join probability model is expressed as:

$$P(X, C, Q) = P(X|C, Q)P(C|Q)P(Q), \quad (3)$$

where it depends on the second preceding or following context $C$. Since $X$ is a continuous variable, $P(X|C, Q)$ is modeled by Gaussian density. The second preceding/following context $C$ is discrete,




(a) (b)

**Fig. 5**. (a) Recognition accuracy rates of phoneme recognition experiments. (b) Number of improvements for each phoneme.

and therefore $P(C|Q)$ is represented by conditional probability table (CPT). When $C$ is observable, the left/right state output probability is simply:

$$P(x_t|c_n, q_i) = P(X = x_t|C = c_n, Q = q_i). \quad (4)$$

However, since the second preceding/following context $C$ ($C_L$ or $C_R$) is assumed hidden during recognition, the left/right state output probability is then calculated by marginalization over $C$:

$$P(x_t|q_i) = \sum_{n=1}^{N} P(c_n|q_i)P(x_t|c_n, q_i), \quad (5)$$

where for simplicity, we use $x_t$, $q_i$, and $c_n$ instead of $\langle X = x_t\rangle$, $\langle Q = q_i\rangle$, and $\langle C = c_n\rangle$, respectively. Analyzing Eq. (5), we can see that it is equivalent to the state output probability of the conventional HMM of Eq. (2) if we treat term $P(c_n|q_i)$ as a mixture weight coefficient for the Gaussian component $P(X|c_n, q_i)$.

For the center state PDF, the BN joint probability model is expressed as:

$$P(X, C_L, C_R, Q) = P(X|C_L, C_R, Q)P(C_L|Q)P(C_R|Q)P(Q), \quad (6)$$

where it depends on both the second preceding context $C_L$ and the 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, the center state output probability is obtained from the $BN_C$ assuming also that both additional variables $C_L$ and $C_R$ are hidden during recognition and take $N_L$ and $N_R$ values:

$$P(x_t|q_i) = \sum_{l=1}^{N_L}\sum_{r=1}^{N_R} P(c_l|q_i)P(c_r|q_i)P(x_t|c_l, c_r, q_i), \quad (7)$$

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. (7) is also 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 the Gaussian component $P(X|c_l, c_r, q_i)$. Using these expressions (Eqs. (5) and (7)), we can perform recognition using existing triphone HMM based decoders without modification.

## 4. EXPERIMENTAL RESULTS AND DISCUSSIONS

Our baseline triphone HMM acoustic model was trained on more than 60 hours of native English speech data from the Wall Street Journal (WSJ0 and WSJ1) speech corpus [12]. 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, $\Delta$ MFCC, and $\Delta$ log power were used as feature parameters. Three states were used as initial HMM for each phoneme. Then shared state HMnet topology was obtained using a successive state splitting (SSS) training algorithm. Since the SSS training algorithm used here was based on the minimum description length (MDL) optimization criterion, the number of shared HMM states is determined automatically by the algorithm. MDL-SSS details can be found in [13]. Here, the length of the HMnet path for each triphone context is restricted to three states. The total number of states is 1,144 with four different versions of Gaussian mixture component numbers per state: 5, 10, 15, and 20.

Using the same database corpus, we obtained time-aligned state segmentation. Then we performed the hybrid pentaphone HMM/BN and trained each $BN_L$, $BN_R$, and $BN_C$ with additional context variables, as described in the previous section. The HMM/BN state topology, the total number of states, and the transition probabilities remain identical to those of the baseline. According to Eqs. (5) and (7), the number of Gaussian components depends on the second preceding/following phonetic context $C$. If we use a 44-phoneme set (including silence) for the English ASR, it means that the total number of Gaussians for each left/right state may reach 44, and the total number of Gaussians for each center state of may reach $44^2$=1,936. 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.

The performance of the models was tested on the ATR Basic Travel Expression Corpus (BTEC)[14]. It consists of travel related expressions, which is quite different from the training corpus. In this study, we randomly selected 200 utterances from 4,080 utterances spoken by 40 different speakers (20 males, 20 females).

First, we evaluated the performance of the pentaphone HMM/BN model in a phoneme recognition task. Recognition results of both pentaphone HMM/BN and the triphone HMM baseline are shown in Fig. 5(a). It can be seen that within the same number of parameters, the performance of pentaphone HM/BN models always performed better than the baseline. To investigate in more detail, we calculated the difference between the number of errors in the baseline results and in the pentaphone HMM/BN results (#error in baseline - #error in pentaphone HMM/BN) for each phoneme. The calculation results, summarized in Fig. 5(b), indicate that most phonemes could get the benefits of incorporating wider context dependencies.

To investigate the consistency of the effect of using pentaphone HMM/BN, we also evaluated the performance of the pentaphone

HMM/BN in LVCSR task. In this case, we use a 20k word dictionary and also a bigram and trigram language model trained on about 150,000 travel related sentences. The word recognition results of both pentaphone HMM/BN and the triphone HMM baseline are shown in Fig. 6. As in the phoneme recognition task, pentaphone HMM/BN outperformed the baseline model in all cases. The best result achieved a 10% relative word error rate (WER) reduction.

However, one might argue that the superior performance of our proposed model is mainly because it has a varied number of mixture components, while the baseline only has a fixed number of mixture components. To investigate this issue, we conducted additional experiments with: (1) the triphone HMM model with a varied number of mixture components per state trained by simply assigning the number of mixture components per state depending on the amount of training data for that state, and (2) the pentaphone HMM/BN with a fixed number of mixture components per state trained by applying data-driven clustering to each state. To minimize the time process, clustering was only performed for Gaussian components of the left and right states. The Gaussian components of the center state were equivalent to the Gaussian components of the center state of the triphone baseline, assuming that the next preceding and following contexts mainly affect the outer states of the model. Their performances were compared with the baseline and the previous pentaphone HMM/BN model with all having about the same 15 mixture components per state, and the results are shown in Fig. 7. The performance of pentaphone HMM/BN with a fixed number is still better than the triphone models with a varied number of mixture components. This shows that by conditioning each Gaussian with additional knowledge of such pentaphone-context dependency, the state PDF becomes more precise, effecting an improvement in performance.

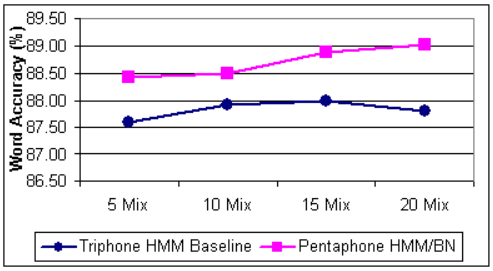



**Fig. 6**. Recognition accuracy rates of the LVCSR experiments.

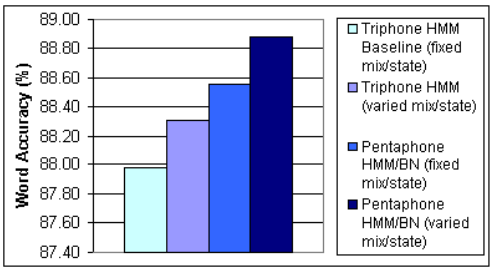



**Fig. 7**. Comparing recognition word accuracy rates of triphone HMM and pentaphone HMM/BN model with a fixed and a varied number of mixture components per state, but having the same 15 mixture components per state on average.

## 5. CONCLUSION

We have presented the possibility of utilizing wide-context dependency which benefits from the HMM/BN modeling framework. This method allows for easy integration of additional information of wide-context dependency into existing HMM-based triphone acoustic models, where the additional knowledge of pentaphone-context dependency is 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. On the 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 second preceding/following context is then assumed hidden, and the state PDF can be calculated by marginalization over those BN joint PDFs. Experimental evaluation in both phoneme recognition and large- vocabulary continuous word recognition tasks, showed that the proposed pentaphone HMM/BN model consistently improved ASR system performance, even when it has the same number of Gaussians as the baseline triphone HMM.

## 6. ACKNOWLEDGEMENTS

Part of this speech recognition research work was supported by the National Institute of Information and Communication Technology (NICT), Japan.

## 7. 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] I. Shafran and M. Ostendorf, "Acoustic model clustering based on syllable structure," *Computer Speech and Language*, vol. 17, no. 4, pp. 311–328, 2003.

[4] R. Messina and D. Jouvet, "Context dependent long unit for speech recognition," in *Proc. ICSLP*, Jeju Island, Korea, 2004, pp. 645–648.

[5] M. Riley, F. Pereira, and M. Mohri, "Transducer composition for context-dependent network expansion," in *Proc. EUROSPEECH*, Rhodos, Greece, 1997, pp. 1427–1430.

[6] M. Schuster and T. Hori, "Efficient generation of high-order context-dependent weighted finite state transducers for speech recognition," in *Proc. ICASSP*, Philadelphia, USA, 2005, pp. 201–204.

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

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

[9] 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.

[10] 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.

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

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

[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.

[14] T. Takezawa, E. Sumita, F. Sugaya, H. Yamamoto, and S. Yamamoto, "Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world," in *Proc. LREC*, Las Palmas, Canary Islands, Spain, 2002, pp. 147–152.