PAPER *Special Section on Statistical Modeling for Speech Processing*

A Hybrid HMM/BN Acoustic Model Utilizing Pentaphone-Context Dependency

Sakriani SAKTI[†a)], **Konstantin MARKOV**[†], ***Nonmembers, and*** **Satoshi NAKAMURA**[†], ***Member***

SUMMARY The most widely used acoustic unit in current automatic speech recognition systems is the triphone, which includes the immediate preceding and following phonetic contexts. Although triphones have proved to be an efficient choice, it is believed that they are insufficient in capturing all of the coarticulation effects. A wider phonetic context seems to be more appropriate, but often suffers from the data sparsity problem and memory constraints. Therefore, an efficient modeling of wider contexts needs to be addressed to achieve a realistic application for an automatic speech recognition system. This paper presents a new method of modeling pentaphone-context units using the hybrid HMM/BN acoustic modeling framework. Rather than modeling pentaphones explicitly, in this approach 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. The advantages of this approach are that we are able to extend the modeled phonetic context within the triphone framework, and we can use a standard decoding system by assuming the next preceding/following context variables hidden during the recognition. To handle the increased parameter number, tying using knowledge-based phoneme classes and a data-driven clustering method is applied. The evaluation experiments indicate that the proposed model outperforms the standard HMM based triphone model, achieving a 9–10% relative word error rate (WER) reduction.

key words: *wide phonetic context model, pentaphone, HMM/BN acoustic model*

1. Introduction

Today's state-of-the-art automatic speech recognition (ASR) systems achieve very good performance in controlled conditions. There are, however, still many challenges to overcome before ASR systems can reach their full potential through widespread use in everyday life. For the best systems, reported results on the 1999 DARPA Broadcast News Benchmark tests showed that error rates on the spontaneous speech portion of the test set were nearly double those of the planned, studio-recorded conditions [1]. There are many factors such as channel effects, speaking style, careless pronunciation, etc., which can cause performance degradation. Experimental results in [2] demonstrated that the style of speech (acoustic variation) is the dominant factor in the recognition error rate. The existing acoustic models (AM) still have a limited capability to handle the coarticulation effects that exist in everyday conversational speech.

Coarticulation is an acoustic and articulatory variability that arises when the articulatory patterns of neighboring speech segments overlap. It is a fundamental part of language sound systems that allows for dynamic transitions between adjacent phoneme segments (both within and across words) that perhaps make speaking easier [3]. As a result, phonemes can have very different waveforms when produced in the context of other phonemes [4].

An acoustic model that can accurately capture these coarticulation effects is obviously needed in large-vocabulary speech recognition systems (LVCSR). The wider the unit models, the better the capturing of the coarticulation effects [5]. Word unit models are impractical for LVCSR systems due to the large amount of training data needed, the large decoding search space, and the inefficiency for expanding the vocabulary system. Syllable-based [6], [7] and multiphone [8] units are smaller than words, both in number and duration, though there are still too many of them and, like words, they lack generality [9]. For example, in the large SWITCHBOARD (SWB) corpus, there are about 9,000 syllables appearing in the training database, but over 8,000 of these have fewer than 100 training tokens [7]. The phonetic units are thus a natural choice since there are only a few of them and their frequency of appearance in the training data is much higher. A standard solution to the coarticulation problem is to extend the phonetic units to include context [10]. Most of the current LVCSR systems use the context-dependent triphone as the fundamental acoustic unit. Context-dependent triphone units have the same structure as context independent phonetic (monophone) units, but are trained on data with immediate preceding and following phonetic context information [9].

Although such triphones have proved to be an efficient choice, it is believed that they are insufficient for capturing all of the coarticulation effects. These effects may come not only from the first preceding/following contexts, but also from further neighboring contexts. In [11], it was found that a vowel may influence not only the preceding consonant but also the vowel before the consonant. Records of /eli/ and /ela/ or /ebi/ and /eba/ showed that the articulatory setting for /e/ was different according to the second vowel in the sequence: the tongue rose higher and nearer to the /i/ in /eli/ and /ebi/ than in tokens in which the last sound constituted an /a/ [12]. Other studies also found that English consonants such as /l/ and /r/ exert long-distance coarticulation effects across syllables, or "resonance" [13], [14]. Thus, by incorporating something wider than the triphone context, more than just one preceding and one following phonetic context is taken into account, which is expected to lead to an im-

Manuscript received July 11, 2005.
Manuscript revised October 14, 2005.

[†] The authors are with ATR Spoken Language Communication Research Laboratories, Kyoto-fu, 619–0288 Japan.
a) E-mail: sakriani.sakti@atr.jp

DOI: 10.1093/ietisy/e89–d.3.954

provement in the performance of such an acoustic model.

Many researchers have tried to improve acoustic models by incorporating a wider-than-triphone context, such as a tetraphone, quinphone/pentaphone, or more [15], [16]. To date, the IBM, Philips/RWTH, and AT&T LVCSR systems have been quite successful in using pentaphone models [17]–[19]. To properly train the model parameters and use them in cross-word decoding, a huge amount of training data and memory space are required. However, such resources are usually not available. If only limited training data is available, context resolution may be lost due to nonrobust parameter estimation and an increased number of unseen contexts. If we also face a memory constraint, the use of the cross-word wide-context model may become cumbersome and sometimes even impossible [20]. For large-scale systems, then, a simple procedure to avoid decoding complexity is to apply the wide-context models in the rescoring pass. In this case, the decoding will use knowledge sources of progressively increasing complexity to decrease the size of the search space [21]. Another possibility is to use only intra-word wide-context units [18]. In [22], 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 [20] was thus conducted in an attempt to simplify the compilation method.

In essence, incorporating wider-than-triphone-context units often leads to additional improvement, but it requires large training data and makes the training and decoding difficult. On the other hand, the simpler model is more reliable but less precise in capturing the coarticualtion effects. Therefore, an efficient modeling of the wide-context unit, which can maintain the balance between the context resolution and training data size, is one important problem that needs to be addressed to achieve the realistic application of an ASR system.

Over the last decade, the Bayesian Network (BN) has become a popular method for encoding uncertainty in artificial intelligence. It has also been found to be very powerful in solving various data analysis problems in areas such as expert systems, decision support systems, and pattern recognition [23]. A BN can readily handle incomplete data sets; it allows one to learn about causal relationships; it is well structured and easy to represent; it facilitates the combination of domain knowledge and data; and lastly, it offers an efficient and principled approach for avoiding over-fitting data [24]. With a BN, since it is possible to associate an arbitrary set of variables with each speech frame or HMM state, it is easy to construct models in which phonetic state information is augmented with other variables [25]. That is why, recently, many researchers in speech recognition use a BN to incorporate additional knowledge, such as articulatory features, sub-band correlation, or speaking style [26]–[29]. Another advantage of a BN is that additional features that are difficult to estimate reliably during recognition may be left hidden, i.e., unobservable.

The approach we propose in this paper is based on the hybrid HMM/BN model [25], which allows us to incorporate a wider-than-triphone context by utilizing the advantages of a BN. The probabilistic dependencies between the triphone context unit and the next preceding/following contexts are learned through a BN, and the wide context state output probability distribution can be modeled. The advantages of this approach are that we are able to extend the modeled phonetic context within the triphone framework, and we can use a standard decoding system by assuming the next preceding/following context variables hidden during the recognition. In this study, it is first assumed that the next preceding and following contexts affect mainly the outer HMM states and we only modify those states' pdfs. Then we try to extend the approach to include the inner states of the triphone HMM model. To improve the robustness of the parameter estimation, the standard approach is to tie some state output probability distributions. In this study, we apply Gaussian tying using both knowledge-based and data-driven clustering techniques.

In the next section, we briefly describe the hybrid HMM/BN background followed by the structure of the hybrid pentaphone HMM/BN model. Parameter reduction with phoneme classes and clustering methods is described in Sect. 4. Details of experiments are presented in Sect. 5, including results and a discussion. A conclusion is drawn in Sect. 6.

2. Hybrid HMM/BN Background

The HMM/BN model is a combination of an HMM and a BN. The temporal characteristics of speech are modeled by the HMM state transitions, while the HMM state probability distributions are represented by the BN. A block diagram of the HMM/BN is shown in Fig. 1, with the HMM on top level and the BN underneath.

This model is described by two sets of probabilities: HMM transition probabilities $P(q_i|q_j)$ and the joint probability distribution of the BN $P(Z_1, \ldots, Z_K)$, where Z_k, $k = 1, \ldots, K$ are the BN variables. The BN joint probability density function (PDF) can be factorized as:

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

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

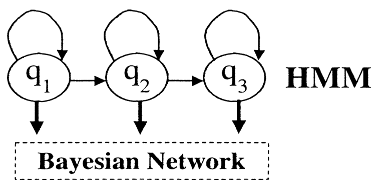

Fig. 1 HMM/BN model structure. HMM transitions model speech temporal characteristics and BN represents state probability distributions.

It is also possible to use different kinds of BN structures for different sets of HMM states. Figure 2 shows a simple example of a BN structure with three variables, where variable Q represents the HMM state, X represents the spectrum observation variable, and Y represents any other additional information, such as pitch, articulatory positions, speaker gender, context information, etc. Here, Q 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 is 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.

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

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

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

In the case of the HMM/BN model, as that of Fig. 2, the state PDF is the BN joint probability model that can be expressed by a chain rule, according to Eq. (1):

$$P(X, Y, Q) = P(X|Y, Q)P(Y|Q)P(Q), \tag{3}$$

thus the state output probability, when all the BN variables are observable, is simply:

$$P(x_t|y_n, q_i) = P(X = x_t|Y = y_n, Q = q_i). \tag{4}$$

However, if the additional variable Y is hidden, then the state output probability is calculated by marginalization over Y:

$$\begin{aligned} P(x_t|q_i) &= \frac{P(x_t, q_i)}{P(q_i)} = \frac{\sum_{n=1}^{N} P(x_t, y_n, q_i)}{P(q_i)} \\ &= \frac{\sum_{n=1}^{N} P(x_t|y_n, q_i)P(y_n|q_i)P(q_i)}{P(q_i)} \\ &= \sum_{n=1}^{N} P(y_n|q_i)P(x_t|y_n, q_i), \end{aligned} \tag{5}$$

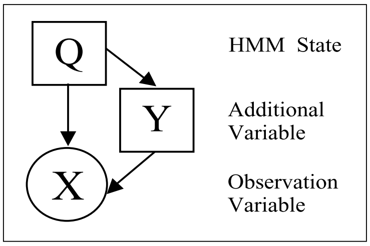

Fig. 2 A simple example of a BN structure with three variables Q, Y, X, where Q represents the HMM state, X represents the spectrum observation variable, and Y represents any additional information.

where for simplicity, we use these x_t, q_i, and y_n notations instead of $\langle X = x_t\rangle$, $\langle Q = q_i\rangle$, and $\langle Y = y_n\rangle$, respectively. Here, we can see that Eq. (5) is equivalent to the state output probability of the conventional HMM of Eq. (2) if we treat the term $P(y_n|q_i)$ as a mixture weight coefficient for the Gaussian component $P(X|y_n, q_i)$. Thus, the existing HMM decoders can work with the HMM/BN model without any modifications.

The training procedure for the hybrid HMM/BN model 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. If all variables are observable during training, only simple ML parameter estimation can be applied; however, if some variables are hidden, then the parameters can be estimated by the standard EM algorithm.

More details about the HMM/BN approach can be found in [25]–[27].

3. Hybrid 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.

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

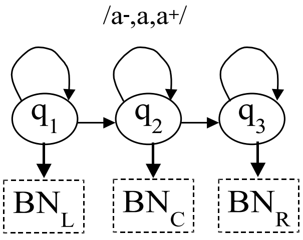

Fig. 3 Hybrid pentaphone HMM/BN structure.

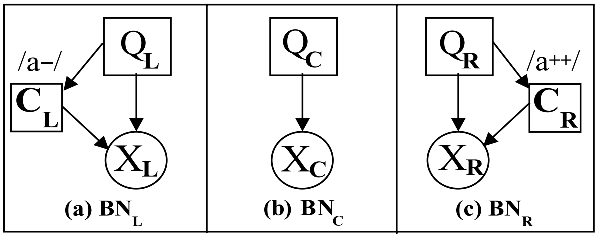

Fig. 4 BN topologies of the left state (a), center state (b), and right state (c) of LR-HMM/BN, for modeling a pentaphone context $/a^{--}, a^{-}, a, a^{+}, a^{++}/$.

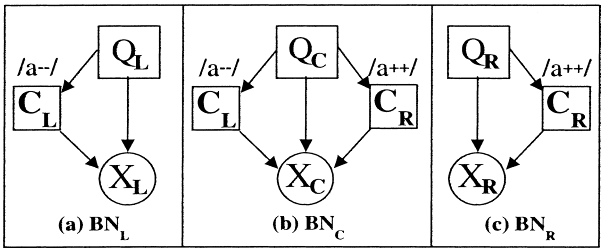

Fig. 5 BN topologies of the left state (a), center state (b), and right state (c) of LRC-HMM/BN, for modeling a pentaphone context $/a^{--}, a^{-}, a, a^{+}, a^{++}/$.

Fig. 4 (a), (b) and (c), respectively. Here, it is first assumed that the next preceding and following contexts mainly affect the outer states of the triphone HMM model, so that only BN_L and BN_R have an additional discrete variable C_L and C_R (as variable Y in the previous section). They are associated with the second preceding and following contexts, respectively. BN_C does not have any additional context variables. Since only the left and right states have additional variables, we call this model LR-HMM/BN.

The state PDF of the pentaphone HMM/BN model is the BN joint probability model, which is expressed as:

$$P(X, C, Q) = P(X|C, Q)P(C|Q)P(Q), \tag{6}$$

where it depends on the second preceding or succeeding context C. When C is observable, the left/right state output probability is simply:

$$P(x_t|c_n, q_i) = P(X = x_t|C = c_n, Q = q_i). \tag{7}$$

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

$$P(x_t|q_i) = \sum_{n=1}^{N} P(c_n|q_i)P(x_t|c_n, q_i), \tag{8}$$

where for simplicity, we use these x_t, q_i, and c_n notations instead of $\langle X = x_t \rangle$, $\langle Q = q_i \rangle$, and $\langle C = c_n \rangle$, respectively. $P(c_n|q_i)$ is the probability that the state q_i has the second preceding/following contexts c_n, and $P(x_t|c_n, q_i)$ is the probability of observation x_t given that we are in the state q_i having the second preceding/following contexts c_n. Here, we can see that Eq. (8) is equivalent to the state output probability of the conventional HMM of Eq. (2) if we treat the term $P(c_n|q_i)$ as a mixture weight coefficient for the Gaussian component $P(X|c_n, q_i)$.

Next, we attempted to incorporate the wide context dependencies into the center state of the triphone HMM model. The state BN topologies for this case are shown in Fig. 5. The BN_L and BN_R are the same as before, while BN_C has two additional context variables: the second preceding (C_L) and the second following (C_R) contexts. Since all states have wide-context variables, we call this model LRC-HMM/BN.

The output probability for the left/right state is obtained as in LR-HMM/BN. Here, 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:

$$P(x_t|q_i) = \sum_{l=1}^{N_L} \sum_{r=1}^{N_R} P(c_l|q_i)P(c_r|q_i)P(x_t|c_l, c_r, q_i), \tag{9}$$

where for simplicity, we use these x_t, q_i, c_l, and c_r notations 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. $P(c_l|Q)P(c_r|q_i)$ are the probabilities that the center state q_i has the second preceding and following contexts (c_l and c_r), and $P(x_t|c_l, c_r, q_i)$ is the probability of observation x_t given that we are in the center state q_i having the second preceding and following contexts, c_l and c_r, respectively. Here, we can see that Eq. (9) is also equivalent to the state output probability of the conventional HMM of Eq. (2) if we treat the 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. (8) and (9)), we can perform recognition using the existing triphone HMM based decoders without any modification.

The training procedure for the hybrid pentaphone HMM/BN model is based on the Viterbi algorithm as described in Sect. 2. Since all variables, including triphone state Q, second preceding (C_L) context, second following (C_R) context and feature variable X, are observable during training, only simple ML parameter estimation is applied on the training of the state BN at step 3 of the algorithm.

4. Parameter Reduction

According to Eqs. (8) and (9), for each value c_n of the second preceding/following phonetic context C, there is a corresponding Gaussian component. An example of observation space modeling by BN_R is shown in Fig. 6. If we use a 44-phoneme set (including silence) for the English ASR, it means that the second preceding/following phonetic context C has 44 possible values ($C = c_1, c_2, \ldots, c_{44}$), thus the total number of Gaussians for each left/right state may become 44, and the total number of Gaussians for each center state of LRC-HMM/BN may become $44^2 = 1{,}936$. If the amount of training data is not enough to obtain a reliable estimate of the increased model parameters, the overall performance may degrade significantly. It is therefore necessary

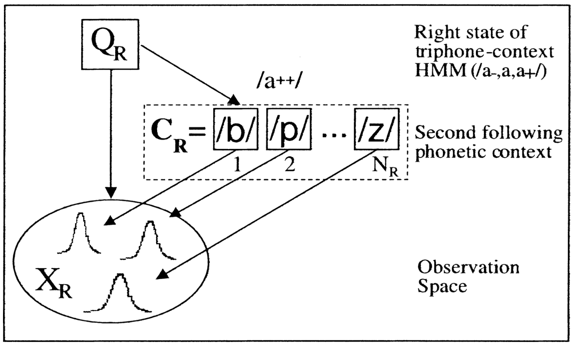
Q_R
/a++/
Right state of triphone-context HMM (/a-,a,a+/)
C_R= /b/ /p/ ··· /z/
1
2
N_R
Second following phonetic context
X_R
Observation Space

Fig. 6 An example of observation space modeling by BN_R, where a different value of C_R corresponds to a different Gaussian.

to reduce the number of Gaussians. Here we attempt both knowledge-based phoneme classes and data-driven clustering techniques.

4.1 Knowledge-Based Phoneme Classes

This is a method where specific knowledge of the unit contexts is explicitly used to guide the classification procedure [30]. Here, we structure the phoneme contexts into a tree based on major distinctions in the manner of articulation. Many phonemes having the same location of articulation tend to have similar effects on the neighboring phonemes. For example, /b/ and /p/ have similar effects on the following vowel, while /n/ and /m/ also have similar effects on the following vowel. The main terminal nodes of the phoneme tree that we used here are: plosives (example: /b/, /p/, /k/, /ch/), nasals (example: /n/, /m/), fricatives (example: /f/, /s/), laterals (example: /l/), trills (example: /r/), and vowels (example: /a/, /i/). Considering also the amount of training data, each of these terminal nodes is divided into more detailed nodes, such as plosive bilabials, plosive velars, and fricative glottals. Based on this tree, we can cluster N (N_L or N_R) second preceding/following contexts into L classes where $L < N$.

4.2 Data-Driven Clustering

Data-driven clustering is also a common approach for parameter tying. Instead of clustering the data based on specific knowledge, they are clustered based on some similarity measure regardless of what phonetic context they represent [23]. Initially, each Gaussian is placed in a separate cluster, then the pair of clusters which would form the smallest resultant cluster when combined are merged. The distance metric is determined by the Euclidean distance between the Gaussian means. This process is repeated until the total number of clusters falls below a certain threshold. With this clustering technique, we can set up any total number of Gaussian components, such that it will correspond to an average of any fixed number of mixture components per state.

5. Experimental Results and Discussion

Our baseline triphone HMM acoustic model is trained on more than 60 hours of native English speech data from the Wall Street Journal (WSJ0 and WSJ1) speech corpus [31]. 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 the initial model for each phoneme. Then, a state level HMnet is obtained using a successive state splitting (SSS) algorithm based on the minimum description length (MDL) criterion in order to gain the optimal state topology in which triphone contexts are shared and tied at the state level. Details about MDL-SSS can be found in [32]. Here, the length of the HMnet path for each triphone context is kept 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. Each Gaussian distribution has diagonal-covariance matrix.

The performance of the models was tested on the ATR Basic Travel Expression Corpus (BTEC) [33], 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). The best baseline HMM performance is 87.98% word accuracy, obtained by a triphone HMM with 15 Gaussians per state.

Using the same database corpus, we obtained time-aligned state segmentation. First, we evaluated the hybrid pentaphone LR-HMM/BN and trained the BN_L/BN_R with second preceding/following contexts as additional discrete variables. The center state BN_C was equivalent to the standard HMM state PDF modeled as a mixture of Gaussians. Thus, as a center state of the HMM/BN model, we used the corresponding five component mixture states from the baseline acoustic model. The HMM/BN state topology, the total number of states, and the transition probabilities are all the same as those of the baseline. The initial HMM/BN model used a 44-phoneme context set for C ($C = c_1, c_2, \ldots, c_{44}$). During training, there were some phoneme contexts c_n which did not exist due to grammatical rules or were unseen in the training data, which after training resulted in about 30 Gaussians on average per left/right state. Since the center-state parameters remain the same as the baseline triphone 5-mixture-component HMM, the final hybrid LR-HMM/BN model has about 24 mixtures per state (on average). Then, as described in Sect. 4, using the knowledge-based phoneme clustering, we reduced the 44-phoneme set into 30, 20, 10, and 6 classes. Keeping the center state with five Gaussians per state resulted in hybrid LR-HMM/BN models with 18, 13, 8, and 5 component mixtures on average, respectively. The results of the pentaphone LR-HMM/BN with different kinds of phoneme class sets are shown in Fig. 7. For comparison, we include the HMM triphone baseline with the 15 component mixtures that performed the best.

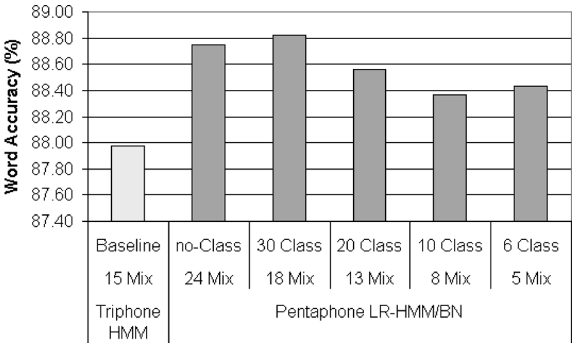

Fig. 7 Recognition accuracy rates of pentaphone LR-HMM/BN using knowledge-based second preceding and following context clustering.

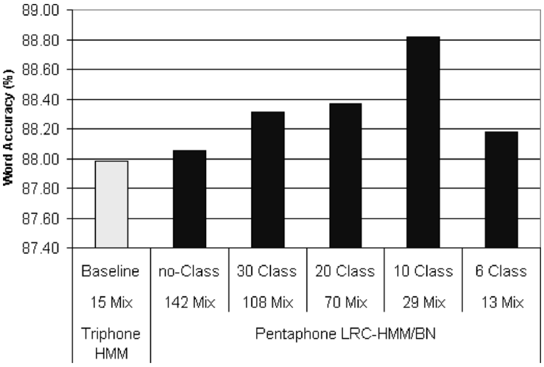

Fig. 8 Recognition accuracy rates of pentaphone LRC-HMM/BN using knowledge-based second preceding and following context clustering.

Next, we also evaluated the hybrid pentaphone LRC-HMM/BN model and trained the BN_C with both second preceding and following contexts as additional discrete variables. The left and right state (BN_L and BN_R) are the same as the hybrid pentaphone LR-HMM/BN. The HMM/BN state topology, the total number of states, and the transition probability are all also the same as those of the baseline. The initial HMM/BN model used a 44-phoneme context set for C ($C = c_1, c_2, \ldots, c_{44}$). During training, there were some phoneme contexts c_n that did not exist due to grammatical rules or were unseen in the training data, which after training resulted in about 412 Gaussians on average per center state and 30 Gaussians on average per left/right state. The average for the final hybrid pentaphone LRC-HMM/BN model was about 142 mixtures per state. To reduce the number of Gaussians, we clustered the 44-phoneme-context set into 30, 20, 10, and 6 classes using knowledge-based phoneme clustering. As a result, the hybrid pentaphone LRC-HMM/BN models had 108, 70, 29, and 13 component mixtures, respectively. The results of the pentaphone LRC-HMM/BN with different kinds of phoneme class sets are shown in Fig. 8.

By only changing the probability distribution of states to incorporate a wider phonetic context through BN (and keeping the other parameters the same), we obtained improved recognition performance. The pentaphone LR-HMM/BN with 30 classes is the best, and further reducing the number of parameters degrades the performance. Nevertheless, the worst performance is still better than the baseline. The pentaphone LRC-HMM/BN with a 44-phoneme set (142 mixtures per state) performed only slightly better than the HMM baseline due to the huge number of parameters. By reducing the number of Gaussians, the resulting performance can be improved from 88.05% to 88.82%. This best performance of the pentaphone LRC-HMM/BN is obtained with 10 classes (29 Gaussians per state). For the optimal size of C_L and C_R using the knowledge-based phoneme clustering, both LRC-HMM/BN and LR-HMM/BN models achieved similar performance.

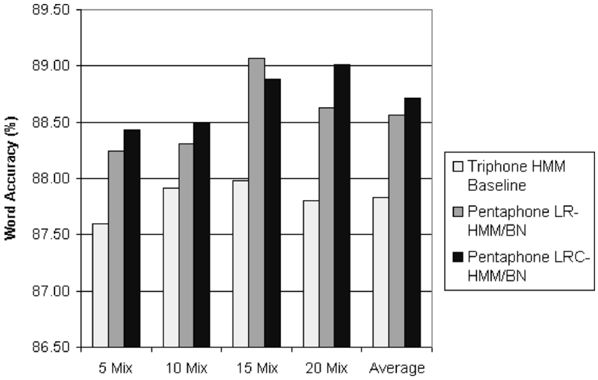

Fig. 9 Recognition accuracy rates of pentaphone LR-HMM/BN and LRC-HMM/BN using data-driven Gaussian clustering.

To be able to compare the pentaphone HMM/BN model and the baseline having exactly the same total number of Gaussians, using data driven clustering we reduced the size of the initial HMM/BN model to correspond to a 5-, 10-, 15-, and 20-mixture component baseline. The center state of the pentaphone LR-HMM/BN also had the corresponding mixture component size. The results of the triphone HMM baseline, the pentaphone LR-HMM/BN, and the pentaphone LRC-HMM/BN are shown in Fig. 9.

It can be seen that within the same number of parameters, both types of pentaphone HMM/BN outperformed the baseline. The best performance of the pentaphone LR-HMM/BN is obtained with 15 Gaussian mixtures, which gives about a 9% relative word error rate (WER) reduction, while the best performance of the pentaphone LRC-HMM/BN is obtained with 20 Gaussian mixtures, which gives about a 10% relative word error rate (WER) reduction. On average, both the LRC-HMM/BN and LR-HMM/BN models also achieved similar performance as before, indicating that both knowledge-based and data-driven clustering techniques are equally efficient in reducing the number of Gaussian components.

Previously, there were experimental results by another researchers showing that a model with a varied number of mixture components often outperforms a model with a fixed number of mixture components, where both models have almost the same total number of Gaussians [34]. To investigate whether the superior performance of our proposed models is mainly not due to that reason, we conducted ad-

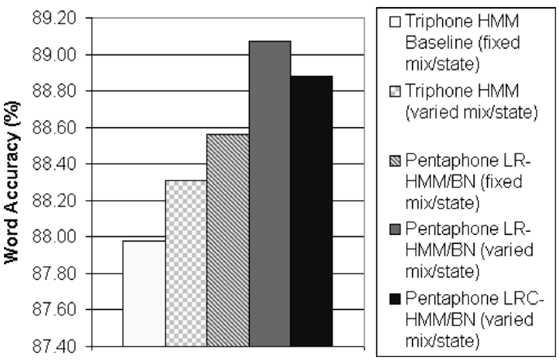

Fig. 10 Comparing recognition accuracy rates of triphone HMM and pentaphone HMM/BN models with a fixed and a varied number of mixture components per state, but having the same 15 mixture components per state on average.

ditional experiments with the triphone HMM model with a varied number of mixture components per state that is trained by simply assigning the number of mixture components per state depending on the amount of training data for that state, and the LR-HMM/BN with a fixed number of mixture components per state trained by applying data-driven clustering for each state. With both having about the same 15 mixture components per state, their performances were compared with the baseline and the previous pentaphone HMM/BN models, and the results of which are shown in Fig. 10. The performance of the LR-HMM/BN with a fixed number is still better than the triphone models with a varied number of mixture components. This indicates that the coarticulation variability is higher than most of the other variability factors. Thus, by explicitly conditioning each Gaussian on such pentaphone-context dependency, instead of just implicitly learning it by the EM algorithm, we can better model the overall PDF, effecting an improvement in performance.

6. Conclusion

We have demonstrated the possibility and benefits of utilizing the wide-context dependency based on the HMM/BN acoustic modeling framework. With this method, we can easily extend the conventional triphone model to cover a wider context where the additional knowledge of pentaphone-context dependency is incorporated into 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 issue of recognition, if we lack an appropriate decoding for the pentaphone HMM/BN model, we still can use the standard decoding system without any modification, while the second preceding/following context is then assumed hidden and the state output probability calculation can be reduced to that of a Gaussian mixture. 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.

Acknowledgement

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

References

[1] D. Pallett, J. Fiscuss, J. Garofolo, A. Martin, and M. Przybocki, "1998 broadcast news benchmark test results: English and non-English word error rate performance measures," Proc. DARPA Broadcast News Workshop, pp.5–12, Virginia, USA, 1999.

[2] M. Weintraub, K. Taussig, K. Hunicke-Smith, and A. Snodgrass, "Effect of speaking style on LVCSR performance," Proc. ICSLP, pp.16–19, Philadelphia, USA, 1996.

[3] R. Scarborough, Coarticulation and the Structure of the Lexicon, PhD Dissertation in Linguistics, University of California at Los Angeles (UCLA), USA, 2004.

[4] L. Rabiner and B.H. Juang, Fundamentals of Speech Recognition, Prentice Hall, New Jersey, USA, 1993.

[5] T. Pfau, M. Beham, W. Reichl, and G. Ruske, "Creating large subword units for speech recognition," Proc. EUROSPEECH, pp.1191–1194, Rhodos, Greece, 1997.

[6] I. Shafran and M. Ostendorf, "Acoustic model clustering based on syllable structure," Comput. Speech Lang., vol.17, no.4, pp.311–328, 2003.

[7] A. Ganapathiraju, J. Hamaker, J. Picone, M. Ordowski, and G. Doddington, "Syllable-based large-vocabulary continuous speech recognition," IEEE Trans. Speech Audio Process., vol.9, no.4, pp.358–366, 2001.

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

[9] P. O'Neill, S. Vaseghi, B. Doherty, W. Tan, and P. McCourt, "Multi-phone strings as subword units for speech recognition," Proc. ICSLP, pp.2523–2526, Sydney, Australia, 1998.

[10] E. Smith, S. Marian, and M. Javier, "Computer recognition of facial actions: A study of co-articulation effects," Proc. 8th Symposium of Neural Computation, California, USA, 2001.

[11] E. Scripture, The Elements of Experimental Phonetics, Charles Scribners Sons, New York, USA, 1902.

[12] B. Kuehner and F. Nolan, "The origin of coarticulation," in Coarticulation: Theory, Data, Techniques, ed. W. Hardcastle and N. Hawlett, pp.7–30, Cambridge University Press, Cambridge, UK, 1999.

[13] S. Heid and S. Hawkins, "An acoustical study of long-domain /r/ and /l/ coarticulation," 5th Seminar on Speech Production: Model and Data, pp.77–80, Kloster Seeon, Germany, 2000.

[14] P. West, "Long distance coarticulatory effects of British English /l/ and /r/: and EMA, EPG and acoustic study," 5th Seminar on Speech Production: Model and Data, pp.105–108, Kloster Seeon, Germany, 2000.

[15] M. Finke and I. Rogina, "Wide-context acoustic modeling in read vs. spontaneous speech," Proc. ICASSP, pp.1743–1746, Munich, Germany, 1997.

[16] L. Bahl, P. de Souza, P. Gopalakrishnan, D. Nahamoo, and M. Picheny, "Decision tree for phonological rules in continuous speech," Proc. ICASSP, pp.185–188, Toronto, Canada, 1991.

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

[18] P. Beyerlein, X. Aubert, R. Haeb-Umbach, M. Harris, D. Klakow, A. Wandemuth, S. Molau, M. Pitz, and A. Sixtus, "The Philips/RWTH system for transcription of broadcast news," Proc. DARPA Broadcast News Workshop, pp.151–155, Virginia, USA, 1999.

[19] A. Ljolje, D. Hindle, M. Riley, and R. Sproat, "The AT&T LVCSR-2000 system," Speech Transcription Workshop, University of Maryland, USA, 2000. http://www.nist.gov/speech/publications/tw00/pdf/cts30.pdf

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

[21] T. Hori, Y. Noda, and S. Matsunaga, "Improved phoneme-history-dependent search method for large-vocabulary continuous-speech recognition," IEICE Trans. Inf. & Syst., vol.E86-D, no.6, pp.1059–1067, June 2003.

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

[23] N. Friedman and M. Goldszmidt, "Learning Bayesian network from data," Technical Report, SRI International, http://www/dsv.su.se/ijcai-99/tutorials/d3.html, 1998.

[24] D. Heckerman, "A tutorial on learning with Bayesian networks," Technical Report, MSR-TR-95-06, Microsoft Research, March 1995.

[25] 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, March 2003.

[26] K. Markov, J. Dang, Y. Lizuka, and S. Nakamura, "Hybrid HMM/BN ASR system integrating spectrum and articulatory features," Proc. EUROSPEECH, pp.965–968, Geneva, Switzerland, 2003.

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

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

[29] T. Stephenson, M. Mathew, and H. Bourland, "Modeling auxiliary information in Bayesian network based ASR," Proc. EUROSPEECH, pp.2765–2768, Aalborg, Denmark, 2001.

[30] X. Huang, A. Acero, and H.W. Hon, Spoken Language Processing, Prentice Hall, New Jersey, USA, 2001.

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

[32] 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, Aug. 2004.

[33] 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," Proc. LREC, pp.147–152, Las Palmas, Canary Islands, Spain, 2002.

[34] V. Valtchev, J. Odell, P.C. Woodland, and S. Young, "MMIE training of large vocabulary speech recognition systems," Speech Commun., vol.22, pp.303–314, 1997.

Sakriani Sakti received her B.E degree in Informatics (cum laude) from Bandung Institute of Technology, Indonesia, in 1999. In 2000, she received an MSc scholarship award from the "DAAD-Siemens Program Asia 21st Century", to study in Communication Technology, University of Ulm, Germany, and received her MSc degree in 2002. During her thesis work, she was an intern student at Speech Understanding Department, DaimlerChrysler Research Center, Ulm, Germany. Since 2003, she has been an engineer at ATR Spoken Language Communication Laboratories, Japan. Her research interests include speech recognition and statistical pattern recognition.

Konstantin Markov was born in Sofia, Bulgaria. After graduating with honors from the St. Petersburg Technical University, he worked for several years as a research engineer at the Communication Industry Research Institute in Sofia. He received his M.Sc. and Ph.D. degrees in electrical engineering from Toyohashi University of Technology, Japan, in 1996 and 1999, respectively. In 1998, he received the Best Student Paper Award from the IEICE Society. In 1999, he joined the research development department of ATR, Japan, and in 2000 became an invited researcher at the ATR Spoken Language Communication (SLC) Research Laboratories. Currently, he is a senior research scientist at the Acoustics and Speech Processing Department of ATR SLC. He is a member of ASJ, IEEE and ISCA. His research interests include signal processing, automatic speech recognition, Bayesian networks and statistical pattern recognition.

Satoshi Nakamura received his B.S. degree in electronic engineering from Kyoto Institute of Technology in 1981 and a Ph.D. degree in information science from Kyoto University in 1992. Between 1981–1993, he worked with the Central Research Laboratory, Sharp Corporation, Nara, Japan. From 1986–1989, he worked with ATR Interpreting Telephony Research Laboratories, and from 1994–2000, he was an associate professor of the Graduate School of Information Science, Nara Institute of Science and Technology, Japan. In 1996, he was a visiting research professor of the CAIP Center of Rutgers University in New Jersey, USA. He is currently director of the ATR Spoken Language Communication Laboratories, Japan. He has also served as an honorary professor at the University of Karlsruhe, Germany, since 2004. His current research interests include speech recognition, speech translation, spoken dialogue systems, stochastic modeling of speech, and microphone arrays. He received the Awaya Award from the Acoustical Society of Japan in 1992, and the Interaction2001 Best Paper Award from the Information Processing Society of Japan in 2001. He served as an associate editor for the Journal of IEICE Information in 2000–2002. He was a member of the Speech Technical Committee of the IEEE Signal Processing Society in 2001–2004. He is a member of the Acoustical Society of Japan (ASJ), Information Processing Society of Japan (IPSJ), and IEEE.