Incorporating Knowledge Sources into a Statistical Acoustic Model for Spoken Language Communication Systems

Sakriani Sakti, Konstantin Markov, *Member*, *IEEE*, and Satoshi Nakamura, *Member*, *IEEE*

Abstract—This paper introduces a general framework for incorporating additional sources of knowledge into an HMM-based statistical acoustic model. Since the knowledge sources are often derived from different domains, it may be difficult to formulate a probabilistic function of the model without learning the causal dependencies between the sources. We utilized a Bayesian network framework to solve this problem. The advantages of this graphical model framework are 1) it allows the probabilistic relationship between information sources to be learned and 2) it facilitates the decomposition of the joint probability density function (PDF) into a linked set of local conditional PDFs. This way, a simplified form of the model can be constructed and reliably estimated using a limited amount of training data. We applied this framework to the problem of incorporating wide-phonetic knowledge information, which often suffers from a sparsity of data and memory constraints. We evaluated how well the proposed method performed on an large-vocabulary continuous speech recognition (LVCSR) task using English speech data that contained two different types of accents. The experimental results revealed that it improved the word accuracy with respect to standard HMM, with or without additional sources of knowledge.

Index Terms—Acoustic modeling, knowledge incorporation, Bayesian network, junction tree, wide-context dependency.

1 Introduction

The growth of information technology has been continuous and is having an ever-larger impact on many aspects of our daily lives. The matter of communication through speech between human beings and information-processing machines such as dialog systems has also become increasingly important [1]. One of the fundamental technologies for achieving a speech-oriented interface is automatic speech recognition (ASR). Many researchers have worked in the area of ASR for almost four decades. The goal is to develop an intelligent machine that can automatically recognize naturally spoken words uttered by humans. However, extracting the underlying linguistic message from a complex acoustic signal is not an easy task due to many sources of variability contained in the signal [2].

Several approaches have been developed to address the problem and these approaches to ASR can generally be classified into two main types, that is, "knowledge-based" and "corpus-based." The former is mainly based on the human ability to interpret spectrograms or other visual representations of the speech signal using knowledge-based rules [3], [4], [5]. However, as there are problems in that, it is difficult to envisage all of the ways in which these rules are interdependent; some rules inevitably compete with others to explain the same phenomenon while others are in direct contradiction [6]. In contrast, the latter approach is usually based on modeling the speech signal using well-defined statistical algorithms that can automatically extract knowledge from the data. This modeling approach has achieved encouraging results and has outperformed the previous knowledge-based approach. This is why most current ASR systems usually use statistical data-driven methods based on hidden Markov models (HMMs). Today's state-of-the-art ASR systems reach very good performance, under controlled conditions.

Despite significant progress in this field, there are still many challenges to overcome before ASR systems can reach their full potential through widespread use in everyday life. For instance, in the presence of unexpected acoustic variability, ASR systems often perform much worse than human listeners [7], [8], [9]. Only a limited level of success can be achieved by only relying on statistical models and mostly ignoring additional knowledge that is available. As many researchers are aware of this problem, various attempts to integrate knowledge-based and statistical approaches more explicitly have been made.

To date, Li et al. [10] have proposed research that enables sources of acoustic phonetic knowledge to be incorporated using neural networks for rescoring purposes. IBM's and AT&T's large-vocabulary continuous speech recognition (LVCSR) systems have also successfully improved acoustic models (AMs) by incorporating the coarticulation effects of longer spans, such as quinphone/pentaphone models [11], [12]. Some researchers have recently attempted to utilize graphical tools such as Bayesian networks (BNs), which can be regarded as a generalization of HMMs, where, in addition to speech spectral information, additional knowledge such as articulatory features, subband correlation, or speaking styles can be easily incorporated [13], [14].

- *The authors are with the NICT/ATR Spoken Language Communication Research Labs., 2-2-2 Hikaridai, Keihanna Science City, Japan 619-0288. E-mail: {sakriani.sakti, konstantin.markov, satoshi.nakamura}@atr.jp.*

Manuscript received 12 Sept. 2006; revised 5 Feb. 2007; accepted 16 Apr. 2007; published online 30 Apr. 2007.
Recommended for acceptance by R.C. Guido, L. Deng, and S. Makino.
For information on obtaining reprints of this article, please send e-mail to: tc@computer.org, and reference IEEECS Log Number TCSI-0362-0906.

0018-9340/07/$25.00 © 2007 IEEE Published by the IEEE Computer Society

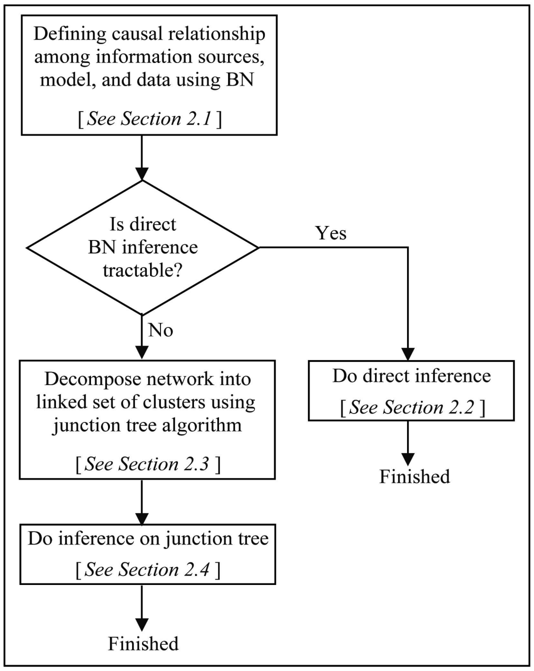

Defining causal relationship among information sources, model, and data using BN
[See Section 2.1]
Is direct BN inference tractable?
Yes
No
Decompose network into linked set of clusters using junction tree algorithm
[See Section 2.3]
Do direct inference
[See Section 2.2]
Finished
Do inference on junction tree
[See Section 2.4]
Finished

Fig. 1. General procedure for incorporating additional knowledge sources.

However, there have often been cases when developing such complex models and achieving optimal performance have not been feasible. This is especially true when there are insufficient resources, that is, the amount of training data and memory space available, to properly train the model parameters. As a result, input space resolution may be lost due to nonrobust estimates and the increased number of unseen patterns. Moreover, decoding with large models may also become cumbersome and sometimes even impossible. The best we can do is to choose a simplified form of the model that can be reliably estimated using the training data available.

In this paper, we propose a method of incorporating additional sources of knowledge in a unified way. We utilized a BN framework to easily integrate any additional knowledge source from any domain. The advantages of this graphical model framework are 1) it allows the probabilistic relationship between information sources to be learned and 2) it facilitates the decomposition of the joint probability density function (PDF) into a linked set of local conditional PDFs. A simplified form of the model can be constructed and reliably estimated using a limited amount of training data in this way.

This framework represents a general approach, meaning that it can be applied to many existing acoustic modeling problems with their respective model-based likelihood functions. Here, we discuss our application of the proposed framework to the problem of incorporating wide-phonetic knowledge information that often suffers from a sparsity of data and memory constraints. We first show how the additional sources of knowledge are incorporated in the HMM state distribution. We then show how the additional sources of knowledge are incorporated in HMM phonetic modeling. Both approaches are experimentally verified in an LVCSR task using English speech data that contains two accents.

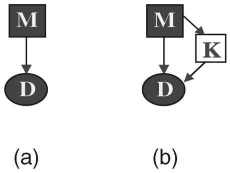


Fig. 2. (a) A BN topology that describes the conditional relationship between data D and the model M. (b) A BN topology that describes the conditional relationship between D, M and an additional knowledge K.

We first describe the general framework for incorporating additional knowledge sources in Section 2. We then briefly describe the conventional HMM AM in Section 3. Sections 4 and 5 show how this framework is used to incorporate additional sources of knowledge at the HMM state and phonetic-model levels, including application to the problem of incorporating wide-phonetic context information. The details on the experiments are presented in Section 6, including the results and a discussion. Finally, conclusions are drawn in Section 7.

2 General Framework for Incorporating Knowledge Sources

In the statistical corpus-based approach, given some observation data D, we train a model M. One key problem of interest is to compute the likelihood, $P(D|M)$, which predicts the data that can be expected given current knowledge about the model.

We can model the PDF, $P(D|M)$, in simple cases by using conditional probability tables (CPT) (if D is discrete) or continuous functions such as Gaussian densities (if D is continuous); the output probability for given data d and model parameter m is then simply calculated as

$$p(d|m) = P(D = d|M = m). \tag{1}$$

Then, assume that we want to incorporate additional knowledge sources into the model. The procedure consists of several steps, as outlined in Fig. 1.

2.1 Defining Causal Relationships between Information Sources

Let us start from a simple case, where the causal relationship between D and M is described using BN, like the one outlined in Fig. 2a, where we have assumed that M is a discrete variable denoted by the square node and D is a continuous variable denoted by the oval node.

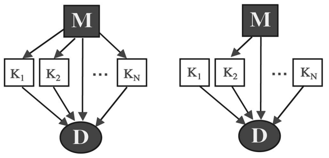
The BN joint probability function can be factorized [15] as

$$P(Z_1, Z_2, \ldots, Z_K) = \prod_{k=1}^{K} P(Z_k|Pa(Z_k)), \tag{2}$$

where $Pa(Z_k)$ denotes the parents of BN variable Z_k so that we obtain

$$P(D, M) = P(D|M)P(M) \tag{3}$$

from Fig. 2a. We thus simply define the conditional relationship between D, M, and K based on our knowledge



(a) (b)

Fig. 3. Several examples of BN topologies that describe the conditional relationship between data D, model M, and several knowledge sources $K_1, K_2, \ldots, K_N$.

of the data to incorporate additional knowledge K into $P(D, M)$ and express the joint probability model in a similar way. The conditional relationship between D, M, and K, for example, can be described by the BN outlined in Fig. 2b and the BN joint probability function becomes

$$P(D, K, M) = P(D|K, M)P(K|M)P(M). \quad (4)$$

Now, let us consider a more detailed example, where we have assumed that there are $K_1, K_2, \ldots, K_N$ knowledge sources. Here, we have assumed that they are all conditionally independent. Fig. 3 outlines two examples of conditional relationship structures between D, M, and $K_1, K_2, \ldots, K_N$. Then, the joint PDF becomes

$$\begin{aligned} &P(D, K_1, \ldots, K_N, M) \\ &= P(D|K_1, \ldots, K_N, M)P(K_1|M) \ldots P(K_N|M)P(M) \end{aligned} \quad (5)$$

for the BN in Fig. 3a, according to (2). If there are some K_i that receive no causal impact from M, as outlined in Fig. 3b (see K_1 and K_N), then the joint probability function becomes

$$\begin{aligned} &P(D, K_1, \ldots, K_N, M) \\ &= P(D|K_1, \ldots, K_N, M)P(K_1)P(K_2|M) \ldots P(K_N)P(M). \end{aligned} \quad (6)$$

As can be seen, different conditional independence assumptions can lead to different probability function decompositions (see (5) and (6)).

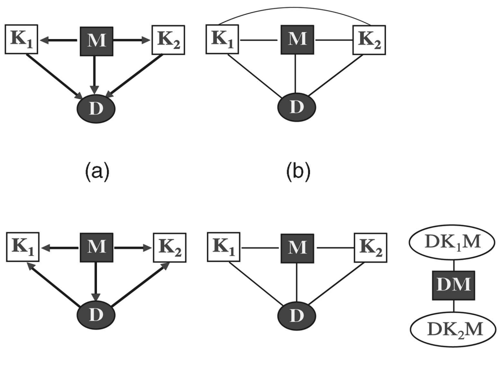
2.2 Direct Inference on BN

Our primary interest during inference is to calculate the global conditional probability, $P(D|K_1, \ldots, K_N, M)$. If the form this PDF takes allows direct calculation, the following two cases can be considered:

1. **All variables can be observed.** In this case, it can simply be calculated as in (1)

$$\begin{aligned} &p(d|k_{1_j}, \ldots, k_{N_j}, m) \\ &= P(D = d|K_1 = k_{1_j}, \ldots, K_N = k_{N_j}, M = m). \end{aligned} \quad (7)$$

2. **Some variables, such as the additional knowledge sources, $K_1, \ldots, K_N$, cannot be observed or are hidden.** In this case, the calculation is done using (5) and by marginalization over all possible K_i : $k_{i_1}, k_{i_2}, \ldots, k_{i_M}$ for all K_i:



(a) (b)

(c) (d) (e)

Fig. 4. (a) BN topology describing the conditional relationship between D, M, K_1, and K_2. (b) Moral and triangulated graph in Fig. 4a. (c) Equivalent BN topology. (d) Moral and triangulated graph in Fig. 4c. (e) Junction tree in Fig. 4d.

$$\begin{aligned} p(d|m) &= \frac{p(d, m)}{p(m)} \\ &= \frac{\sum_{1_j=1}^{M_1} \ldots \sum_{N_j=1}^{M_N} p(d, k_{1_j}, \ldots, k_{N_j}, m)}{p(m)} \\ &= \sum_{1_j=1}^{M_1} \ldots \sum_{N_j=1}^{M_N} p(d|k_{1_j}, \ldots, k_{N_j}, m)p(k_{1_j}|m) \ldots p(k_{N_j}|m), \end{aligned} \quad (8)$$

where, for simplicity, we have used d, m, and k_{i_j} instead of $\langle D = d \rangle$, $\langle M = m \rangle$, and $\langle K_i = k_{i_j} \rangle$.

However, the calculation of global conditional probability $P(D|K_1, \ldots, K_N, M)$ is occasionally not trivial due to too many variables and/or computational complexity. In such cases, directed graphs need to be decomposed into clusters of variables on which the relevant computations can be carried out. This can be done with the junction tree algorithm [15], which will be briefly described in Section 2.3.

2.3 Junction Tree Decomposition

Let us consider a simple case where we only incorporate two additional knowledge sources, K_1 and K_2. The causal relationship between D, M, K_1, and K_2 is described by the BN in Fig. 4a. Here, M, K_1, and K_2 are discrete variables denoted by the square nodes and D is a continuous variable denoted by the oval node.

The following graphical transformations are then applied to obtain a junction tree [15], [16]:

1. Construct an undirected graph from BN by marrying the parents (adding a link between any pair of variables with a common child) and dropping the direction of the links. The resulting graph is called a **moral graph**.
2. Selectively add arcs to the moral graph to form a **triangulated graph**.
3. Form a subset containing $Pa(A) \bigcup A$, which is called a **cluster/clique**, for all variables A with $Pa(A) \neq 0$ in the triangulated graph.
4. Build a **junction tree**, starting with clusters/cliques as nodes, in which each link between two cliques is

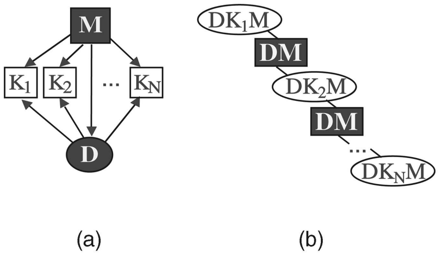


Fig. 5. (a) Equivalent BN topology of the BN shown in Fig. 3a. (b) Corresponding Junction tree.

labeled by using the **separator** of a nonempty intersection between these cliques.

Fig. 4b outlines a moral and triangulated version of BN in Fig. 4a. However, we can only obtain one cluster/clique with the full set of variables $\{D, M, K_1, K_2\}$ from this triangulated graph and cannot decompose any further. Fortunately, since K_1 and K_2 are assumed to be independent, we can obtain an equivalent graph, as in Fig. 4c, by reversing some arrows. Fig. 4d outlines the moral and triangulated version of this graph. We can then identify the clusters/cliques and obtain the junction tree outlined in Fig. 4e, where the cluster sets are represented by the oval nodes, and the separator sets are represented by the square nodes.

The joint probability distribution is then defined as the product of all cluster potentials divided by the product of the separator potentials [16] as

$$P(U) = \frac{\prod_i \phi_{C_i}}{\prod_j \phi_{S_i}}, \tag{9}$$

where U is the "universe" representing all the variables in the graph, ϕ_{C_i} is the cluster potential (the probability over cluster C_i), and ϕ_{S_i} is the separator potential (the probability over separator S_i). Thus, the joint probability function, $P(D, K_1, K_2, M)$, becomes

$$P(D, K_1, K_2, M) = \frac{P(D, K_1, M)P(D, K_2, M)}{P(D, M)} \tag{10}$$

according to Fig. 4e, where $P(D, K_1, M)$ and $P(D, K_2, M)$ are the cluster potentials and $P(D, M)$ is the separator potential.

The equivalent BN topology of the BN outlined in Fig. 3a can be described as in Fig. 5a based on similar assumptions and considerations. The corresponding junction tree is outlined in Fig. 5b, where there are N clusters of variables $\{\{D, K_1, M\}, \{D, K_2, M\}, \ldots \{D, K_N, M\}\}$ and $N-1$ separators $\{D, M\}$; the joint probability function of (5) can then be decomposed as

$$\begin{aligned} &P(D, K_1, \ldots, K_N, M) \\ &= \frac{\prod_{i=1}^{N} P(D, K_i, M)}{\prod_{i=1}^{N-1} P(D, M)} = \frac{\prod_{i=1}^{N} P(D, K_i, M)}{P(D, M)^{N-1}}. \end{aligned} \tag{11}$$

This indicates a new way of representing the joint probability function, $P(D, K_1, \ldots, K_N, M)$, as a composition of several local joint probability functions $P(D, K_1, M), \ldots, P(D, K_N, M)$, which correspond to the probability of observational data D given the specific additional knowledge of $K_1, K_2, \ldots, K_N$.

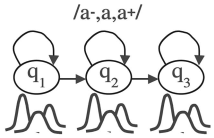


Fig. 6. The conventional HMM AM with Gaussian mixture density used to model the triphone $/a^-, a, a^+/$.

2.4 Junction Tree Inference

We obtain

$$P(D, K_i, M) = P(D|K_i, M)P(K_i|M)P(M) \tag{12}$$

for all $P(D, K_i, M)$ using the chain rule so that (11) becomes

$$\begin{aligned} &P(D, K_1, \ldots, K_N, M) \\ &= \frac{\prod_{i=1}^{N} P(D, K_i, M)}{P(D, M)^{N-1}} \\ &= \frac{\prod_{i=1}^{N} \{P(D|K_i, M)P(K_i|M)P(M)\}}{\{P(D|M)P(M)\}^{N-1}} \\ &= \frac{\prod_{i=1}^{N} P(D|K_i, M)}{P(D|M)^{N-1}} P(K_1|M) \ldots P(K_N|M)P(M). \end{aligned} \tag{13}$$

Comparing this with (5), we can see that

$$P(D|K_1, \ldots, K_N, M) = \frac{\prod_{i=1}^{N} P(D|K_i, M)}{P(D|M)^{N-1}}, \tag{14}$$

which indicates that $P(D|K_1, \ldots, K_N, M)$ can be decomposed into separate terms corresponding to the probability of observing data D given the specific additional knowledge of $K_1, K_2, \ldots, K_N$.

It will now be much easier to define, estimate, and calculate several simple $P(D|K_i, M)$ than a single but complex $P(D|K_1, \ldots, K_N, M)$.

The output probability during inference for given data d, model parameter m, and additional knowledge source k_{1_j} is then calculated as

$$p(d|k_{1_j}, \ldots, k_{N_j}, m) = \frac{\prod_{i=1}^{N} P(D = d|K_i = k_{i_j}, M = m)}{P(D = d|M = m)^{N-1}}. \tag{15}$$

3 Conventional HMM AM

Let us now define some notations related to the conventional HMM. We denote an HMM phonetic model of triphone context $/a^-, a, a^+/$ with λ and the HMM state variable with Q. X is an observation variable and $X_s = X_t, \ldots, X_{t+m}$ is an observation data segment of length m. The standard HMM structure is outlined in Fig. 6, where

1. the short-term spectral characteristics are modeled with a mixture of Gaussians and
2. the temporal speech characteristics are governed by HMM state transitions.

The HMM state output probability, $p(x_t|q_i)$, is usually calculated from the state PDF, $P(X|Q)$, as

$$\begin{aligned} p(x_t|q_i) &= P(X = x_t|Q = q_i) \\ &= \sum_{m=1}^{M} b_m \mathcal{N}(x_t; \mu_m, \Sigma_m), \end{aligned} \tag{16}$$

where b_m is the mixture weight for the $m\text{th}$ mixture in state q_i and $\mathcal{N}(.)$ is a Gaussian function with mean vector μ_m and covariance matrix Σ_m. The HMM segmental likelihood, $P(X_s|\lambda)$, is then calculated from the joint probability of observation and the state sequence, taken over all state sequences (total likelihood) or approximately over just the most likely state sequence (Viterbi path) [2].

4 Incorporating Knowledge Sources at HMM State Level

4.1 Common Considerations

Model M is currently our triphone HMM state Q and D is observation variable X. Following the theoretical framework described in Section 2, we proceed with the next two steps:

1. **Defining the causal relationship.** The structure of the topology is similar to the one in Fig. 2a and the triphone HMM state PDF is now represented by the BN joint probability function, which is similar to (3):

   $$P(X, Q) = P(X|Q)P(Q). \tag{17}$$

   We can simply follow (5) so that

   $$\begin{aligned} &P(X, K_1, \ldots, K_N, Q) \\ &= P(X|K_1, \ldots, K_N, Q)P(K_1|Q) \ldots P(K_N|Q)P(Q) \end{aligned} \tag{18}$$

   to incorporate additional knowledge sources $K_1, K_2, \ldots, K_N$ into our HMM state distribution $P(X, Q)$ (assuming that all $K_1, K_2, \ldots, K_N$ are independent given Q).
2. **Inference.** Our primary interest is to calculate the HMM state output probability, $P(X|K_1, \ldots, K_N, Q)$, which can easily be modeled with a Gaussian function. We can thus directly obtain state output. If all additional knowledge sources $K_1, \ldots, K_N$ are assumed to be hidden as described in Section 2.2, the state output probability is obtained as in (8) by marginalization over all possible $K_i : k_{i_1}, k_{i_2}, \ldots, k_{i_M}$ for all K_i, $1 \leq i \leq N$:

   $$\begin{aligned} &p(x_t|q_t) = \\ &\sum_{1_j=1}^{M_1} \cdots \sum_{N_j=1}^{M_N} p(x_t|k_{1_j}, \ldots, k_{N_j}, q_t)p(k_{1_j}|q_t) \ldots p(k_{N_j}|q_t). \end{aligned} \tag{19}$$

   Here, we can see that (19) is also equivalent to the state output probability of the conventional HMM in (16) if we treat term $p(k_{1_j}|q_t) \ldots p(k_{N_j}|q_t)$ as a mixture weight coefficient for the Gaussian component $p(x_t|k_{1_j}, \ldots, k_{N_j}, q_t)$. Since expressions such as (19) represent a mixture of Gaussians, we can undertake recognition using existing HMM-based decoders without the need for any modifications. Also, since BN is only used to infer the state output likelihood, this allows us to retain our HMM-based triphone AM topology, where HMM state transitions are still used to govern temporal speech characteristics. This approach is also known as the hybrid HMM/BN modeling framework and is described in [17], [18]. After this, we will also call the model obtained by incorporating additional knowledge at the state level the HMM/BN model.

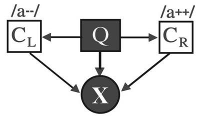


Fig. 7. The BN-C topology for modeling a pentaphone context $/a, a^-, a, a^+, a^{++}/$, where state PDF has additional variables C_L and C_R representing the second preceding and following contexts, respectively.

Parameter learning of this model can be adopted from the general training of the HMM/BN model [17]. This is based on the forward-backward algorithm, where each training iteration consists of BN training and HMM transition probability updates. BN training is done using standard statistical methods. Maximum likelihood (ML) parameters estimation is applied if all variables are observable during training; however, if some are hidden, the parameters can then be estimated with the standard expectation maximization (EM) algorithm.

4.2 Incorporation of Wide-Phonetic Context Information

The most widely used acoustic unit in ASR systems is currently still the triphone, which includes the immediate preceding and following phonetic context. Although triphones have proved to be an efficient choice, wider phonetic contexts seem to be more appropriate for capturing longer spans of coarticulation effects; however, these often suffer from the problem of a sparsity of data and memory constraints. Here, we will explain how to apply our framework, described in the previous section, to the problem of incorporating wide-phonetic knowledge information.

Assume that we need to extend our conventional HMM λ of triphone context $/a^-, a, a^+/$ into a pentaphone context such as $/a, a^-, a, a^+, a^{++}/$. We therefore incorporate the additional second preceding and succeeding contexts, C_L $(/a/)$ and C_R $(/a^{++}/)$, into the triphone state PDF by inserting two new variables into BN based on our approach.

The conditional relationship between the triphone HMM state Q, observation data X, and the two additional variables, C_L and C_R, is described by the BN topology, as outlined in Fig. 7. We will call this the BN-C topology.

The HMM state PDF is currently represented by the BN joint probability, which, according to (18), can be decomposed as

$$\begin{aligned} &P(X, C_L, C_R, Q) \\ &= P(X|C_L, C_R, Q)P(C_L|Q)P(C_R|Q)P(Q), \end{aligned} \tag{20}$$

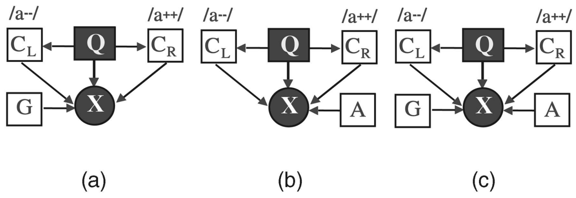


Fig. 8. (a) BN-CG topology with additional variables G, C_L, and C_R, (b) BN-CA topology with additional variables A, C_L, and C_R, and (c) BN-CGA topology with additional variables A, G, C_L, and C_R.

where X depends on both second preceding context C_L and second following context C_R. Since X is continuous and C_L, C_R, and Q are discrete variables, $P(X|C_L, C_R, Q)$ is modeled with a Gaussian function and each $P(C_L|Q)$ or $P(C_R|Q)$ is represented by a CPT.

The state output probability can be obtained from $P(X|C_L, C_R, Q)$ and, assuming that the additional context variables, C_L and C_R, cannot be observed (hidden) during recognition, as in (19),

$$p(x_t|q_i) = \sum_{c_l=1}^{N_L} \sum_{c_r=1}^{N_R} p(c_l|q_i)p(c_r|q_i)p(x_t|c_l, c_r, q_i), \tag{21}$$

which is equivalent to the state output probability of the conventional HMM in (16) if we treat term $p(c_l|q_i)p(c_r|q_i)$ as a mixture weight coefficient for the Gaussian component, $p(x|c_l, c_r, q_i)$. Thus, here, a Gaussian PDF is trained for all combinations of c_l, c_r, and q_i.

We can also further extend the pentaphone BN with other additional knowledge variables such as gender or accent information using this framework. Fig. 8 describes several examples of conditional relationship structures between triphone HMM state Q, observation data X, the two additional variables, C_L and C_R, and the gender, G, or accent, A, variables. The BN topology becomes the one described in Fig. 8a by extending BN-C with an additional variable of gender G and is called BN-CG. The BN topology becomes the one in Fig. 8b by extending BN-C with the additional accent variable A and is called BN-CA. The BN topology in Fig. 8c is extended with both accent and gender variables and is called BN-CGA.

The HMM state PDF for the BN-CGA example (see Fig. 8c) is expressed as

$$\begin{aligned} &P(X, C_L, C_R, Q, A, G) \\ &= P(X|C_L, C_R, Q, A, G)P(C_L|Q)P(C_R|Q) \\ &\quad P(Q)P(A)P(G), \end{aligned} \tag{22}$$

where X depends on accent A, gender G, the second preceding context, C_L, and the second following context, C_R. The state output probability can also be obtained from $P(X|C_L, C_R, Q, A, G)$ in a way similar to that in (21):

$$\begin{aligned} p(x_t|q_i) = &\sum_{a=1}^{N_A} \sum_{g=1}^{N_G} \sum_{c_l=1}^{N_L} \sum_{c_r=1}^{N_R} p(a)p(g)p(c_l|q_i)p(c_r|q_i) \\ &p(x_t|c_l, c_r, q_i, a, g). \end{aligned} \tag{23}$$

Here, we also treat the term $p(a)p(g)p(c_l|q_i)p(c_r|q_i)$ as a mixture weight coefficient for the Gaussian component $p(x|c_l, c_r, q_i, a, g)$ so that each Gaussian PDF is trained for each combination of c_l, c_r, q_i, a, and g.

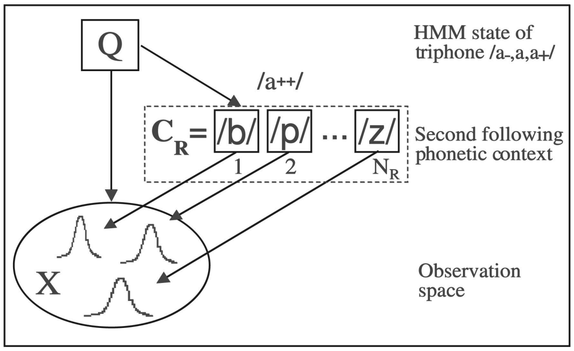


Fig. 9. An example of observation space modeling by BN, where a different value of C_R corresponds to a different Gaussian.

Both (21) and (23) represent a mixture of Gaussians as used in the standard triphone HMM AM. We can thus undertake recognition using existing triphone HMM-based decoders without having to modify them. Parameter learning of the proposed model is done as mentioned in the previous section. The ML parameter estimation is used since all variables, including triphone state Q, accent A, gender G, second preceding (C_L) context, second following (C_R) context, and variable X can be observed during training.

We can reduce the number of parameters with clustering techniques, such as knowledge-based or data-driven clustering, if there is an insufficient amount of training data to obtain reliable estimates for all model parameters. For example, for each value c_l/c_r of the second preceding/following phonetic context, C_R/C_R, there is a corresponding Gaussian component according to (21) and (23). Fig. 9 outlines the observation space for BN with additional C_R only. If we use a 44-phoneme set (including silence) for English ASR, this 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 state with BN-C topology (see Fig. 7) may become $44^2 = 1,936$ and even much more for states with BN-CG, BN-CA, and BN-CGA topologies.

Here, we group the phoneme contexts based on major distinctions in the manner of articulation to reduce the sizes of parameters. Table 1 lists examples of knowledge-based phoneme classes adapted from that in [19].

More details and a discussion on the possibilities of pentaphones based on HMM/BN approaches can be found in [20], [21].

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

5 Incorporating Source of Knowledge at the Phonetic Model Level

5.1 Common Considerations

Again following the theoretical framework described in Section 2, model M is our current HMM phonetic model, λ, and D is segment X_s, we proceed with the next two steps:

1. **Defining the causal relationship.** The structure of the topology is similar to the one in Fig. 2a and the probability function of HMM phonetic units is now represented by the BN joint probability function, similar to (3)

$$P(X_s, \lambda) = P(X_s|\lambda)P(\lambda). \tag{24}$$

To incorporate additional knowledge sources $K_1, K_2, \ldots, K_N$ into our HMM phonetic model, $P(X_s, \lambda)$ (assuming that all $K_1, K_2, \ldots, K_N$ are independent given λ), we can simply follow (5) so that

$$\begin{aligned} &P(X_s, K_1, \ldots, K_N, \lambda) \\ &= P(X_s|K_1, \ldots, K_N, \lambda)P(K_1|\lambda) \ldots P(K_N|\lambda)P(\lambda). \end{aligned} \tag{25}$$

2. **Inference.** Our primary interest now is to calculate $P(X_s|K_1, \ldots, K_N, \lambda)$ given input segment X_s. However, it is difficult to obtain a simple functional form for this conditional PDF because it involves an HMM model, λ, and a segment, X_s, of variable duration. Thus, here, we need to decompose $P(X_s|K_1, \ldots, K_N, \lambda)$ by the junction tree algorithm, as described in Section 2.3. It can be decomposed as

$$P(X_s|K_1, \ldots, K_N, \lambda) = \frac{\prod_{i=1}^{N} P(X_s|K_i, \lambda)}{P(X_s|\lambda)^{N-1}}, \tag{26}$$

according to (14), which indicates a new way of representing HMM phonetic likelihood $P(X_s|K_1, \ldots, K_N, \lambda)$ through the composition of several less complex dependencies, that is, $P(X_s|K_1, \lambda), \ldots, P(X_s|K_N, \lambda)$, which correspond to the likelihood of segment observation data X_s given the respective specific additional knowledge of $K_1, K_2, \ldots,$ or K_N.

5.2 Incorporation of Wide-Phonetic Context Information

Let us apply the approach described in the previous section to the same task of incorporating wide-phonetic knowledge information, where we extend triphone context $/a^-, a, a^+/$ into a pentaphone $/a, a^-, a, a^+, a^{++}/$. Structurally, the conventional HMM of a triphone-context unit model can be described as in Fig. 10a and that of a pentaphone-context unit models can be described as in Fig. 10b.

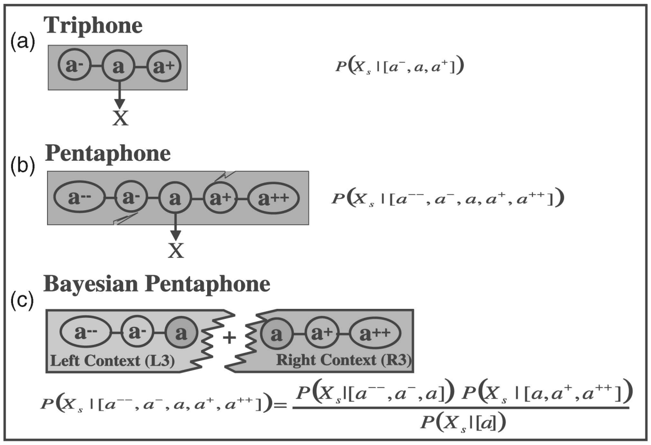

(a)
Triphone
a-
a
a+
X
$P(X_s \mid [a^-, a, a^+])$
(b)
Pentaphone
a--
a-
a
a+
a++
X
$P(X_s \mid [a^{--}, a^-, a, a^+, a^{++}])$
(c)
Bayesian Pentaphone
a--
a-
a
+
a
a+
a++
Left Context (L3)
Right Context (R3)
$P(X_s \mid [a^{--}, a^-, a, a^+, a^{++}]) = \frac{P(X_s \mid [a^{--}, a^-, a]) \, P(X_s \mid [a, a^+, a^{++}])}{P(X_s \mid [a])}$

Fig. 10. (a) The conventional triphone model, (b) the conventional pentaphone model, and (c) the Bayesian pentaphone model composition C1L3R3, consisting of the preceding/following triphone-context unit and center-monophone unit.

We incorporate the additional second preceding C_L of $/a/$ and succeeding contexts C_R of $/a^{++}/$ into probability function $P(X_s|\lambda)$. The conditional relationship between X_s, λ, and C_L and C_R is described by BN similar to the one in Fig. 4a. The final junction tree decomposition is also similar to the one in Fig. 4e, where M is our current HMM phonetic model, λ, and D is segment X_s. The conditional probability function is then defined as

$$P(X_s|C_L, C_R, \lambda) = \frac{P(X_s|C_L, \lambda)P(X_s|C_R, \lambda)}{P(X_s|\lambda)}, \tag{27}$$

according to (26). Since λ is associated with the triphone $/a^-, a, a^+/$, the second preceding C_L with $/a/$ and the second succeeding C_R with $/a^{++}/$, we can write

$$\begin{aligned} &P(X_s|C_L, C_R, \lambda) \\ &= P(X_s|a, a^{++}, [a^-, a, a^+]) \\ &= P(X_s|[a, a^-, a, a^+, a^{++}]) \end{aligned} \tag{28}$$

and (28) becomes

$$\begin{aligned} &P(X_s|[a, a^-, a, a^+, a^{++}]) \\ &= \frac{P(X_s|a, [a^-, a, a^+])P(X_s|a^{++}, [a^-, a, a^+])}{P(X_s|[a^-, a, a^+])} \\ &= \frac{P(X_s|[a, a^-, a, a^+])P(X_s|[a^-, a, a^+, a^{++}])}{P(X_s|[a^-, a, a^+])}. \end{aligned} \tag{29}$$

This indicates that a pentaphone model can be composed of $p(X_s|[a, a^-, a, a^+])$, $p(X_s|[a^-, a, a^+, a^{++}])$, and $p(X_s|[a^-, a, a^+])$, which correspond to the likelihood of segment X_s given the left/preceding-tetraphone-context, right/following-tetraphone-context, and the center-triphone-context units. However, developing tetraphone models for $[a, a^-, a, a^+]$ and $[a^-, a, a^+, a^{++}]$ may also be difficult due to the sparsity of data.

Instead, let us use (28) and adjust λ to represent a monophone, $/a/$, and the second preceding and succeeding contexts, C_L and C_R, to respectively represent $/a, a^-/$ and $/a^+, a^{++}/$. Then,

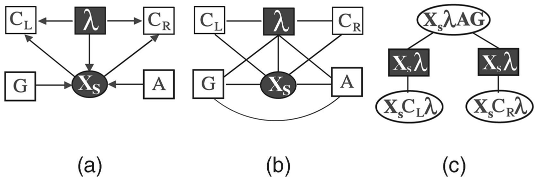


Fig. 11. (a) BN topology describing the conditional relationship between X_s, λ, C_L, C_R, A, and G. (b) Moral and triangulated graph of that in Fig. 11a. (c) The corresponding Junction tree.

$$\begin{aligned} &P(X_s|[a, a^-, a, a^+, a^{++}]) \\ &= \frac{P(X_s|[a, a^-], a)P(X_s|[a^+, a^{++}], a)}{P(X_s|[a])} \\ &= \frac{P(X_s|[a, a^-, a])P(X_s|[a, a^+, a^{++}])}{P(X_s|[a])}, \end{aligned} \tag{30}$$

which indicates that the pentaphone-context, $/a, a^-, a, a^+, a^{++}/$, is composed of $p(X_s|[a, a^-, a])$, $p(X_s|[a, a^+, a^{++}])$, and $p(X_s|[a])$, which correspond to the likelihood of observation X_s given the left/preceding-triphone-context unit (L3), the right/following-triphone-context unit (R3), and the monophone unit (C1). We call this composition C1L3R3 and it is shown structurally in Fig. 10c.

As can be seen, the number of context units to be estimated is reduced from N^5 to $(2N^3 + N)$, without loss of context coverage, where N is the number of phones. If we use a 44-phoneme set for English ASR, the total number of different contexts that need to be estimated in the pentaphone model is $44^5 =\sim 165,000,000$ context units. A composition with triphone-context units reduces the complexity to about 170,000 context units.

Analyzing (29) and (30), we can see that (27) can be used as a starting point for deriving other compositions of the HMM phonetic model as well. In the case where we assume that λ is monophone unit $/a/$, and C_L and C_R are the ones preceding and following context unit $/a^-/$ and $/a^+/$, respectively, we can obtain the same factorization as one that has been proposed [22], [23] and that is known as the Bayesian triphone:

$$P(X_s|[a^-, a, a^+]) = \frac{P(X_s|[a^-, a])P(X_s|[a, a^+])}{P(X_s|[a])}, \tag{31}$$

where the triphone model is constructed from monophone and biphone models. After this, any models composed in this way will also be called Bayesian models.

The extended version of the Bayesian triphone, the so-called Bayesian wide-phonetic context model, can also be found in our previous study [24], [25]. This approach allows us to model a wide range of phonetic contexts from less context-dependent models simply based on Bayes' rule. However, difficulties arise when different types of knowledge sources need to be incorporated.

In contrast, the current unified framework gives us a more appropriate means of incorporating various kinds of knowledge sources. For example, we can easily further extend C1L3R3 with other additional knowledge variables such as gender or accent information. We can extend C1L3R3 with gender information only (C1L3R3-G), with accent information only (C1L3R3-A), or with both (C1L3R3-AG). For the case of C1L3R3-AG, the BN topology, when the moral and triangulated graph, and also its corresponding junction tree become as that shown in Fig. 11, then the conditional probability function is obtained as

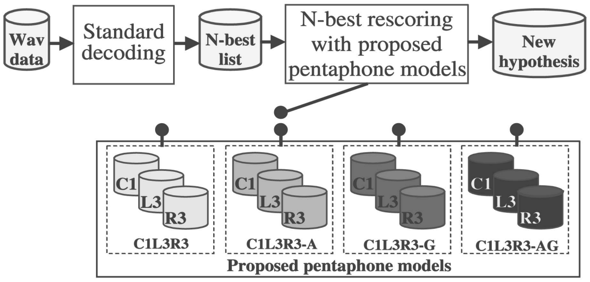


Fig. 12. The rescoring procedure with Bayesian pentaphone models.

$$\begin{aligned} &P(X_s|C_L, C_R, \lambda, A, G) \\ &= P(X_s|\lambda, A, G)\frac{P(X_s|C_L, \lambda)}{P(X_s|\lambda)}\frac{P(X_s|C_R, \lambda)}{P(X_s|\lambda)} \\ &= \frac{P(X_s|\lambda, A, G)P(X_s|C_L, \lambda)}{P(X_s|\lambda)}\frac{P(X_s|\lambda, A, G)P(X_s|C_R, \lambda)}{P(X_s|\lambda)} \\ &\quad \cdot \frac{1}{P(X_s|\lambda, A, G)} \\ &= \frac{P(X_s|C_L, \lambda, A, G)P(X_s|C_R\lambda, A, G)}{P(X_s|\lambda, A, G)}. \end{aligned} \tag{32}$$

Thus, following the same setting as C1L3R3 for λ, C_L, and C_R, the pentaphone likelihood of C1L3R3-AG becomes

$$\begin{aligned} &P(X_s|[a, a^-, a, a^+, a^{++}], A, G) \\ &= \frac{P(X_s|[a, a^-, a], A, G)P(X_s|[a, a^+, a^{++}], A, G)}{P(X_s|[a], A, G)}, \end{aligned} \tag{33}$$

which indicates that $P(X_s|[a, a^-, a, a^+, a^{++}], A, G)$ can be simplified by factorizing it into $P(X_s|[a], A, G)$, $P(X_s|[a, a^-, a], A, G)$, and $P(X_s|[a, a^+, a^{++}], A, G)$.

The implementation of the proposed pentaphone models into an ASR system requires a special decoder that can work with several models. This can be avoided if the proposed pentaphone models are applied by rescoring the N-best list generated by a standard triphone-based HMM system. There is a block diagram of such a rescoring procedure in Fig. 12.

Word-level N-best recognition is carried out using a conventional HMM AM and standard Viterbi decoding for all utterances in the test data. All N-best hypotheses include an acoustic score, a language model (LM) score, and a Viterbi segmentation of all phonemes. Every phoneme segment in each hypothesis is then rescored using the proposed pentaphone models, as seen in Fig. 13.

There might be some phonetic contexts that have not appeared during training. For such contexts, the proposed pentaphone context model is not able to produce any output probability during recognition. We simply assign a small numeric value as an output probability to handle this problem. Flooring is applied to all component models since this rescoring involves the output probability from the preceding, the following, and the center models.

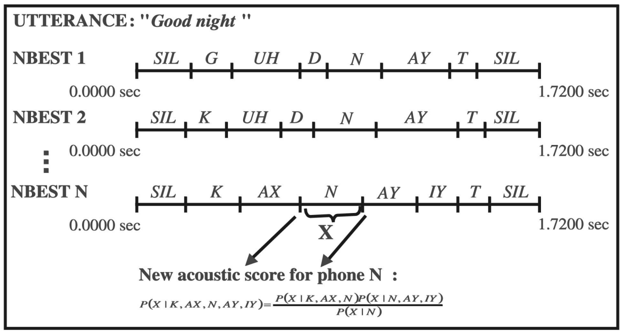


Fig. 13. The N-best rescoring mechanism.

The estimates of parameters even for the proposed pentaphone model may become unreliable if there is an insufficient amount of training data, as may occur with the state output. We used deleted interpolation to improve the reliability of the model, which allows us to fall back to a more reliable model when the supposedly more precise model is, in fact, unreliable [26]. The concept involves interpolating two separately trained models, one of which is more reliably trained than the other. However, instead of interpolating two models, we applied this approach to interpolating two phonetic likelihoods, where the phonetic likelihood of the proposed Bayesian pentaphone model, $P(X_s|\lambda_{bayPenta})$, is the precise one, whereas the triphone likelihood, $P(X_s|\lambda_{triphn})$, is the more reliable one; therefore, the interpolation phonetic likelihood, $P(X_s|\lambda)$, is obtained as

$$P(X_s|\lambda) = \alpha P(X_s|\lambda_{bayPenta}) + (1-\alpha)P(X_s|\lambda_{triphn}), \quad (34)$$

where α represents the weight of the HMM phonetic likelihood of the proposed pentaphone model and $(1-\alpha)$ represents the weight of the HMM phonetic likelihood of the triphone model. If there is a sufficiently large amount of training data, $P(X_s|\lambda_{bayPenta})$ becomes more reliable and α is expected to tend toward 1.0. However, if there is not, α will tend toward 0.0 so as to fall back to the more reliable model, $P(X_s|\lambda_{triphn})$.

All left/right contexts will be filled by silence at the beginning/end utterances. Since we assumed that there would be no long silences between adjacent words, the last phonetic context from the previous word will also affect the beginning phonetic context of the current word. This rescoring mechanism thus behaves the same way for all segments within and in-between words (crossword model). The new scores are then combined with the LM score for the current hypothesis. The hypothesis achieving the highest total utterance score from the N-best is selected as the new recognition output.

6 Experiments

The accented English speech corpus of the Advanced Telecommunication Research (ATR) Institute International (Japan) was used in these experiments. The text material was based on the basic domain of expressions used in travel. The speech database consisted of American (US) and Australian (AUS) English accents with about 45,000 utterances (~ 44 speech hours) spoken by 100 speakers (50 males and 50 females) for each accent. We used 90 percent of the data or about 40,000 utterances (20,000 utterances by 40 speakers for each male and female) as the training data. We randomly selected 200 utterances for evaluation, spoken by 20 different speakers (10 males and 10 females) from the remaining 10 percent of mixed accented data (US and AUS). Bigram and trigram language models were trained on about 150,000 travel-related sentences. The available pronunciation dictionary consisted of about 37,000 words and was based on US pronunciations.

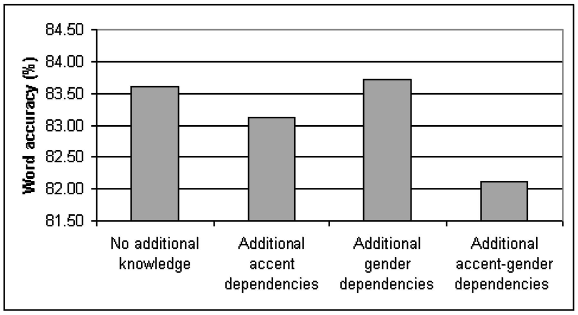


Fig. 14. Comparing recognition word accuracy rates of the triphone baseline models having the same five mixture components per state on the average.

A sampling frequency of 16 kHz, a frame length of 20 ms, a frame shift of 10 ms, and 25-dimensional feature parameters consisting of 12-order MFCC, Δ MFCC, and Δ log power were used as the feature parameters. Three states were used as the initial HMM for all phonemes. A triphone AM with a shared-state HMnet topology was then obtained using a successive state splitting (SSS) training algorithm. The number of shared states was determined automatically by the algorithm since the SSS algorithm used here was based on the minimum description length (MDL) optimization criterion. Details on the MDL-SSS can be found elsewhere [27]. The SSS topology training was done using all training data. The number of states was 2,126 in total and models with four different versions of Gaussian mixture components per state, that is, 5, 10, 15, and 20, were obtained.

It is also possible to incorporate additional knowledge such as gender and accent in the conventional triphone AM by training gender and/or accent dependent AMs. Only a procedure of embedded training was conducted with a specific accent or gender training data to ensure the same structure for the topology for all models. Thus, in total, we obtained one single triphone AM (without any additional knowledge), two accent-dependent triphone AMs (for both US and AUS), two gender-dependent triphone AMs (for both males and females), and four accent-gender-dependent triphone AMs (for US males and females and AUS males and females).

How well these baseline models performed with five mixture components per state is plotted in the graph in Fig. 14. The triphone baseline without any additional knowledge achieved a word accuracy of 83.60 percent. However, only gender-dependent models could improve performance slightly. The performance of the other models only decreased. This especially decreased to a word accuracy of 82.11 percent for the accent-gender-dependent models.

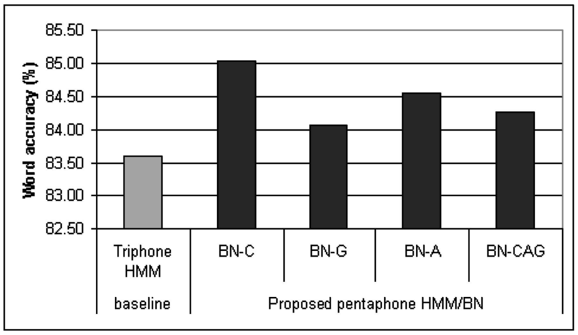


Fig. 15. Comparing recognition word accuracy rates of the pentaphone HMM/BN models using different BN topologies (BN-C, BN-CG, BN-CA, and BN-CGA, as shown in Figs. 7 and Figs. 8a, 8b, and 8c respectively), but having the same five mixture components per state on the average.

This might be due to the amount of training data, which was much smaller compared to the other baseline models.

6.1 Performance When Incorporating Knowledge Sources at HMM State Level

The proposed pentaphone models were trained using the same amount of training data on all accent data labeled with phoneme-class context variables, as described in Section 4.2. The model state topology, the total number of states, and the transition probabilities were all identical to the triphone HMM baseline. Therefore, they all had similar complexity in terms of the number of parameters. The main difference was only in the probability distribution of states, where each Gaussian was explicitly conditioned on C_L or C_R. All Gaussian components in the HMM baseline, in contrast, were learned implicitly by the EM algorithm, without any "meaningful" interpretation of the mixture index. There were some phoneme context classes of C_L or C_R which did not exist due to grammatical rules or did not appear in the training data, which, after training, resulted in about 50 Gaussians per state on the average. We used a data-driven clustering technique and reduced the size of the pentaphone models to correspond to 5, 10, 15, and 20 mixture components per state to avoid unreliably estimated parameters and to be able to compare the performance with the baseline system by having exactly the same total number of Gaussians.

We first evaluated how well the pentaphone models BN-C, BN-CG, BN-CA, or BN-CGA performed, using the same test data as for the baseline. The results for all four models having the same five mixture components per state on the average are plotted in Fig. 15. As can be seen, we obtained improved recognition using all types of BNs by only changing the probability distribution of states to incorporate various type of knowledge sources. However, the incorporation of gender and accent variables did not improve the recognition rate of the proposed models any further. This problem may again be related to the limited amount of training data for each accent or gender dependent model. That is why the best performance was obtained using BN-C achieving a word accuracy of 85.03 percent.

We evaluated it on a test set of matching accents, where the test data were 200 randomly selected utterances from each accent (US and AUS) to investigate what effect using BN-C had in more detail. The results obtained with models with different numbers of mixture components are summarized in Table 2.

TABLE 2
Recognition Accuracy Rates (%) for the Proposed Pentaphone HMM/BN Model Using BN-C (See Fig. 7) on a Test Set of Matching Accents with Different Number of Mixture Components

Mixture number	US accent		AUS accent	
	Triphn baseline	Proposed BN-C	Triphn baseline	Proposed BN-C
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

It can be seen that the proposed pentaphone models always performed better than the baseline within the same number of parameters. The best performance for the US pentaphone HMM/BN was obtained with 10 Gaussian mixtures, which resulted in a relative reduction in WER of about 8 percent, and the best performance for the AUS pentaphone was obtained with 20 Gaussian mixtures, which resulted in a relative reduction in WER of about 11 percent. We also evaluated the performance of these pentaphone models on a test set of mismatching accents, for example, the US speech trained model was tested on the AUS speech test data and vice versa. The results obtained using the models with 15 mixture components are summarized in Table 3. The results from evaluating matching accents have also been included to enable easy comparison. We can see that the pentaphone model on mismatching accents still consistently outperforms the standard HMM triphone model.

6.2 Performance When Incorporating Knowledge Sources at HMM Phonetic Model Level

We investigated [24], [25] several different ways of decomposing pentaphone models and found that the best was C1L3R3 composition. Here, we describe additional experiments only using the C1L3R3 model.

All components of all accented pentaphone models were trained separately using the same amount of training data and the same SSS training algorithm. There were a total of 3,660 states (sum of C1: 132 states, L3: 1,746 states, and R3: 1,782 states) with four different versions of Gaussian mixture component numbers per state, that is, 5, 10, 15, and 20. An embedded training procedure was then carried out for pentaphones C1L3R3-A, C1L3R3-G, and C1L3R3-AG with specific accent or gender training data.

TABLE 3
Recognition Accuracy Rates (%) for the Proposed Pentaphone HMM/BN Model Using BN-C (See Fig. 7) on a Test Set of Mismatching Accents with 15 Mixture Components

Accented test set	US accent		AUS accent	
	Triphn baseline	Proposed BN-C	Triphn baseline	Proposed BN-C
US test	84.78	85.55	75.22	76.96
AUS test	64.78	65.43	83.46	84.18

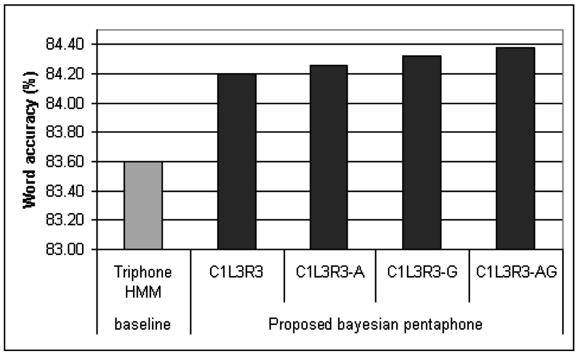


Fig. 16. Comparing recognition accuracy rates of different Bayesian pentaphone models, C1L3R3, C1L3R3-A, C1L3R3-G, and C1L3R3-AG, having the same five mixture components per state.

We first evaluated what effect incorporating additional knowledge sources would have on multiaccented test data. The results for the proposed pentaphones C1L3R3, C1L3R3-A, C1L3R3-G, and C1L3R3-AG having five mixture components are summarized in Fig. 16. The rescoring was done using a 10-best list and a 0.3 weight parameter, α, for deleted interpolation. As can be seen, the more knowledge sources we incorporated, the better the performance. The proposed pentaphone C1L3R3 model improved performance relative to the baseline and the best performance that was achieved was a word accuracy of 84.38 percent with C1L3R3-AG, which incorporated additional knowledge of accent A, gender G, second preceding context C_L, and succeeding context C_R. Performance did not decrease when gender and accent were incorporated, as was the case for pentaphone HMM/BN, which is probably due to the use of deleted interpolation.

We next investigated how well C1L3R3-AG performed on all accented test data in more detail, with the N-best $(\mathrm{N} = 10)$ list. The weight parameter, λ, for deleted interpolation was the same (0.3). Here, we measured both the relative improvement (Rel-Imp) and the relative improvement to rescoring (Rel-Resc-Imp) as used in [10]:

$$RelRescImp = \frac{RescoringResult - Baseline}{NbestListUpperBound - Baseline}, \qquad (35)$$

where the N-best list upper bound is the N-best recognition result.

The results obtained with models with different numbers of mixture components are summarized in Table 4. As can be seen, the proposed pentaphone model consistently improved the performance of the ASR system. The largest Rel-Resc-Imp was achieved with 15 mixture models for both US and AUS accents (37.92 percent for the US model and 38.04 percent for the AUS model).

We also evaluated how well the proposed pentaphone C1L3R3-AG model performed on a test set of mismatching accents. The results obtained using a model with 15 mixture components are summarized in Table 5. The results from evaluating matching accents have also been included to enable easy comparison. We can see that the pentaphone C1L3R3-AG model we propose also consistently outperforms the standard triphone model with mismatching accents.

6.3 Comparison of Different Models

Last, we conducted additional experiments with a conventional pentaphone HMM model with 2,202 states, which was trained from scratch using MDL-SSS, to investigate whether the superior performance of our proposed models is mainly due only to the wide-phonetic context. Pentaphone models that were dependent on accent and gender were also obtained using a procedure of embedded training with specific accent or gender training data. They were implemented by rescoring the N-best list, as was the case for the Bayesian pentaphone.

The results for all models with five mixture components per state are plotted in Fig. 17. As can be seen, the proposed pentaphone C1L3R3 model improved performance relative to the baseline and this was better than just rescoring with the conventional pentaphone HMM. This might be because, given the amount of training data, the training of the conventional pentaphone model using the MDL-SSS algorithm resulted in a model with 2,202 total states, which is

TABLE 4
Recognition Accuracy Rates (%) for the Proposed Bayesian Pentaphone C1L3R3-AG (See (33)) on a Test Set of Matching Accents with Different Number of Mixture Components

	US accent					AUS Accent				
Mixture number	**Upper bound**	**Triphn baseline**	**Proposed C1L3R3-AG**	*Rel Imp*	*Rel Resc-Imp*	**Upper bound**	**Triphn baseline**	**Proposed C1L3R3-AG**	*Rel Imp*	*Rel Resc-Imp*
5 mix	87.52	84.30	85.19	*5.67*	*27.64*	85.79	82.33	83.76	*8.09*	*41.33*
10 mix	87.94	84.66	85.79	*7.37*	*34.45*	85.37	82.21	82.81	*3.37*	*18.99*
15 mix	87.76	84.78	85.91	*7.42*	*37.92*	86.93	83.46	84.78	*7.98*	*38.04*
20 mix	87.78	85.25	85.91	*4.47*	*26.09*	86.39	82.63	83.58	*5.47*	*25.27*

TABLE 5
Recognition Accuracy Rates (%) for the Proposed Bayesian Pentaphone C1L3R3-AG Model (See (33)) on a Test Set of Mismatching Accents with 15 Mixture Components

	US accent					AUS accent				
Accented test set	**Upper bound**	**Triphn baseline**	**Proposed C1L3R3-AG**	*Rel Imp*	*Rel Resc-Imp*	**Upper bound**	**Triphn baseline**	**Proposed C1L3R3-AG**	*Rel Imp*	*Rel Resc-Imp*
US Test	87.76	84.78	85.91	*7.42*	*37.92*	80.60	75.22	77.31	*8.43*	*38.85*
AUS Test	71.76	64.78	68.12	*9.48*	*47.85*	86.93	83.46	84.78	*7.98*	*38.04*

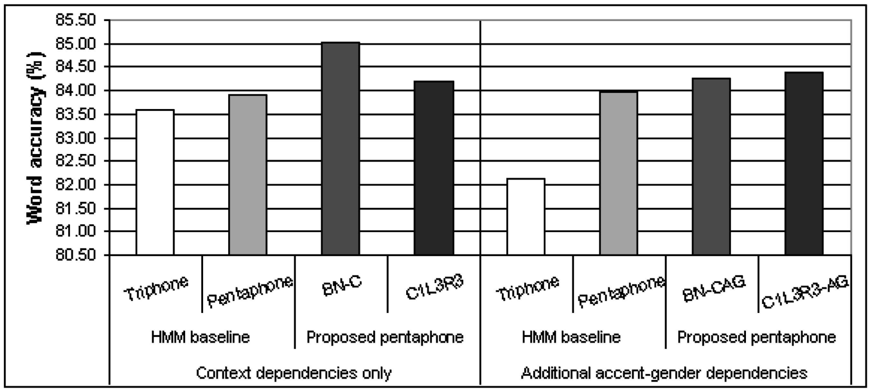


Fig. 17. Comparing the recognition accuracy rates of different systems triphone HMM baseline, pentaphone HMM baseline, and the proposed pentaphone models.

not that different from the total number of states in the triphone HMM. The resolution of context was reduced as there seemed to be too many different pentaphone contexts sharing the same Gaussian components. Thus, approximating a pentaphone model using the composition of several less context-dependent models could help to increase the resolution of context and improve performance. The best performance that was achieved was a word accuracy of 85.03 percent with BN-C.

7 Conclusion

We presented a general framework for incorporating additional knowledge sources into HMM-based statistical acoustic models. We also demonstrated the implementation of this framework by incorporating wide-phonetic context information into a triphone HMM. This was first done at HMM state level by means of BN. If the additional knowledge sources are assumed to be hidden during recognition, our approach allows the use of the standard decoding system without modification. Second, we incorporated the wide-phonetic context acoustic modeling at the HMM phonetic model level by constructing a model from several other models that had narrower contexts. As this technique of composition led to a reduction in the number of context units to be estimated, the resolution of contexts could be considerably improved since only less context-dependent models needed to be estimated. We applied these wide-context-model compositions at the postprocessing stage with N-best rescoring. The experimental results revealed that the wide-phonetic context models developed with the proposed framework improved word accuracy with respect to standard triphone models. Additional knowledge of the second preceding context, C_L, and the succeeding context, C_R, was appropriate to incorporate at HMM state level, whereas additional knowledge of accent A and gender G, was more appropriate to incorporate at the HMM phonetic model level.

References

[1] J. Holmes and W. Holmes, *Speech Synthesis and Recognition.* Taylor and Francis, 2001.

[2] W.J. Holmes and M. Huckvale, "Why Have HMMs Been So Successful for Automatic Speech Recognition and How Might They Be Improved," *Speech, Hearing, and Language,* vol. 8, pp. 207-219, 1994.

[3] D.H. Klatt, *Review of the ARPA Speech Understanding Project,* vol. 62, pp. 1345-1366, Acoustical Soc. Am., 1977.

[4] V.W. Zue and R.A. Cole, "Experiments on Spectrogram Reading," *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP '79),* pp. 116-119, 1979.

[5] J. Johannsen, J. MacAllister, T. Michalek, and S. Ross, "A Speech Spectrogram Expert," *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP '83),* pp. 746-749, 1983.

[6] S.E. Levinson, "Structural Methods in Automatic Speech Recognition," *Proc. IEEE,* vol. 73, pp. 1625-1650, Nov. 1985.

[7] R.P. Lippmann, "Speech Recognition by Machines and Humans," *Speech Comm.,* vol. 22, pp. 1-15, 1997.

[8] 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, 1999.

[9] M. Weintraub, K. Taussig, K. Hunicke-Smith, and A. Snodgrass, "Effect of Speaking Style on LVCSR Performance," *Proc. Int'l Conf. Spoken Language Processing (ICSLP '96),* pp. 16-19, 1996.

[10] J. Li, Y. Tsao, and C.-H. Lee, "A Study on Knowledge Source Integration for Candidate Rescoring in Automatic Speech Recognition," *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP '05),* pp. 837-840, 2005.

[11] 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 Univ., 2000.

[12] A. Ljolje, D. Hindle, M. Riley, and R. Sproat, "The AT&T LVCSR-2000 System," *Proc. Speech Transcription Workshop,* 2000.

[13] G. Zweig and S. Russell, "Probabilistic Modeling with Bayesian Networks for Automatic Speech Recognition," *Proc. Int'l Conf. Spoken Language Processing (ICSLP '98),* pp. 3010-3013, 1998.

[14] K. Daoudi, D. Fohr, and C. Antoine, "A New Approach for Multi-Band Speech Recognition Based on Probabilistic Graphical Models," *Proc. Int'l Conf. Spoken Language Processing (ICSLP '00),* pp. 329-332, 2000.

[15] F. Jensen, *An Introduction to Bayesian Network.* UCL Press, 1998.

[16] C. Huang and A. Darwiche, "Inference in Belief Networks: A Procedural Guide," *Int'l J. Approximate Reasoning,* vol. 11, pp. 1-158, 1994.

[17] K. Markov and S. Nakamura, "Forward-Backwards Training of Hybrid HMM/BN Acoustic Models," *Proc. Int'l Conf. Spoken Language Processing (ICSLP '06),* 2006.

[18] K. Markov and S. Nakamura, "Modeling Successive Frame Dependencies with Hybrid HMM/BN Acoustic Model," *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP '05),* pp. 701-704, 2005.

[19] J.J. Odell, "The Use of Context in Large Vocabulary Speech Recognition," PhD dissertation, Cambridge Univ., 1995.

[20] S. Sakti, S. Nakamura, and K. Markov, "A Hybrid HMM/BN Acoustic Model Utilizing Pentaphone-Context Dependency," *IEICE Trans. Information and Systems,* vol. E89-D, no. 3, pp. 953-961, 2006.

[21] S. Sakti, S. Nakamura, and K. Markov, "Incorporation of Pentaphone-Context Dependency Based on Hybrid HMM/BN Acoustic Modeling Framework," *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing,* 2006.

[22] J. Ming, P.O. Boyle, M. Owens, and F.J. Smith, "A Bayesian Approach for Building Triphone Models for Continuous Speech Recognition," *IEEE Trans. Speech and Audio Processing,* vol. 7, no. 6, pp. 678-684, Nov. 1999.

[23] J. Ming and F. Jack Smith, "Improved Phone Recognition Using Bayesian Triphone Models," *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP '98),* pp. 409-412, 1998.

[24] S. Sakti, S. Nakamura, and K. Markov, "Improving Acoustic Model Precision by Incorporating a Wide Phonetic Context Based on a Bayesian Framework," *IEICE Trans. Information and Systems,* vol. E89-D, no. 3, pp. 946-953, 2006.

[25] S. Sakti, S. Nakamura, and K. Markov, "Incorporating a Bayesian Wide Phonetic Context Model for Acoustic Rescoring," *Proc. European Conf. Speech Comm. and Technology (EUROSPEECH '05),* pp. 1629-1632, 2005.

[26] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing.* Prentice Hall, 2001.

[27] T. Jitsuhiro, T. Matsui, and S. Nakamura, "Automatic Generation of Non-Uniform HMM Topologies Based on the MDL Criterion," *IEICE Trans. Information and Systems,* vol. E87-D, no. 8, pp. 2121-2129, 2004.

Sakriani Sakti received the BE in informatics (cum laude) from the Bandung Institute of Technology, Indonesia, in 1999. She received the MSc degree in communication technology from the University of Ulm, Germany, in 2002. She received a scholarship award from the "DAAD-Siemens Program Asia 21st Century" in 2000 to study for the MS degree. She was an intern student at the Speech Understanding Department of the DaimlerChrysler Research Center in Ulm during the work she did on her thesis. She joined the Advanced Telecommunication Research (ATR) Spoken Language Communication (SLC) Laboratories, Japan, in 2003 as an invited engineer. She is currently a research engineer at both the National Institute of Information and Communications Technology (NICT) and ATR-SLC. She is also a PhD student in the Dialogue Systems Group of the Department of Information Technology at the University of Ulm. Her research interests include speech recognition and statistical pattern recognition.

Konstantin Markov the degree (with honors) from the St. Petersburg Technical University and the MSc and PhD degrees in electrical engineering from the Toyohashi University of Technology, Japan, in 1996 and 1999, respectively. He worked for several years as a research engineer at the Communication Industry Research Institute in Sofia. He received the Best Student Paper Award from the IEICE Society in 1998. He joined the Research Development Department of Advanced Telecommunications Research (ATR) in Japan in 1999 and commenced duties as an invited researcher at the ATR Spoken Language Communication (SLC) Research Laboratories in 2000. He is currently a senior research scientist in the Acoustics and Speech Processing Department at ATR-SLC. He is a member of the Acoustical Society of Japan (ASJ), the IEEE, and the International Speech Communication Association (ISCA). His research interests include signal processing, automatic speech recognition, Bayesian networks, and statistical pattern recognition.

Satoshi Nakamura received the BS degree in electronic engineering from the Kyoto Institute of Technology in 1981 and the PhD degree in information science from Kyoto University in 1992. He worked at the Central Research Laboratory of Sharp Corporation in Nara, Japan, from 1981 to 1993. He also worked with ATR Interpreting Telephony Research Laboratories from 1986 to 1989. He was an associate professor at the Graduate School of Information Science of the Nara Institute of Science and Technology, Japan, from 1994 to 2000. He was a visiting research professor at the CAIP Center of Rutgers University, Piscataway, New Jersey, in 1996. He is currently the director of the Advanced Telecommunication Research Spoken Language Communication (ATR-SLC) 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 of the *Journal of IEICE Information* from 2000 to 2002. He was also a member of the Speech Technical Committee of the IEEE Signal Processing Society from 2001 to 2004. He is a member of the Acoustical Society of Japan (ASJ), the Information Processing Society of Japan (IPSJ), and the IEEE.