IEICE TRANS. INF. & SYST., VOL.E89–D, NO.3 MARCH 2006

PAPER Special Section on Statistical Modeling for Speech Processing

Improving Acoustic Model Precision by Incorporating a Wide Phonetic Context Based on a Bayesian Framework

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

SUMMARY Over the last decade, the Bayesian approach has increased in popularity in many application areas. It uses a probabilistic framework which encodes our beliefs or actions in situations of uncertainty. Information from several models can also be combined based on the Bayesian framework to achieve better inference and to better account for modeling uncertainty. The approach we adopted here is to utilize the benefits of the Bayesian framework to improve acoustic model precision in speech recognition systems, which modeling a wider-than-triphone context by approximating it using several less context-dependent models. Such a composition was developed in order to avoid the crucial problem of limited training data and to reduce the model complexity. To enhance the model reliability due to unseen contexts and limited training data, flooring and smoothing techniques are applied. Experimental results show that the proposed Bayesian pentaphone model improves word accuracy in comparison with the standard triphone model.

key words: *Bayesian framework, wide phonetic context model, acoustic rescoring*

1. Introduction

Bayesian statistical method provides a complete paradigm for both statistical inference and decision making under uncertainty [1]. It uses a probabilistic framework which encodes our beliefs or actions in situations of uncertainty. In its simplest form, if H denotes a hypothesis and D denotes data, the Bayes' theorem states that:

$$p(H|D) = \frac{p(D|H)p(H)}{p(D)}, \tag{1}$$

where $p(H|D)$ is the probabilistic statement of belief about H after obtaining D or the so-called posterior conditional distribution, and $p(H)$ is regarded as a probabilistic statement of belief about H before obtaining data D or the so-called prior distribution. Having specified $p(D|H)$ and $p(D)$, the mechanism of the theorem provides a solution to the problem of how to learn from data [2].

Based on the posterior distribution estimation, this enables the selection of an appropriate model structure which excludes over-trained models. It offers a robust classification based on a predictive posterior distribution, which mitigates the effects of over-training [3]. Information from several models can also be combined based on the Bayesian framework to achieve better inference and to better account for modeling uncertainty [4]. By utilizing these benefits, the Bayesian framework can help in many application areas, especially where the problem is uncertain and the available data is limited.

In current automatic speech recognition (ASR) systems, there are still many challenges to overcome before they can reach their full potential through widespread use in everyday life. One of the shortcomings of the mainstream acoustic modeling is its limited capability to handle the coarticulation effects that exist in conversational speech. The triphone acoustic unit, which includes the immediate preceding and following phonetic contexts, is the most widely used in current acoustic models. 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 [5], it was found that a vowel may influence not only the preceding consonant but also the vowel before the consonant. Other studies also found that English consonants such as /l/ and /r/ exert long-distance coarticulation effects across syllables, or "resonance" [6], [7]. Thus, by incorporating something wider than the triphone context, more than just one preceding and one following phonetic contexts are taken into account. The performance of such an acoustic model is expected to improve.

Many researchers have tried to improve acoustic models by incorporating a wider-than-triphone context, such as a tetraphone, quinphone/pentaphone, or more [8], [9]. The IBM, Philips/RWTH, and AT&T LVCSR systems have been quite successful in using pentaphone models [10]–[12]. 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 non-robust 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 [13]. For large-scale systems, 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 [14].

In essence, incorporating wider-than-triphone-context units often leads to additional improvement, but it requires a large amount of training data and makes the training and

Manuscript received July 11, 2005.
Manuscript revised September 30, 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.3.946

decoding difficult. On the other hand, the simpler model is more reliable but less precise in capturing the coarticulation effects. Therefore, an efficient modeling of the wide-context unit, which can maintain the balance between the context resolution and training data size, is an important problem that needs to be addressed for a realistic application of an ASR system.

The approach we adopted here is to utilize the benefits of the Bayesian framework, which modeling a wider-than-triphone context by approximating it using several less context-dependent models. This approach is an extension of the method proposed in [15], [16] where a triphone model is constructed from monophone and biphone models. Such a composition is developed in order to alleviate the crucial issue of limited training data. This approach allows us to model a wide phonetic context from less context-dependent models, without training the whole large model from scratch. With this composition technique, the loss of context resolution can be considerably lowered since only less context-dependent models need to be estimated. In this work, we use the conventional HMM system to generate an N-best hypothesis list. Then, we apply the Bayesian wide context models to rescore the N-best list. During the rescoring process, there might be some phonetic contexts which have not been seen during the training process. To enhance the model reliability with respect to the unseen contexts and limited training data, flooring and smoothing techniques are used.

In the next section, we briefly describe the Bayesian framework for constructing a wide phonetic context. First, we describe the original Bayesian triphone model, then the Bayesian pentaphone model including a general representation of the Bayesian wide phonetic context model. In Sect. 3, we describe approaches to enhance the model reliability with respect to unseen contexts and limited training data. A detailed explanation of deleted interpolation as one of the enhancing techniques is given in Sect. 4. The use of the Bayesian wide phonetic context model in the N-best rescoring mechanism is described in Sect. 5. Details of the experiments are presented in Sect. 6, including the results and discussion. A conclusion is drawn in Sect. 7.

2. Bayesian Wide Phonetic Context

2.1 Bayesian Triphone Model

Following the theoretical framework of [15], [16], a phone-level observation is denoted by X and a context-dependent triphone model Q is denoted by $/a^-, a, a^+/$, with a being some phone and a^- and a^+ being its preceding and following phonemes, respectively. The problem of triphonic acoustic modeling can be expressed as the estimation of the probability density function (pdf) $p(X|Q) = p(X|a^-, a, a^+)$ of X generated from triphone $/a^-, a, a^+/$. Using the Bayesian principle:

$$p(X|a^-, a, a^+) = \frac{p(X, a^-, a, a^+)}{p(a^-, a, a^+)} = \frac{p(a^-, a^+|a, X)p(a, X)}{p(a^-, a^+|a)p(a)}. \quad (2)$$

Assuming that a^- and a^+ are independent given a and X, $p(a^-, a^+|a) \approx p(a^-|a)p(a^+|a)$, $p(a^-, a^+|a, X) \approx p(a^-|a, X)p(a^+|a, X)$, and Eq. (2) becomes:

$$p(X|a^-, a, a^+) \approx \frac{p(a^-|a, X)p(a^+|a, X)p(a, X)}{p(a^-|a)p(a^+|a)p(a)}. \quad (3)$$

By multiplying both the numerator and denominator by $p(a, X)p(a)$, and applying the Bayes rule, Eq. (3) becomes:

$$\begin{aligned} & p(X|a^-, a, a^+) \\ & \approx \frac{p(a^-|a, X)p(a, X)}{p(a^-|a)p(a)} \frac{p(a^+|a, X)p(a, X)}{p(a^+|a)p(a)} \frac{p(a)}{p(a, X)} \\ & \approx \frac{p(X|a^-, a)p(a^-, a)}{p(a^-, a)} \frac{p(X|a, a^+)p(a, a^+)}{p(a, a^+)} \frac{1}{p(X|a)} \\ & \approx \frac{p(X|a^-, a)p(X|a, a^+)}{p(X|a)}. \end{aligned} \quad (4)$$

This indicates a new way of representing a triphone model by models of less context dependency, i.e., $p(X|a^-, a)$, $p(X|a, a^+)$ and $p(X|a)$, which correspond to the pdfs of the observation X given the preceding/following context biphone and context-independent monophone units, respectively.

This composition leads to a reduction of the number of context units to be estimated from N^3 to $(N^2 + N)$, without loss of context coverage, where N is the number of phones. Since the derivation of Eq. (4) is closely related to Bayesian statistics, it is called the Bayesian triphone model.

Graphically, the conventional triphone unit, where a full triphone model is trained from scratch, is shown in Fig. 1 (a), and the composition of the Bayesian triphone unit is shown in Fig. 1 (b). In this paper, to distinguish different context units and compositions, we use the following naming scheme. The conventional triphone context unit will be called C3. The triphone by composition will be called C1L2R2, since it is composed of the left/preceding biphone context unit (L2), the right/following biphone context unit (R2), and the center context independent monophone unit (C1).

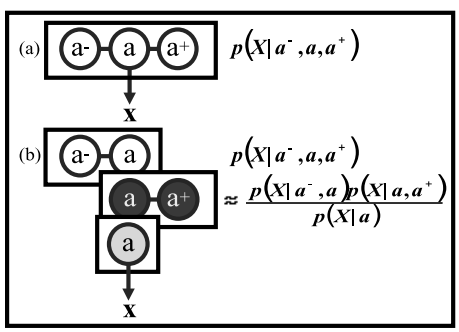

Fig. 1 Bayesian triphone model composition. (a) is C3, the conventional triphone unit and (b) is C1L2R2, the Bayesian triphone composed of the preceding/following biphone-context unit and center-monophone unit.

2.2 Bayesian Pentaphone Model

Here, we extend the approach from the previous section to compose a wider context, the pentaphone model. It includes not only the immediate preceding and following phonetic contexts, but also the second preceding and following phonetic contexts. The pdf of X generated from the pentaphone $/a^{--}, a^{-}, a, a^{+}, a^{++}/$ context unit becomes:

$$\begin{aligned} & p(X|a^{--}, a^{-}, a, a^{+}, a^{++}) \\ &= \frac{p(X, a^{--}, a^{-}, a, a^{+}, a^{++})}{p(a^{--}, a^{-}, a, a^{+}, a^{++})} \\ &= \frac{p(a^{--}, a^{-}, a^{+}, a^{++}|a, X)p(a, X)}{p(a^{--}, a^{-}, a^{+}, a^{++}|a)p(a)} \\ &\approx \frac{p(a^{--}, a^{-}|a, X)p(a^{+}, a^{++}|a, X)p(a, X)}{p(a^{--}, a^{-}|a)p(a^{+}, a^{++}|a)p(a)} \\ &\approx \frac{p(X|a^{--}, a^{-}, a)p(X|a, a^{+}, a^{++})}{p(X|a)}. \end{aligned} \tag{5}$$

The result indicates that a pentaphone model can be composed of several less context dependent models, i.e., $p(X|a^{--}, a^{-}, a), p(X|a, a^{+}, a^{++})$ and $p(X|a)$, which correspond to the pdfs of the observation X given the left/preceding-triphone-context (L3), right/following-triphone-context (R3) and center monophone base unit (C1), respectively. We call it composition C1L3R3 and the graphical representation is shown in Fig. 2 (b). If we treat the monophone unit $/a/$ as one base unit A, and the preceding context unit $/a^{--}, a^{-}/$ and following context unit $/a^{+}, a^{++}/$ as A^{-} and A^{+}, respectively, then we can derive Eq. (5) in the following way:

$$\begin{aligned} & p(X|A^{-}, A, A^{+}) \\ &\approx \frac{p(X|A^{-}, A)p(X|A, A^{+})}{p(X|A)} \\ &\approx \frac{p(X|[a^{--}, a^{-}], a)p(X|a, [a^{+}, a^{++}])}{p(X|a)} \\ &\approx \frac{p(X|a^{--}, a^{-}, a)p(X|a, a^{+}, a^{++})}{p(X|a)}. \end{aligned} \tag{6}$$

The result shows that $p(X|A^{-}, A, A^{+})$ can represent the model composition in a more general way, where A can be any context unit, and the A^{-} and the A^{+} are its one or more preceding and following contexts, respectively. Hereafter, we will use this term as a general representation and to derive other compositions of the Bayesian pentaphone model.

If we set A to be a triphone unit $/a^{-}, a, a^{+}/$, A^{-} to be the second preceding context $/a^{--}/$, and A^{+} to be the second following context $/a^{++}/$, then $p(X|A^{-}, A, A^{+})$ will become:

$$\begin{aligned} & p(X|A^{-}, A, A^{+}) \\ &\approx \frac{p(X|A^{-}, A)p(X|A, A^{+})}{p(X|A)} \\ &\approx \frac{p(X|a^{--}, [a^{-}, a, a^{+}])p(X|[a^{-}, a, a^{+}], a^{++})}{p(X|[a^{-}, a, a^{+}])} \end{aligned}$$

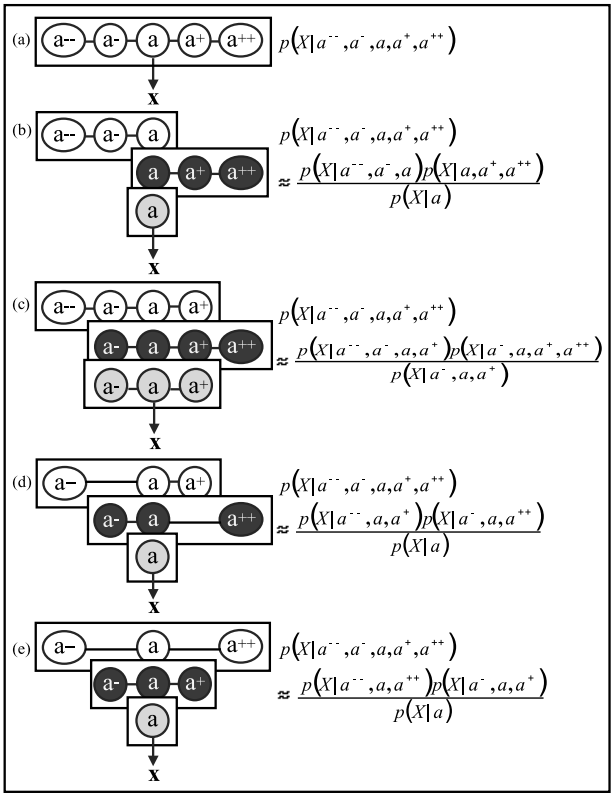

Fig. 2 Bayesian pentaphone model composition. (a) is C5, the conventional pentaphone model, (b) is Bayesian C1L3R3, which is composed of the preceding/following triphone-context unit and center-monophone unit, (c) is Bayesian C3L4R4, which is composed of the preceding/following tetraphone-context unit and center-triphone-context unit, (d) is Bayesian C1Lsk3Rsk3, which is composed of the preceding/following skip-triphone-context unit and center-monophone unit, and (e) is Bayesian C1C3Csk3, which is composed of the center skip-triphone-context unit, center triphone-context unit and center-monophone unit.

$$\approx \frac{p(X|a^{--}, a^{-}, a, a^{+})p(X|a^{-}, a, a^{+}, a^{++})}{p(X|a^{-}, a, a^{+})}. \tag{7}$$

This indicates that a pentaphone model can also be composed of $p(X|a^{--}, a^{-}, a, a^{+})$, $p(X|a^{-}, a, a^{+}, a^{++})$ and $p(X|a^{-}, a, a^{+})$ (see Fig. 2 (c)), which correspond to the pdfs of the observation X given the the left/preceding-tetraphone-context unit (L4), right/following-tetraphone-context unit (R4) and center-triphone-context unit (C3). This composition is called composition C3L4R4.

By approximating the probability distribution of composition C3L4R4 with more reduced models, such as $p(X|a^{--}, a^{-}, a, a^{+})$ with $p(X|a^{--}, a, a^{+})$, $p(X|a^{-}, a, a^{+}, a^{++})$ with $p(X|a^{-}, a, a^{++})$, and $p(X|a^{-}, a, a^{+})$ with $p(X|a)$, Eq. (7) becomes:

$$\begin{aligned} & p(X|A^{-}, A, A^{+}) \\ &\approx \frac{p(X|A^{-}, A)p(X|A, A^{+})}{p(X|A)} \\ &\approx \frac{p(X|a^{--}, a^{-}, a, a^{+})p(X|a^{-}, a, a^{+}, a^{++})}{p(X|a^{-}, a, a^{+})} \\ &\approx \frac{p(X|a^{--}, a, a^{+})p(X|a^{-}, a, a^{++})}{p(X|a)}. \end{aligned} \tag{8}$$

This approximation gives another way of composing pentaphone models, where a pentaphone model is composed of $p(X|a^{--}, a, a^{+})$, $p(X|a^{-}, a, a^{++})$ and $p(X|a)$. These correspond to the pdfs of the observation X given the left/preceding-skip-triphone-context (Lsk3), right/following-skip-triphone-context (Rsk3) and center monophone base unit (C1), respectively. The composition in Eq. (8) is called composition C1Lsk3Rsk3, which is shown in Fig. 2 (d).

The above algorithms, such as compositions C1L3R3, C3L4R4, and C3Lsk3Rsk3, always follow the general representation $p(X|A^{-}, A, A^{+})$, where the wide context model $/A^{-}, A, A^{+}/$ is composed of left context dependent $/A^{-}/$, right context dependent $/A^{+}/$, and center base context unit $/A/$. However, there can be some alternatives for composing wide context models other than those described above. For example, a wide context model is composed of several less context dependent models, where the center base unit $/A/$ in each model is the center point of each phonetic context. Then, the pdf of X generated from the pentaphone $/a^{--}, a^{-}, a, a^{+}, a^{++}/$ context unit can be approximated as follows:

$$\begin{aligned} & p(X|a^{--}, a^{-}, a, a^{+}, a^{++}) \\ &= \frac{p(X, a^{--}, a^{-}, a, a^{+}, a^{++})}{p(a^{--}, a^{-}, a, a^{+}, a^{++})} \\ &= \frac{p(a^{--}, a^{-}, a^{+}, a^{++}|a, X)p(a, X)}{p(a^{--}, a^{-}, a^{+}, a^{++}|a)p(a)} \\ &\approx \frac{p(a^{--}, a^{++}|a, X)p(a^{-}, a^{+}|a, X)p(a, X)}{p(a^{--}, a^{++}|a)p(a^{-}, a^{+}|a)p(a)} \\ &\approx \frac{p(X|a^{--}, a, a^{++})p(X|a^{-}, a, a^{+})}{p(X|a)}. \end{aligned} \tag{9}$$

The result indicates that a pentaphone model can be composed of $p(X|a^{--}, a, a^{++})$, $p(X|a^{-}, a, a^{+})$ and $p(X|a)$, which correspond to the pdfs of the observation X given the center-skip-triphone-context (Csk3), center-triphone-context (C3) and center monophone base unit (C1), which is called composition C1C3Csk3 (Fig. 2 (e)).

In these compositions, the number of context units to be estimated is reduced from N^5 to $(2N^3 + N)$ for composition C1L3R3, C1Lsk3Rsk3, and C1C3Csk3, and to $(2N^4 + N^3)$ for composition C3L4R4, 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 $44^5 =\sim 165$ million context units. Composition with triphone-context-units reduces the complexity to about 170 thousand context units, but composition with tetraphone-context-units reduces the complexity to only about 7.5 million context units.

3. Enhancing Model Reliability

For some phonetic contexts that have not been seen during training, the Bayesian wide context model is not able to produce any output probability during recognition. To handle this problem we simply assign a small numeric value as an output probability. Since the Bayesian wide context model score involves the output probability from several less context-dependent models, this flooring mechanism is applied for each model.

If the amount of training data is not large enough, the parameter estimation of the Bayesian wide context model $p(X|A^{-}, A, A^{+})$ may become unreliable, and so will the state output. The common approach to improve the model reliability is to apply a smoothing technique, such as back-off or interpolation smoothing. In this study, we try three different approaches:

1. "No decision":
   In this case, no smoothing technique is applied. We always accept the output value from Bayesian wide context model $p(X|A^{-}, A, A^{+})$ as the final output, so that
   $$p(X|Q) = p(X|A^{-}, A, A^{+}). \tag{10}$$
2. "Hard decision":
   Here, we only accept the output value from Bayesian wide context model $p(X|A^{-}, A, A^{+})$ when it is bigger than the output from the base model $p(X|A)$. Otherwise we fall back or to $p(X|A)$. It is similar to the back-off technique, but in this case, the back-off weight is just 0 or 1.
   $$p(X|Q) = \begin{cases} p(X|A^{-}, A, A^{+}), \\ \qquad \text{if } p(X|A^{-}, A, A^{+}) \geq p(X|A) \\ p(X|A), \\ \qquad \text{otherwise} \end{cases} \tag{11}$$
3. "Soft decision":
   Here, we use deleted interpolation, which is described in the next section.

4. Deleted Interpolation

Deleted interpolation (DI) is an efficient technique which allows us to fall back to the more reliable model when the supposedly more precise model is, in fact, unreliable [17]. The concept involves interpolating two (or more) separately trained models, one of which is more reliably trained than the other. So the interpolation model, $p(X|Q)$, is obtained as

$$p(X|Q) = \lambda p(X|Q_{precise}) + (1-\lambda)p(X|Q_{reliable}), \tag{12}$$

where λ represents the weight of the precise model, and $(1 - \lambda)$ represents the weight of the reduced, but more reliable, model. If the amount of training data is large enough, $p(X|Q_{precise})$ becomes more reliable and λ is expected to tend to 1.0. But if it is not, λ will tend to 0.0 so as to fall back to the more reliable model $p(X|Q_{reliable})$.

In our case, the Bayesian wide context model is the precise one, while the base model is the more reliable one, so Eq. (12) becomes:

$$p(X|Q) = \lambda p(X|A^{-}, A, A^{+}) + (1 - \lambda)p(X|A). \tag{13}$$

The optimal value of interpolation weights can be estimated using a development set rather than training or using the cross-validation method [17]. In this method, the training data is divided into M parts, and models are trained from each combination of $M - 1$ parts, with the deleted part serving as development data to estimate the interpolation weights. These M sets of interpolation weights are then averaged to obtain the final weight.

5. The Use of the Bayesian Wide Phonetic Context Model

To avoid decoding complexity, Bayesian wide context models are applied by rescoring an N-best list generated by a standard decoding system without any modification. The block diagram of the rescoring procedure is shown in Fig. 3.

For each utterance in the test data, an N-best recognition (on word level) of the baseline system is performed using a conventional HMM model and standard two-pass Viterbi decoding. Each N-best hypothesis includes an acoustic score, a language model (LM) score and a Viterbi segmentation of each phoneme. Then every phoneme segment in each hypothesis is rescored using the wider context model as shown in Fig. 4. In each rescoring, we applied a "no decision", "hard decision" and "soft decision" mechanism to enhance the model reliability (see Sect. 3).

At the beginning/end of the utterance, all left/right contexts will be filled by silence. Since we assume that there is no long silence between adjacent words, the last phonetic context from the previous word will also affect the beginning phonetic context from the current word. Thus, this rescoring mechanism is performed the same way for each segment within and in-between words (cross word model). Then, the new scores are combined with the LM score for this hypothesis. The hypothesis achieving the highest total utterance score among the N-best is selected as the new recognition output.

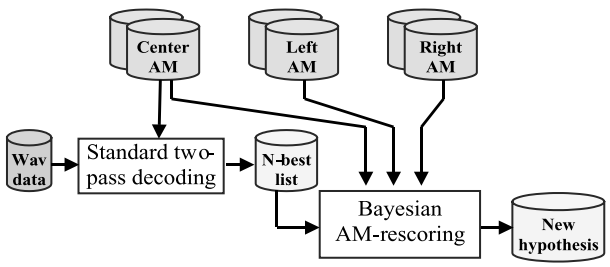

Fig. 3 Rescoring procedure.

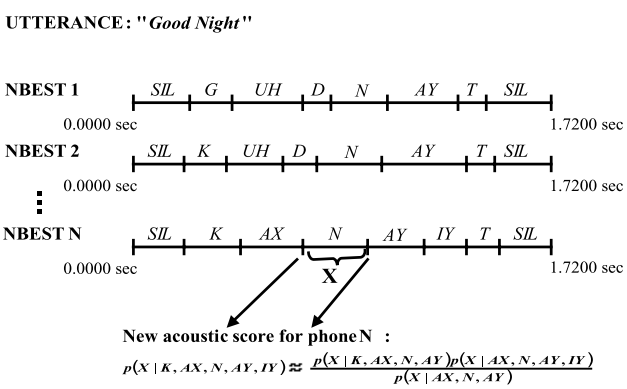

Fig. 4 N-best rescoring mechanism.

6. Experimental Results and Discussion

As training data, we use more than 60 hours of native English speech data from the Wall Street Journal (WSJ0 and WSJ1) speech corpus [18]. A sampling frequency of 16 kHz, a frame length of a 20-ms Hamming window, a frame shift of 10 ms, and 25 dimensional feature parameters consisting of 12-order MFCC, Δ MFCC and Δ log power are used as feature parameters. Three states were used as initial HMM for each phoneme. Then, a shared state HMnet topology was obtained using a successive state splitting (SSS) training algorithm. Since the SSS training algorithm used here based on the minimum description length (MDL) optimization criterion, the number of shared HMM states is determined automatically by the algorithm. Details about MDL-SSS can be found in [19]. Each component model of the Bayesian wide-phonetic context models is trained separately using the same SSS training algorithm, the same amount of training data, and the same 15 Gaussian mixture component numbers per state as our optimum choice.

The performance of the models was tested on the ATR Basic Travel Expression Corpus (BTEC) [20], which is quite different from the training corpus. The full BTEC test set1 consists of 4,080 read speech utterances spoken by 40 different speakers (20 Males, 20 Females). In this study, in order to reduce the training time, we simply selected 1,000 utterances spoken by 20 different speakers (10 Males, 10 Females), and used them as a development set to find the optimum λ parameter of the deleted interpolation. Two hundred randomly selected utterances spoken by 40 different speakers (20 Males, 20 Females) were used as a test set.

In the first experiment, a context independent monophone system with 132 total states was used as a baseline. Rescoring was done with the Bayesian triphone C1L2R2 (with 2,700 states, sum of C1: 132 st., L2: 1,313 st., and R2: 1,255 st.) as described in Sect. 5. For comparison, we also rescored using conventional biphone C2 (with 1,313 states) and triphone C3 (with 2,009 states) models. In each rescoring, we applied a "no decision", "hard decision" and "soft decision" mechanism (see Sect. 3). The recognition results for all models, obtained by each decision mechanism, are shown in Fig. 5. Bayesian C1L2R2 could achieve a significant improvement of up to 5.6% relative to the baseline. Its performance is better than just biphone C2, but still worse than the triphone C3. According to our preliminary experiments, the WSJ database is more or less suitable to train a conventional triphone model without losing the context resolution, and the optimum model was the one with about 2,000–3,000 total states and 15 Gaussian mixture components. Thus, C3 with 2,009 states and 15 mixture compo-

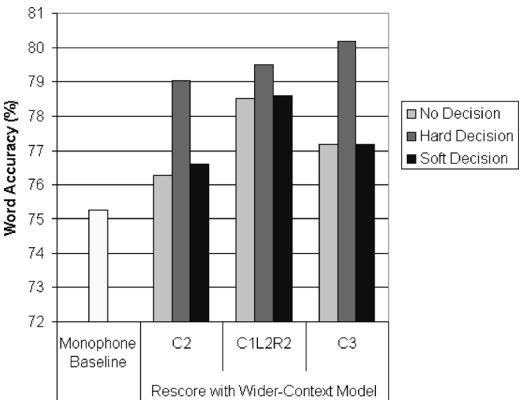

Fig. 5 Recognition accuracy rates of the Bayesian triphone model.

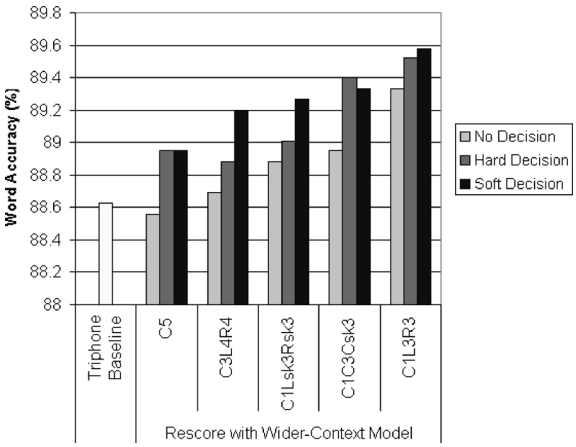

Fig. 6 Recognition accuracy rates of the Bayesian pentaphone models.

nents per state is optimum in terms of parameter number and context resolution. That might be the reason why it yields the best result.

Their best performances were obtained using the "hard decision" mechanism. They are better than the performance using "no decision" and "soft decision" with the optimal weight parameter $\lambda = 0.5$. This might be due to the following reasons. Considering the amount of training data and the number of parameters, the triphone model is optimum and much more precise than the monophone model. But it might give an unreliable estimation if there are some unseen phonetic contexts in the testing data. Thus, the "hard decision" or back-off smoothing became the optimum choice, since it only falls back to the monophone model if the output from the triphone model is unreliable. On the other hand, using the "no decision" mechanism, which always accepts the output value from the triphone model, may contain some unreliable outputs due to unseen contexts, and using the "soft decision" mechanism always interpolating the triphone model with the monophone model with equal weight ($\lambda = 0.5$) may hurt the recognition accuracy of the triphone model.

Next, we experimented with wider context models, where we used the context-dependent triphone system with 2,009 total states as the baseline to generate new N-best lists for rescoring. We tested four types of Bayesian pentaphone models: C1L3R3, C3L4R4, C1Lsk3Rsk3, and C1C3Csk3, which have a symmetric composition and simpler implementation than other possible Bayesian compositions. Those models are composed as described in Sect. 2.2. The C1L3R3 model has 3,175 states (sum of C1: 132 st., L3: 1,524 st., R3: 1,519 st.), the C3L4R4 model has 6,052 states (sum of C3: 2,009 st., L4: 2,021 st., R4: 2,022 st.), the C1Lsk3Rsk3 model has 3,333 states (sum of C1: 132 st., Lsk3: 1,587 st., Rsk3: 1,614 st.), and the C1C3Csk3 model has 3,250 states (sum of C1: 132 st., C3: 2,009 st., Csk3: 1,109 st.). As a comparison, we also rescored with a conventional full pentaphone model C5 with 2,040 total states, trained from scratch. The recognition results for all models, obtained by each decision mechanism, are shown in Fig. 6. The result shows that all pentaphone models could also achieve improvement relative to the baseline. Here, the conventional pentaphone C5 gives a worse performance than the Bayesian pentaphone models. This might be because of the following reason. Given the amount of WSJ training data, the optimum pentaphone model achieved with the MDL-SSS algorithm has 2,040 total states, which is not so different from the total number of states in triphone C3. It seems that there are many different pentaphone contexts sharing the same Gaussian components, so that the context resolution is reduced. Thus approximating a pentaphone model using the Bayesian composition of several less context-dependent models such as triphone models could help to reduce the loss of context resolution and improve the performance.

Of the Bayesian pentaphone models, the C1L3R3 model gives the best result and the worst is the C3L4R4 model. The reason for this might be that the WSJ training data is also not enough to properly train the model parameters of the tetraphone components L4 and R4. Their total number of states are only slightly different than the total number of states of triphone C3. So, as it happened in pentaphone C5, there might be many tetraphone contexts which share the same Gaussian components and the context resolution is reduced. Another reason might be that the triple phoneme overlap between the L4 and R4 component models is too big, so developing a composition among them could not give an optimum solution. So these might be the reasons why the C3L4R4 became the worst. But the other Bayesian C1L3R3, C1Lsk3Rsk3, C1C3Csk3 models basically have a similar number of total states and the total amount of training data would be enough to train the triphone contexts. All of them also only have a single phoneme overlap between the model components. However, considering the context phonetic dependency, the dependency between adjacent phonetic contexts may have much stronger effects than the dependency between skipped phonetic contexts. This mean that the more adjacent phonetic contexts they have, the better the model. Thus, C1C3Csk3 is better than C1Lsk3Rsk3, and the C1L3R3 model is the best among all models.

In this case, the best performance was obtained by the

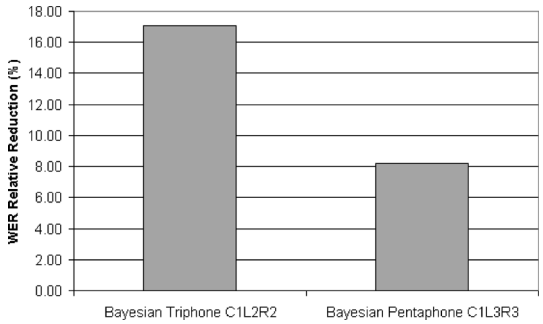

Fig. 7 A word error rate (WER) relative reduction of the Bayesian triphone C1L2R2 model with respect to the monophone baseline and a WER relative reduction of Bayesian pentaphone C1L3R3 model with respect to the triphone baseline.

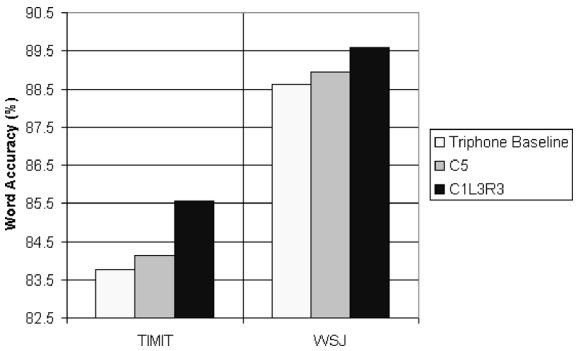

Fig. 8 Recognition accuracy rates of the conventional pentaphone C5 and the proposed Bayesian pentaphone C1L3R3 models with different amounts of training data.

"soft decision" mechanism using deleted interpolation. This shows that, when the estimation of the pentaphone model is less reliable, it is useful to interpolate the pentaphone model and the triphone model estimations, because the triphone model can often provide useful information. The optimal weight parameter λ was about 0.3. Having a weight factor of 0.3 means that the contribution of the pentaphone model is only about 30% of the total score.

Figure 7 shows a comparison between a word error rate (WER) relative reduction of the Bayesian triphone C1L2R2 model with respect to the monophone baseline and a WER relative reduction of the Bayesian pentaphone C1L3R3 model with respect to the triphone baseline. The error rate reduction of the Bayesian pentaphone model is smaller (about half of the error rate reduction of the Bayesian triphone model), probably due to the following reasons. First, the coarticulation effect from the second preceding and following contexts is less than the coarticulation effect from the first preceding and following contexts. Second, the variations in the read speech data due to longer coarticulation effects might be less than in conversational speech. This can also be seen from the weight factor of the deleted interpolation, which can be interpreted as a confidence factor with 30% only. However, with this relatively small contribution, the results show that it still can help to improve the recognition performance.

To show the consistency of the effect of using the Bayesian composition, we did another evaluation of experiments on fewer training data. Here, we chose the TIMIT acoustic-phonetic continuous speech corpus [21] as another American-English, phonetically-rich corpus, but smaller than the WSJ database corpus. It contains only about seven hours of read speech (6,300 utterances in total). Each component acoustic model was trained using the SSS algorithm as before. In this case, the triphone baseline has 434 states, the conventional pentaphone C5 has 440 states, and the proposed Bayesian pentaphone C1L3R3 has 850 states (sum of C1: 132 st., L3: 369 st., R3: 349 st.). These models were tested using the same BTEC test set with "soft decision" only. The optimal weight parameter λ was also 0.3. The results are shown in Fig. 8. As can be seen, with fewer training data, the performance difference between the proposed C1L3R3 model and the conventional pentaphone C5 model became more significant.

7. Conclusion

We have demonstrated the possibility of improving acoustic model performance by incorporating a wide phonetic context based on the Bayesian framework. This method allows us to construct wider context models from several other models that have a narrower context. This composition technique leads to a reduction of the number of context units to be estimated, so the loss of context resolution can be considerably reduced since only less context-dependent models need to be estimated. We apply these wide-context-model compositions at the post-processing stage with N-best rescoring, so we can use the standard decoding system without any modification. The recognition results showed that ASR system performance can be improved by rescoring with Bayesian wide-phonetic-context models.

Acknowledgement

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

References

[1] J. Bernardo, "Bayesian statistic," UNESCO Encyclopedia of Life Support Systems (EOLSS), ftp://matheron.uv.es/pub/personal/bernardo/BayesStat.pdf, 2001.
[2] J. Bernardo and A. Smith, Bayesian Theory, John Wiley & Sons, Chichester, West Sussex, England, 1994.
[3] S. Watanabe, Y. Minami, A. Nakamura, and N. Ueda, "Variational Bayesian estimation and clustering for speech recognition," IEEE Trans. Speech Audio Process., vol.12, no.4, pp.365–381, 2004.
[4] D. Heckerman, C. Meek, and G. Cooper, "A Bayesian approach to causal discovery," Tech. Rep. MSR-TR-97-05, Microsoft Research, Advanced Technology Division, Microsoft Corporation, Redmond, WA, USA, 1997.
[5] E. Scripture, The Elements of Experimental Phonetics, Charles Scribners Sons, New York, USA, 1902.

[6] 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.
[7] 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.
[8] M. Finke and I. Rogina, "Wide context acoustic modeling in read vs. spontaneous speech," Proc. ICASSP, pp.1743–1746, Munich, Germany, 1997.
[9] 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.
[10] 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.
[11] 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.
[12] 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
[13] 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.
[14] 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.
[15] 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 Audio Process, vol.7, no.6, pp.678–684, Nov. 1999.
[16] J. Ming and F.J. Smith, "A Bayesian triphone model," Comput. Speech Lang., vol.13, pp.195–206, 1999.
[17] X. Huang, A. Acero, and H.W. Hon, Spoken Language Processing, Prentice Hall, New Jersey, USA, 2001.
[18] 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.
[19] 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.
[20] 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.
[21] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren, "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM," Technical Report NISTIR 4930, NIST, 1993.

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 Communication Technology at the University of Ulm, Germany, and received her MSc degree in 2002. During her thesis work, she was an intern student at the 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.

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.

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.