Incorporating a Bayesian Wide Phonetic Context Model for Acoustic Rescoring

Sakriani Sakti, Satoshi Nakamura, Konstantin Markov

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

Abstract

This paper presents a method for improving acoustic model precision by incorporating wide phonetic context units in speech recognition. The wide phonetic context model is constructed from several narrower context-dependent models based on the Bayesian framework. Such a composition is performed in order to avoid the crucial problem of a limited availability of training data and to reduce the model complexity. To enhance the model reliability due to unseen contexts and limited training data, flooring and deleted interpolation techniques are used. Experimental results show that this method gives improvement of the word accuracy with respect to the standard triphone model.

1. Introduction

Coarticulation effect is a term where phonemes can have very different waveforms when produced in the context of other phonemes [1]. Efficient modeling of the coarticulation effect is one important problem that needs to be addressed for realistic application of an automatic speech recognition (ASR) system. A standard solution to the coarticulation problem in speech is to extend the analysis units to include context [2]. A triphone unit which includes the immediate preceding and following phonetic contexts is most widely used in the current ASR systems.

Although such triphones have proved to be an efficient model, a wider phonetic context seems to be appropriate for capturing coarticulation effects more precisely. By incorporating something wider than the triphone context, such as a pentaphone (or more), more than just one preceding and one following phonetic context dependencies are taken into account, so the performance of such an acoustic model is expected to improve. This, however, faces the general problem of un-robust parameter estimation due to the limited speech training data and increased numbers of unseen context. As a consequence, the complexity of the model, as well as the number of parameters that need to be estimated may increase dramatically.

Some researcher have tried to use wider-than-triphone units, such as syllables or multi-phone units that give better overall recognition rates [3, 4], but difficulties arise in using them properly with conventional decoding. Another study proposed compiling wide context dependent models into a network of Weighted Finite State Transducers (WFT) [5], so that the decoding process is completely decoupled from dealing with the wide context. However, when higher order models are used, difficulties lie in the compilation it self. The work in [6] then tried to simplify the compilation method. For large scale systems, a simple procedure is to apply wide context models in rescoring pass.

The approach we adopted in this paper is to model a wider-than-triphone context by approximating it using several less context-dependent models. This approach is an extension of the method proposed in [7, 8] where a triphone model is constructed from monophone and biphone models. Such a composition is performed in order to alleviate the crucial issue of limited training data. The approach allows us to model wide phonetic context from less context-dependent models, without training the whole large model from scratch. In this work, we use the standard decoding system with no modification to generate a N-best hypotheses list. Then, we apply the proposed approach to acoustic rescoring. It is a process where decoding will use knowledge sources of progressively increasing complexity to decrease the size of the search space [9]. First it will use the conventional hidden Markov model (HMM) to find the N-best sequences word based on the Viterbi algorithm, then it will rescore them using a wider context model. To enhance the model reliability due to unseen contexts and limited training data, flooring and deleted interpolation techniques are used.

In the next section, we briefly describe the Bayesian framework for constructing a wide phonetic context. The approaches to enhance the model reliability are described in Section 3 and detail explanation of deleted interpolation technique is described in Section 4. The acoustic rescoring mechanism is described in Section 5. Details of experiments are presented in Section 6, including results and discussion. A conclusion is drawn in Section 7.

2. Bayesian Wide Phonetic Context

Following to the theoretical framework of [7, 8], 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 $/a^-, a, a^+/$. Using the Bayesian principle

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

It is assumed that a^- and a^+ are independent given a and X, so that $p(a^-, a^+|a) = p(a^-|a)p(a^+|a)$ and $p(a^-, a^+|a, X) = p(a^-|a, X)p(a^+|a, X)$. Using the Bayes rule, Equation (1) can

be easily transformed to:

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

This equation 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-context, following-context and context-independent units, respectively. Since the derivation of Equation (2) is closely related to Bayesian statistics, it is called the Bayesian triphone model.

Using similar consideration, we extend this framework into a wider context, such as a pentaphone model. It includes not only the immediate preceding and following phonetic contexts, but also the second preceding and following phonetic context. The pdf of X generated from the pentaphone $/a^{--}, a^-, a, a^+, a^{++}/$ becomes:

$$\begin{aligned} p(X \quad | \quad & a^{--}, a^-, a, a^+, a^{++}) \\ = \quad & \frac{p(X, a^{--}, a^-, a, a^+, a^{++})}{p(a^{--}, a^-, a, a^+, a^{++})} \\ = \quad & \frac{p(X|a^{--}, a^-, a)p(X|a, a^+, a^{++})}{p(X|a)} \end{aligned} \quad (3)$$

The result indicates that a pentaphone model can be decomposed into models of less context dependency, 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 preceding-triphone-context, following-triphone-context and center unit, respectively. In this case the center unit is still a context-independent (monophone) unit. In order to have a triphone unit as a center base, the overlapping between preceding-context and following-context should perform a triphone context (see Figure 1). The pdf of X generated from $/a^{--}, a^-, a, a^+, a^{++}/$ becomes:

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

In this case a pentaphone model is composed by the preceding-tetraphone-context unit, following-tetraphone-context unit and center-triphone-context unit.

If we treat the triphone unit $/a^-, a, a^+/$ as one base unit A, and the second preceding/following contexts is the immediate context from our base unit A (A^- and A^+), then:

$$\begin{aligned} p(X \quad | \quad & a^{--}, a^-, a, a^+, a^{++}) \\ = \quad & \frac{p(X|a^{--}, a^-, a, a^+)p(X|a^-, a, a^+, a^{++})}{p(X|a^-, a, a^+)} \\ = \quad & \frac{p(X|a^{--}, [a^-, a, a^+])p(X|[a^-, a, a^+], a^{++})}{p(X|[a^-, a, a^+])} \\ = \quad & \frac{p(X|A^-, A)p(X|A, A^+)}{p(X|A)} \\ = \quad & p(X|A^-, A, A^+) \end{aligned} \quad (5)$$

The same result will happen if we treat the monophone unit $/a/$ as one base unit A, the preceding context unit $/a^{--}, a^-/$ and the following context unit $/a^+, a^{++}/$ as A^- and A^+, respectively. Now, we can see the basic formula in a more general way, where A can be any context unit of the acoustic modeling, 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.

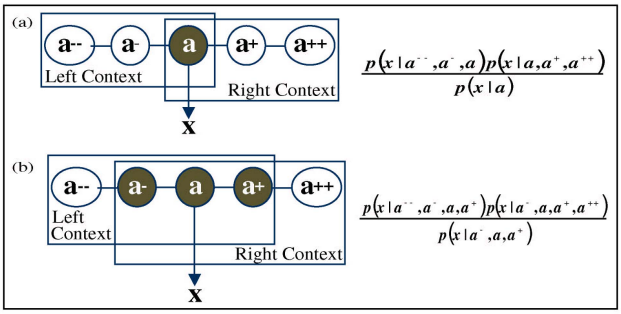

Figure 1: *Bayesian pentaphone model composition. (a) is composed of the preceding/following triphone-context unit and center-monophone unit and (b) is composed of the preceding/following tetraphone-context unit and center-triphone-context unit.*

3. Enhancing Model Reliability

During the rescoring process, there might be some phonetic contexts which have to be estimated but have not been seen during the training process. For such contexts, during recognition the Bayesian wide context model is not able to produce any output probability. To handle this problem, in such cases we simply assign a small numeric value as an output probability. Since the Bayesian wide context model rescoring involves the output probability from preceding, following and center model, 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 be the state output. In this study, to improve the model reliability we try three different approaches:

1. No decision:
   Always accept the output value from Bayesian rescoring $p(X|Q) = p(X|A^-, A, A^+)$ as the final output.
2. Hard decision:
   Accept $p(X|A^-, A, A^+)$ when it is bigger than $p(X|A)$ which is the output from the base model. Otherwise we fall back to $p(X|A)$.
3. Soft decision:
   Use deleted interpolation 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 [10]. 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

$$p(X|Q) = \lambda p(X|A^-, A, A^+) + (1 - \lambda)p(X|A) \quad (6)$$

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, $p(X|A^-, A, A^+)$ becomes more reliable and we expect λ to tend to 1.0. But if it is not, we expect λ to tend to 0.0 so as to fall back to the more reliable model $p(X|A)$. The optimal value of interpolation weights are estimated using development set other than training or using the cross-validation method [10]. 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 unseen data to estimate the interpolation weights. These M sets of interpolation weights are then averaged to obtain the final weights. In this study, in order to reduce the training time, we estimated the interpolation weights using a development set.

5. Acoustic Rescoring Mechanism

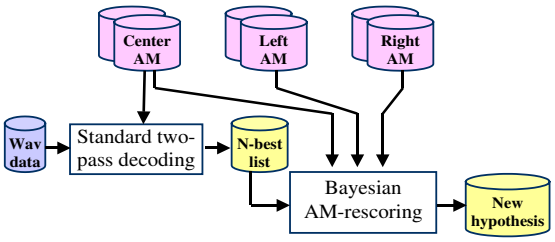

Figure 2: *Rescoring procedure.*

The block diagram of the rescoring procedure is shown in Figure 2. On each utterance in the test data, a N-best recognition (on word level) is performed using a conventional HMM model and a standard two-pass decoding system based on Viterbi algorithm without modification. The system will result N-best hypothesis list including the acoustic score, the language modeling (LM) score and the Viterbi segmentation of each phoneme. Then for every phoneme segment in each hypothesis, we rescore the acoustic precision using wider context model as shown in Figure 3. These updated scores is combined with LM score for this hypothesis. Then the hypothesis achieving the highest total utterance score among the N-best is selected as the new recognition output.

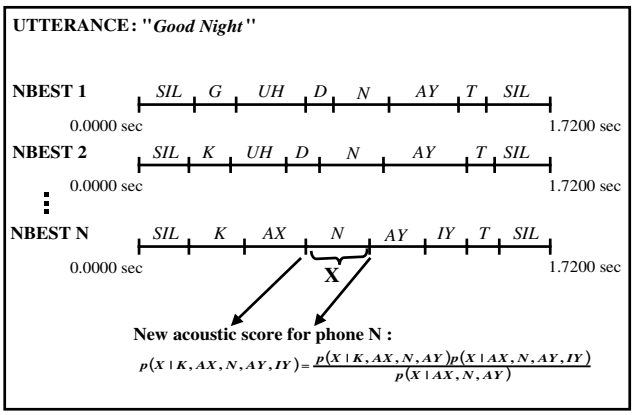

Figure 3: *N-best rescoring mechanism.*

6. Experimental Results and Discussion

Our baseline triphone HMM acoustic model is trained on more than 60 hours of native English speech data from the Wall Street Journal (WSJ0 and WSJ1) speech corpus [11]. A sampling frequency of 16 kHz, a frame length of a 20-ms Hamming window, a frame shift of 10 ms, and 25 dimensional feature parameters consisting of 12-order MFCC, Δ MFCC and Δ log power are used as feature parameters. Three states were used as the initial model for each phoneme. Then, a state level HMnet is obtained using a successive state splitting (SSS) algorithm based on the minimum description length (MDL) criterion in order to gain the optimal structure in which triphone contexts are shared and tied at the state level. Details about MDL-SSS can be found in [12]. Three acoustic models, one for each preceding-context unit $/A^-, A/$, following-context unit $/A, A^+/$ and center-context unit $/A/$, are trained separately.

The performance of this Bayesian rescoring approach was tested on the ATR Basic Travel Expression Corpus (BTEC)[13], which is quite different than 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, we select 1,000 utterances spoken by 20 different speakers (10 Males, 10 Females) used as a development set to find the optimum λ parameter of deleted interpolation. And we randomly selected 200 utterances spoken by 40 different speakers (20 Males, 20 Females) used as a test set.

First, we performed Bayesian triphone rescoring where the baseline system was a context independent monophone system. The Bayesian acoustic rescoring, named rescore-C1L2R2 meaning that it composed from left/preceding-biphone-context unit (L2), right/following-biphone-context unit (R2) and center-monophone-context unit (C1), was done as described in Section 5. As a comparison, we also rescored using conventional biphone and triphone model, named rescore-C2 (center-biphone-context unit) and rescore-C3 (center-triphone-context unit), respectively. For each rescoring, we applied a no decision, hard decision and soft decision mechanism (see Section 3). The best recognition results obtained by each rescoring method are summarized in Figure 4.

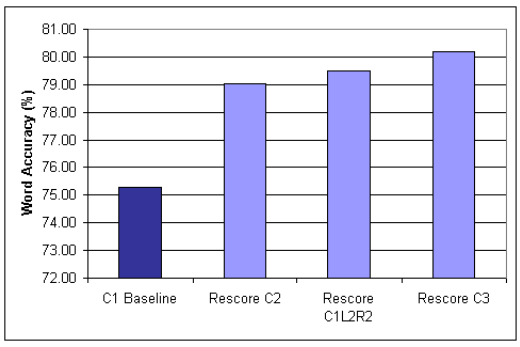

Figure 4: *Recognition results of monophone rescoring.*

In this case, the best performance from each method is obtained by the hard decision mechanism. The result shows that Bayesian rescore-C1L2R2 achieve an improvement of up to 5.6% relative to the baseline. Its performance is better than just rescoring with biphone-context model (rescore-C2). Since the training data is large enough to train the triphone model, the rescore-C3 thus yields the best result among them.

Now we extend this framework into a wider context, to perform Bayesian pentaphone rescoring where the baseline system is a context-dependent triphone system. As described in Section 2, we will have two types of Bayesian pentaphone rescoring, one that uses the left/preceding-triphone-context (L3), the right/following-triphone-context (R3) and the center-monophone-context unit (C1), named as rescore-C1L3R3, and the other one that uses the left/preceding-tetraphone-context (L4), the right/following-tetraphone-context (R4) and the center-triphone-context unit (C3), named as rescore-C3L4R4. As a comparison, we also rescore with a conventional pentaphone model, named as rescore-C5, where we trained a full pentaphone model from scratch. The best recognition results obtained by each rescoring method are summarized in Figure 5.

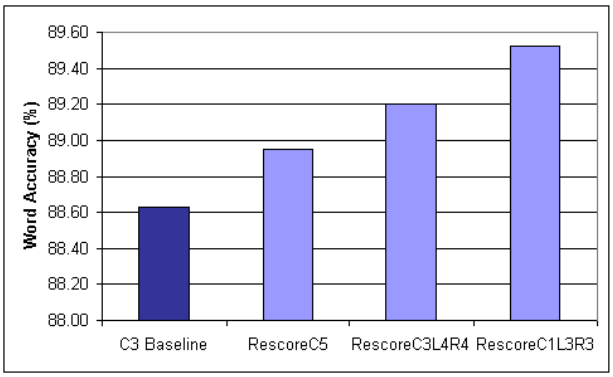

Figure 5: *Recognition results of triphone rescoring.*

In this case, the best performance from each method is obtained by the soft decision mechanism using deleted interpolation. The average of the weight parameter is about 0.3. The result shows that the Bayesian rescore-C1L3R3 and rescore-C3L4R4 could also achieved improvement relative to the baseline. The results of rescore-C5 are worse than applying the Bayesian rescoring technique. Here, the improvement is not as much as in monophone rescoring, 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 training data are not enough to train the full pentaphone model. This can be seen also from the weight factor of the deleted interpolation, which can be interpreted as a confidence factor. Having a weight factor of 0.3 means the contribution of the pentaphone model is only about 30% of the total score.

7. Conclusion

We have demonstrated the possibility to improve acoustic model performance by incorporating wide phonetic context based on Bayesian framework. This method allows to construct wider context models from several other models that have narrower context. We also can use the standard decoding system without any modification, since the new model is applied at the post-processing stage by N-best rescoring. The recognition results show that ASR system performance is improved after Bayesian rescoring.

8. Acknowledgement

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

9. References

[1] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, New Jersay, USA, 1993.

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

[3] I. Shafran and M. Ostendorf, "Use of higher level linguisitic structure in acoustic modeling for speech recognition," in *Proc. ICASSP*, Istanbul, Turkey, 2000, pp. 1021–1024.

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

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

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

[7] Ji Ming, P. O Boyle, M. Owens, and F. Jack 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, November 1999.

[8] Ji Ming and F. Jack Smith, "Improved phone rescognition using Bayesian triphone models," in *Proc. ICASSP*, Seatle, USA, 1998, pp. 409–412.

[9] 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, 2003.

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

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

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

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