LVCSR Based on Context-Dependent Syllable Acoustic Models

Jian ZHANG†, Longbiao WANG†, and Seiichi NAKAGAWA†

†Department of Information and Computer Sciences, Toyohashi University of Technology, Japan
E-mail: †{zhangjian, wang, nakagawa}@slp.ics.tut.ac.jp

Abstract We propose an effective and accurate inter-word context-dependent modeling for large vocabulary continuous speech recognition (LVCSR). As well known, intra-word context-dependent modeling can be realized by describing the context-dependent syllables in the dictionary. However, it usually suffers from the limitation of less accuracy because it does not model inter-syllable pronunciation variations. In our laboratory, a combinational use of linear lexicon and tree-structured lexicon in a 1-best approximation search algorithm for LVCSR was proposed. We only need to make branches for the head syllable according to the contexts and the paths are merged at the second syllable for the linear lexicon. For the tree-structured lexicon, branches are made in a similar way. At the end node of a word, the language scores have to be compensated considering the inter-word context, but the scores of contexts other than that of the best history are lost because of the merge at the second syllable. To solve this problem, we introduce the ’likelihood difference index’. We also investigate the effect of rescoring of the acoustic model (AM) and the language model (LM) in the 2nd pass. The proposed algorithms were evaluated on JNAS and CSJ corpora. The proposed algorithms obtained a remarkable improvement of recognition performance, and the rescoring of the context-dependent syllable acoustic models in the 2nd pass mode also achieved a further improvement even the same acoustic models were used in the 1st pass.

Key words LVCSR, context-dependent acoustic models, acoustic and language model rescoring, linear lexicon, tree-structured lexicon

1 Introduction

Large vocabulary continuous speech recognition (LVCSR) based on hidden Markov model (HMM), acoustic modeling and N-gram language modeling has been developed to a practical level and applied to practical systems such as dictations and broadcast news captioning, and further improvement of the technology will make it be used more widely.

HMM-based LVCSR generally has to take computational cost proportional to its vocabulary size if all the HMM states of the words in the vocabulary are expanded on the memory. A tree-structured lexicon in which prefixes expressed in sub-word units such as phones or syllables are shared among words is often used to reduce the HMM states [1]. In our laboratory, an efficient and accurate search algorithm using the 1-best approximation tree-structured lexicon search and linear lexicon search was proposed [4] [5]. Context-independent HMMs (CIHMMs) were used as acoustic models in our LVCSR system. However, these CIHMMs do not model the various pronunciation variations, so the speech recognition performance was degraded.

In this paper, we propose an effective and accurate inter-word context dependent modeling for our LVCSR. A two-pass context-dependent HMMs based LVCSR is very easy to realize. Therefore, context-independent HMMs were used in the first pass. And then, in the second pass, rescoring using context-dependent HMMs was applied on N-best hypotheses generated by the first pass. As well known, intra-word context-dependent modeling in an one-pass mode can be realized by describing the context-dependent HMMs in the dictionary. However, it usually suffers from the limitation of less accuracy because it does not model various inter-HMM pronunciation variations. The inter-word context-dependent modeling was implemented corresponding to the linear lexicon search and the 1-best approximation tree-structured lexicon search. We only need to make branches for the head syllable according to the contexts and the paths are merged at the second syllable for the linear lexicon. For the tree-structured lexicon, branches are made in a similar way. At the end node of a word, the language scores have to be compensated considering the inter-word context, but the scores of contexts other than that of the best history are lost because of the merge at the second syllable. To solve this problem, we introduce the ’likelihood difference index’.
We also investigate the effect of rescoring of the acoustic model and the language model in the 2nd pass mode. For language model rescoring, a traditional trigram language models are used in the 2nd pass. For acoustic model rescoring, either the context-independent syllable-based HMMs or context-dependent syllable-based HMMs is used in the 2nd pass. The Viterbi algorithm is used to calculate the optimal acoustic likelihoods of N-best hypotheses generated by the first pass. Furthermore, the same acoustic models are used in the first pass and second pass to evaluate the search precision of our LVCSR.

Section 2 overviews our LVCSR by combining the linear lexicon search algorithm with tree-structured lexicon in a 1-best approximation search algorithm. An effective and accurate inter-word context-dependent modeling is proposed in Section 3. Section 4 describes the experimental results of proposed context-dependent modeling algorithm and the effect of rescoring the AM and the LM. Finally, Section 5 summarizes the paper.

2 LVCSR system

2.1 Conventional methods

2.1.1 Linear lexicon and tree-structured lexicon

HMM-based LVCSR takes more computational cost when the vocabulary size becomes larger. When using a linear lexicon illustrated in Fig. 1 (a), the number of expanded HMM states gets large proportionally to the vocabulary size and then the computational costs also proportional to the vocabulary size. The tree-structured lexicon illustrated in Fig. 1 (b) is conventionally used to reduce the number of expanded HMM states.

When using N-gram language models, the likelihood of a hypothesis at time frame i which is generated by connecting a word w to the word sequence W is:

$$P_{W,w}(i) = \max_j \{ P_{W,j} + Q_w(j,i) + \text{lang}(w|W) \}, \quad (1)$$

where \( j < i \), \( P_{W,j} \) is the sum of the likelihoods of the hypothesis W at time j, \( Q_w(j,i) \) is the acoustic likelihood of the word w for period \([j,i]\) and \( \text{lang}(w|W) \) is the language probability of word w given the word history W. If we use a linear lexicon, \( \text{lang}(w|W) \) can be considered at the time of connection of w and W, but if the lexicon is structured as a tree, the root node and many other nodes corresponding to word prefixes are shared among words and thus \( \text{lang}(w|W) \) cannot be added to \( P_{W,w} \) at the time point on the fly.

2.1.2 Approximations of word history dependency for tree-structured lexicon search

The likelihood and boundary of a word depend on the word history. So we have to execute the forward decoding procedure independently for all new hypotheses generated by concatenating a word to all the possible hypotheses. Since many hypotheses are generated in this process, we have to make an assumption on the word history dependency and bundle the hypotheses using an approximation based on the assumption.

The word-pair approximation [3] assumes that the word boundaries depend on just one previous word. Under this assumption, Viterbi forwarding is executed on previous word-dependent HMM networks. This means that HMM networks have to be copied according to the word histories and the Viterbi search has to be executed on the networks in parallel.

On the other hand, the 1-best approximation [2] does not assume any word-history dependencies. This approximation reduces the computational cost very much because system does not need copies of the HMM network and uses only one reentrant network. We can determine the preceding word (that is, the word history) by back-tracing from the leaf of the network and thus we can apply the bigram probability \( \text{P}(w_i|w_{i-1}) \). Even if only one hypothesis is expanded to new hypotheses on the HMM network at every time frame, we can obtain approximately other hypotheses keeping other histories \( w_{i-1}s \) at each frame and compensating the scores by two ways: (1) the differences of history scores between the score of the history ended with \( w_{i-1} \) and that ended with \( w_i \), (2) the difference of language scores between \( \text{P}(w_i|w_{i-1}) \) and \( \text{P}(w_i|w_{i-1})\). But this approximation is not correct. The HMM network is shared by the hypotheses expanded at all the time frame and thus many paths are overridden and wiped off by the other path. So the path which is optimal when considering the language score which can be applied at the leaf node may be rejected by this mechanism. In this way, the optimal word sequence is not guaranteed to be obtained by tree-structured lexicon search under the 1-best approximation.

2.1.3 Recognition using linear lexicon

We compare the linear lexicon search and the tree-structured lexicon search. Linear lexicons for vocabulary words are connected each other from the tails to the heads in an HMM network. Using this network, any sub-word HMMs are not shared among words unlike the tree lexicon and deter-
mine the word associated to each node. This enables that the system applies the bigram language scores at the first node of each word whereas the tree-structured lexicon does not. This means that the system can use language look-ahead, resulting in efficient constraints of the search space.

Fig. 2 (a) shows an example of the new word attachment to hypotheses using a linear lexicon. The word "a-as-ga-o" is attached to the hypothesis D which maximizes the sum of the total score of hypothesis D and the language score $P(a-as-ga-o|D)$. On the contrary, the HMM network is attached to the hypothesis A in the case of a tree-structured lexicon as illustrated in Fig. 2 (b) because of its highest likelihood among the hypotheses not considering the word probabilities when given the histories.

The HMM trellis is shared among words and histories under the 1-best approximation method and thus conflicts often occur. The HMM nodes in linear lexicons are also shared among histories and thus the optimal N-best hypotheses are not guaranteed to be obtained, but there are no conflicts among words and the first best hypothesis is guaranteed to be obtained.

But a linear lexicon does not allow node sharing among words and thus the computational cost needed is much higher than that for a tree-structured lexicon.

2.2 Combination use of dynamically expanded linear lexicon and static tree-structured lexicon

When using bigram language models, the probabilities of word pairs which are not or rarely found in training data are smoothed by unigram back-off. The combination of the bigram and the 1-best approximation search on a tree-structured lexicon looses the optimal search path, but unigram does not.

A new search method using search on dynamically expanded linear lexicon along with the 1-best approximation search on a tree-structured lexicon was proposed in our laboratory [4][5]. That is, we basically used the 1-best approximation tree-structured lexicon search and also used the linear lexicon search in parallel. Only the small number of words $N_{lin}$, which were dynamically selected, are expanded in a linear lexicon network.

In our dynamic selection of the words to be expanded in a linear network, we evaluated all the vocabulary words at every time frame using the following equations:

$$Q_{tree}(w,t) = \max_v (P_{hyp}(v,t) + \text{lang}(w|v)), \quad (2)$$

$$Q_{lin}(w,t) = \max_{s_{w}} (P_{state}(s_{w},t)), \quad (3)$$

where $P_{hyp}(v,t)$ is the likelihood of the final HMM state of the hypothesis which ends with the word $v$, $P_{state}(s_{w},t)$ is the likelihood of the HMM state $s_{w}$ in word $w$ which is expanded on a linear HMM network at the time frame $t$ and $\text{lang}(w|v)$ is a look-ahead probability which is the appearance probability of word $w$ given a word history $v$. $Q_{tree}(w,t)$ in Equation (2) means the maximum value among the sums of likelihood of preceding word sequence ended by word $v$ in a tree-structured network and the look-ahead probability. $Q_{lin}(w,t)$ in Equation (3), which means the maximum value among the HMM states of the word $w$, is calculated for the word $w$ which is already expanded in the linear HMM network. If a word $w$ corresponds to both $Q_{tree}(w,t)$ and $Q_{lin}(w,t)$, then the maximum value of $Q_{tree}(w,t)$ and $Q_{lin}(w,t)$ is selected as the evaluation value for the word.

The $N_{lin}$-best words are selected based on the values. The words selected and not expanded on the linear network yet will be expanded. The words selected and already expanded on the network will be kept on the network. The words not selected on the linear network will be removed from the network. Using the linear lexicon, the optimal path lost by the 1-best approximation search on the tree-structured lexicon will be likely to be kept. On the other hand, the paths with low linguistic probability will be pruned out from the linear network without acoustic evaluation because of the language look-ahead, but such paths also have chance to be kept on the tree-structured network.

3 Inter-word context-dependent modeling

Because Japanese syllables consist of a consonant and vowel pair, there are relatively small number (approx. 120) of syllables and we only need to consider small variations (approx. 7 or 8 including vowels and pause) of left context. So we can use full syllable modeling considering C-V context dependency and also easily extend the CDHMMs to full left CDHMMs to consider V-C context dependency.

As well known, inter-word context-dependent modeling can be realized by describing the CD syllables in the dictionary. But inter-word modeling should be realized by recognition
process.

An approximate context-dependent modeling method is illustrated in Fig. 3. For the linear lexicon (Fig. 3 (a)), phonetic context (that is, ending of a word) of the preceding word $v$ with the maximum summation score of likelihood ending with the word $v$ and a look-ahead language probability is used to generate the context-dependent HMMs. For the 1-best approximation tree-structured lexicon, as the language probability cannot be applied on the head of the tree, phonetic context of the preceding word $v$ with the maximum score ending with the word $v$ is applied to all the head HMMs which is described in Fig. 3 (b). We call this method approximate context-dependent HMM search algorithm. Although this method is very easy to implement, the context-dependent acoustic scores do not need to be selected to the optimal preceding word. So this approximation is not correct and accurate.

To solve above problem, the context-dependent modeling method is modified as Fig. 4. As described in Fig. 4 (a), we only need to make branches for the head syllable according to the contexts and the paths are merged at the second syllable for the linear lexicon. For the tree-structured lexicon, branches are made in a similar way. At the end node of a word, the language scores have to be compensated considering the inter-word context, but the scores of contexts other than that of the best history are lost because of the merge at the second syllable. To solve this problem, we introduce the 'likelihood difference index' (Fig. 4(b)). The likelihood differences between phonetic contexts are calculated at the merging points and the difference index is inherited to the succeeding nodes. For example, Fig. 4(b) shows that $i-a$ ("a" with the left context "i") after 1st syllable matching is the best. Using this index, language scores are accurately compensated considering inter-word phonetic contexts at the end node of the words.

4 Experiments

We evaluated the effectiveness of the inter-word context-dependent modeling and that of the acoustic and language model rescoring, respectively.

4.1 Experimental conditions

To evaluate the performance, we used two kinds of data sets. The first set is 100 sentences from Japanese newspaper read speech in JNAS corpus. The second set is 4 lecture speech utterances from the Corpus of Spontaneous Japanese (CSJ) uttered by 4 male speakers at the meeting of the Acoustic Society of Japan (ASJ). The 4 lectures were described as follows: a01m0007, a01m0035, a01m0074 and a05m0031. All of speech data were sampled with a sampling frequency of 16 kHz, and the signal was pre-emphasized by a factor of 0.98. A Hamming window of 25 ms length was applied and shifted with the step of 10 ms. 38 dimensional feature vectors were used including 12 dimensional MFCCs, their first and second deviation coefficients and the first and second deviations of log power.

For JNAS evaluation data set, 27992 utterances read by 175 male speakers (JNAS corpus) were used to train 116 Japanese context-independent syllable HMMs (strictly speaking, mora HMMs[6]) including short pause and silence. For CSJ evaluation data set, 116 CSJ context-independent HMMs were adapted from the 116 JNAS context-independent HMMs by...
Table 1: Speech recognition results of JNAS based on context-dependent HMMs (%)  

<table>
<thead>
<tr>
<th>AM in the 1st pass</th>
<th>1st pass (bigram)</th>
<th>2nd pass (trigram)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1-best</td>
<td>N-best</td>
</tr>
<tr>
<td>CIHMM</td>
<td>88.7</td>
<td>86.8</td>
</tr>
<tr>
<td>Intra-word CDHMM</td>
<td>89.5</td>
<td>87.3</td>
</tr>
<tr>
<td>Approx. CDHMM</td>
<td>90.6</td>
<td>89.0</td>
</tr>
<tr>
<td>Exact CDHMM</td>
<td>91.0</td>
<td>88.6</td>
</tr>
</tbody>
</table>

MAP [7] [8] using 987 lectures from the CSJ corpus uttered by 987 male speakers. Using these context-independent HMMs (CIHMMs) as base models, we also trained 928 context-dependent HMMs (CDHMMs) with 8 left contexts (5 vowels, silence, /N/, and short pause including /q/). Each continuous density HMM had 5 states, and 4 of them had pdfs of output probability. Each pdf consisted of four Gaussians with full-covariance matrices.

For the JNAS evaluation data set, bigram and trigram language models with a 20000 word vocabulary size were trained from the text of 45 months Mainichi Japanese newspaper with 90 million words. For the CSJ evaluation data set, bigram language model with about 17000 word vocabulary size was trained using a text made by correctly transcribing lecture speech of CSJ corpus.

We used a 2-pass decoder as the baseline. In the first pass, the CIHMMs and bigram language models were used. In the second pass, rescoring using trigram (and CDHMMs in the case of context-dependent modeling) was applied on 200-best hypotheses generated by the first pass. Beam width at the word end was set to 30 according to preliminary experiments. The number of the linear lexicon was set to \( N_{\text{lin}} = 250 \) for the JNAS test set and \( N_{\text{lin}} = 500 \) for the CSJ test set.

4.2 Recognition results of inter-word context-dependent modeling

Speech recognition results of JNAS based on context-dependent HMMs are shown in Table 1. We compared our proposed exact CDHMM search algorithm\(^2\) to CIHMM search algorithm, intra-word CDHMM search algorithm and approximate CDHMM search algorithm. Since the intra-word CIHMM search algorithm did not model the context-dependent pronunciation variations, only a slight improvement was achieved than the CIHMM search algorithm. The result of approximate CDHMM search algorithm was better than that of intra-word CDHMM search algorithm because the likelihood of the head syllable was calculated using CDHMM instead of CIHMM. Even the ‘likelihood difference index’ function has not been implemented, the proposed exact CDHMM search algorithm achieved significant improvement than other methods because it accurately modeled the inter-word context-dependent pronunciation variations. The proposed method achieved a relative error reduction of 24.1% on Cor. and 23.2% on Acc. over the CIHMM search algorithm.

We also evaluated the proposed method using CSJ corpus. The results of one-pass bigram are shown in Table 2. The tendency of CSJ was similar to that of JNAS. That is, the proposed method is robust on both read speech (JNAS) and spontaneous lecture speech (CSJ). The proposed method achieved a relative error reduction of 16.8% on Cor. and 11.6% on Acc. over the CIHMM search algorithm.

4.3 The effect of acoustic and language model rescoring

In this section, we investigated the effect of rescoring of the acoustic model and the language model in the 2nd pass. A traditional trigram language models were used in the 2nd pass for language model rescoring. Acoustic models rescores the N-best hypotheses using either the context-independent syllable-based HMMs or context-dependent syllable-based HMMs was also performed in the 2nd pass. The Viterbi algorithm was used to calculate the optimal acoustic likelihoods of the N-best hypotheses generated by the first pass, and the final result was determined by the maximum summation score of the 2nd pass acoustic likelihood and language probability. Furthermore, the same acoustic models were used in the first pass and the second pass to evaluate the search precision of our LVCSR.

The effect of acoustic and language model rescoring for

\(^2\)In this paper, the function of ‘likelihood difference index’ (Fig. 4 (b)) was not implemented.
Table 3 The effect of acoustic and language model rescoring for JNAS (%)

<table>
<thead>
<tr>
<th>AM in the 1st pass (bigram)</th>
<th>1st pass (CIHMM &amp; bigram)</th>
<th>2nd pass (CDHMM &amp; bigram)</th>
<th>2nd pass (CDHMM &amp; bigram)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIHMM</td>
<td>88.7</td>
<td>86.8</td>
<td>89.0</td>
</tr>
<tr>
<td>Approx. CDHMM</td>
<td>90.6</td>
<td>89.0</td>
<td>-</td>
</tr>
<tr>
<td>Exact CDHMM</td>
<td>91.0</td>
<td>88.6</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 4 The effect of acoustic model rescoring for CSJ (%). Bigram was used as language model.

<table>
<thead>
<tr>
<th>AM in the 1st pass</th>
<th>test data</th>
<th>1st pass (bigram)</th>
<th>2nd pass (CDHMM &amp; bigram)</th>
<th>2nd pass (CDHMM &amp; bigram)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Cor.</td>
<td>Acc.</td>
<td>Cor.</td>
</tr>
<tr>
<td>CIHMM</td>
<td>a01m00007</td>
<td>67.4</td>
<td>61.8</td>
<td>70.8</td>
</tr>
<tr>
<td></td>
<td>a01m0035</td>
<td>50.5</td>
<td>43.1</td>
<td>53.8</td>
</tr>
<tr>
<td></td>
<td>a01m0074</td>
<td>67.5</td>
<td>61.9</td>
<td>72.0</td>
</tr>
<tr>
<td></td>
<td>a05m0031</td>
<td>55.3</td>
<td>51.3</td>
<td>59.9</td>
</tr>
<tr>
<td></td>
<td>Ave.</td>
<td>60.2</td>
<td>54.5</td>
<td>64.1</td>
</tr>
<tr>
<td>Approx. CDHMM</td>
<td>a01m0007</td>
<td>71.4</td>
<td>65.6</td>
<td>72.5</td>
</tr>
<tr>
<td></td>
<td>a01m0035</td>
<td>56.0</td>
<td>48.8</td>
<td>57.7</td>
</tr>
<tr>
<td></td>
<td>a01m0074</td>
<td>72.6</td>
<td>67.5</td>
<td>74.0</td>
</tr>
<tr>
<td></td>
<td>a05m0031</td>
<td>62.9</td>
<td>59.1</td>
<td>64.1</td>
</tr>
<tr>
<td></td>
<td>Ave.</td>
<td>65.7</td>
<td>60.3</td>
<td>67.1</td>
</tr>
</tbody>
</table>

JNAS is shown in Table 3. Language model rescoring was very effective in all these cases. For acoustic model rescoring, the performance was improved significantly if more detailed acoustic models (CDHMMs) were used in the 2nd pass and CIHMMs were used in the 1st pass. When CDHMMs were used in both 1st pass and 2nd pass, the improvement became smaller. In the case of CIHMMs used in both two passes, only 0.3% absolute improvement was achieved in the 2nd pass. That is to say, the search precision of our LVCSR system in the first pass is relatively good.

The effect of acoustic model rescoring for CSJ is shown in Table 4. The similar tendency was obtained. The result of 1-pass approximate CDHMM search algorithm was better than that of the 2-pass CDHMM method. When CIHMMs were used to recalculate the acoustic likelihood, only 0.4% absolute improvement was achieved in the 2nd pass.

5 Conclusion

In this paper, we proposed an effective and accurate inter-word context-dependent modeling for LVCSR. A combinational use of linear lexicon and tree-structured lexicon in the 1-best approximation search algorithm for LVCSR based on CIHMMs proposed in our laboratory was extended to context-dependent modeling. We made branches for the head syllable according to the contexts and the paths are merged at the second syllable for the linear lexicon. For the tree-structured lexicon, branches were made in a similar way. At the end node of a word, the language scores have to be compensated considering the inter-word context, but the scores of contexts other than that of the best history are lost because of the merge at the second syllable. To solve this problem, we introduced the 'likelihood difference index'. We also investigated the effect of the 2-pass decoder with the comparatively complicated language models and acoustic models. We evaluated the proposed algorithms on JNAS and CSJ corpora. The proposed algorithms obtained a remarkable improvement of recognition performance, and the rescoring of the context-dependent syllable acoustic models in the 2nd pass mode also achieved a further improvement even the same acoustic models were used in the 1st pass.

References


*All likelihoods of 100 sentences were improved significantly