A Hybrid Approach to Robust Word Lattice Generation
Via Acoustic-Based Word Detection

Icksang Han, Chiyoun Park, Jeongmi Cho, Jeongsu Kim

Samsung Advanced Institute of Technology, Samsung Electronics Co., Ltd.,
San #14-1, Nongseo-Dong, Kiheung-Gu, Yongin-Si, Kyungki-Do, 446-712, Korea
{hanis, chiyoun.park, jmcho007, jeongsu.kim}@samsung.com

Abstract

A large-vocabulary continuous speech recognition (LVCSR) system usually utilizes a language model in order to reduce the complexity of the algorithm. However, the constraint also produces side-effects including low accuracy of the out-of-grammar sentences and the error propagation of misrecognized words. In order to compensate for the side-effects of the language model, this paper proposes a novel lattice generation method that adopts the idea from the keyword detection method. By combining the word candidates detected mainly from the acoustic aspect of the signal to the word lattice from the ordinary speech recognizer, a hybrid lattice is constructed. The hybrid lattice shows 33% improvement in terms of the lattice accuracy under the condition where the lattice density is the same. In addition, it is observed that the proposed model shows less sensitivity to the out-of-grammar sentences and to the error propagation due to misrecognized words.

Index Terms: LVCSR, word lattice, out-of-grammar, word detection, robust speech recognition

1. Introduction

Thanks to decades of research in the field of speech recognition, the speech recognizers have achieved high recognition rate in the lab condition. The speech recognizers in real-life applications, however, do not exhibit such high performances due to various reasons. Some of the major factors that degrade the performance of a speech recognition system includes: mismatch in the training and testing condition in terms of the acoustic modeling, and out-of-vocabulary (OOV) words or out-of-grammar (OOG) word sequences in terms of the language modeling. Among those, the linguistic factors are getting more importance as the system gets more complicated, such as large vocabulary continuous speech recognition (LVCSR), dialogue system, voice search, and spoken document retrieval (SDR).

The problem of OOV words is an important issue for voice search applications, where proper nouns and newly coined words are often used as inputs. In order to resolve the problem caused by the OOVs, subword-based[1] or word fragment-based methods[2] have been commonly used. In the case of SDR, there also have been approaches that construct the lattices for search indexing based on the word fragments in conjunction with a word-based model [3].

On the other hand, the OOG problem can be more serious than that of OOV in a continuous speech recognition (CSR) based dialogue system, especially when the sentences consist mainly of commonly used words. Most of the LVCSR systems heavily rely on the language models in order to reduce the memory space and computational complexity. However, OOG sen-

tences cannot be easily predicted by the language model, and this leads to an increase in the recognition error rate. Moreover, when a word is misrecognized, the error is likely to propagate to neighboring words due to the high dependency on the language model.

There have been attempt to avoid the problem of error propagation due to language modeling, one of which is the attention shift decoding method[4, 5]. This method first looks for a more reliable region within the lattice where it is more acoustically salient and then perform rescoring of the lattice based on the islands of reliability. But locating the reliable region itself remains a difficult problem.

Therefore, in this paper, we propose a method that can solve the error propagation and OOG problems without having to identify the regions of reliability. The proposed method borrows the idea from the keyword detection methods, so that the words that would have been ignored in the ordinary CSR system can be detected and be used for recognition.

This paper is organized as follows. In the next section, the hybrid word lattice generation method will be explained. Experiments on the proposed algorithm will be presented in three different settings and their results will be discussed in Section 3. Finally, the conclusion will follow in Section 4.

2. Hybrid lattice generation

Figure 1 shows a block diagram of the proposed lattice generation algorithm.

Figure 1: Schematic diagram of the proposed lattice generation algorithm
2.1. Word candidate detection

In this section, the new procedure will be explained in detail. The word candidate detection module can be divided into two stages. In the first step, individual word candidates are detected based on the acoustic score alone. In the second step, the likelihoods of the selected candidates are refined by connecting frequently occurring word sequences.

2.1.1. Individual word candidates

The individual word candidate detection is performed in three steps: phone recognition, word matching, and rescoring. The phonetic recognition step is performed first, since it is not plausible to try to match all possible words to all subsequence of the input signal. After the phone recognition module outputs the best matching phonetic sequence based on the given utterance, the word matching module matches phonetic representation of each vocabulary word to the recognized phone sequence as shown in Figure 2. A phone confusion matrix is used for measuring the similarity between the phone sequences. Only the words whose matching scores are above a certain threshold are rescoring based on the actual utterance signal.

By rescoring the more likely word candidates, the following two goals are achieved. First, a more complex context-dependent phone model can be applied in the rescoring process so that a better estimate of the acoustic model-based score can be calculated. Second, because the endpoints of the words that have been selected from the word recognition process have been limited to the endpoints of the recognized phones, the rescoring module need to relocate the precise endpoints of the words that lead to a higher acoustic score. The iterative Viterbi decoding method[6] is applied in order to refine the location of the words' endpoints.

2.1.2. Candidate sequences

The individual word candidate matching process generates an excessive number of short words as a result. This is because if a word consists of only a few phones, such as ‘o’ in Korean, it becomes more likely that the word is matched to a subsequence of a different words, such as ‘roma[Rome]’. Moreover, the matching acoustic score of ‘o’ are often higher than that of ‘roma’, which reduces the possibility of longer words being selected at the same time increasing the introduction of meaningless candidates.

Therefore, the word candidates are scored based not only on the acoustic model scores, but also on the frame length of the words. The length-adjusted score can be calculated as the following:

\[
\text{Score}_{\text{AM}} = \text{Score}_{\text{AM}} - \frac{\omega_N}{N}
\]

where \(\text{Score}_{\text{AM}}\) is the average log-likelihood per frame based on the acoustic model and \(N\) is the number of frames in the given word.

By applying this scoring method, the incorrect short words can be removed from the candidates, but it also eliminates the correct candidates with short length. This causes a problem especially when the sentence consists mainly of short words. In order to avoid such problem, the words that can be connected by a certain language model are connected together to form a word sequence. By considering concatenated words, the word sequence ‘bi ga wa[It is raining]’ can be recognized as a candidate even if each word ‘bi[rain],’ ‘ga[is]’ and ‘wa[coming]’ is ignored due to its shortness.

The word sequences are scored according to the following equation.

\[
\text{Score}_{\text{seq}} = \text{Score}_{\text{AM}} + \omega_{\text{LM}} \cdot \text{Score}_{\text{LM}} - \frac{\omega_N}{N}
\]

where \(\text{Score}_{\text{LM}}\) is the n-gram language model score for the word sequence.

The ratio of words that are correctly detected within the top \(N\) candidates based on different scoring methods is compared in Figure 3. When the score is adjusted according to the word length, a higher portion of the true words are detected within a few high scoring candidates, but when considering a larger number of candidates, the adjusted score performs poorly due to the elimination of the short words. When the word sequence is also taken into account, true words can still be detected within the high scoring candidates, at the same time maintaining the high performance when more candidates are considered.
2.2. Word lattice and hybrid lattice generation

The word lattice generation procedure corresponds to the blue dotted path in Figure 1. The CSR lattice generation module is equivalent to the first pass of the two-stage decoder described in [7]. The word lattice covers the whole utterance, but the generation of such a word lattice for LVCSR usually involves pruning based on a language model in order to reduce the complexity in computation time and memory space. Therefore, when a word cannot be predicted by the language model or when there has been misrecognition in the previous word, the word may not be included in the generated lattice.

On the other hand, the word candidate detection procedure in Section 2.1 does not perform such pruning, and so the detected candidates include all the words that are more clearly pronounced and have high acoustic scores. However, it is not guaranteed that the word candidates cover the whole utterance. Thus, it may not be possible to reconstruct the sentence based on these word candidates alone.

Therefore, by combining the recognized word candidates into the ordinary CSR-generated word lattice, the weakness of both methods can be complemented, and so one can be ready to start recognizing the whole sentence based on the clearly pronounced word candidates. Figure 4 illustrates this effect. When the word A is misrecognized as A’, ordinary CSR lattice generation process may propagate the error into the next word B. Similarly, if the word sequence C-D has not been modeled in the language model, the word D may not be correctly recognized even if C is detected. Combining the acoustic-based word candidate detection into the word lattice can recover this problem.

3. Experiments

3.1. Settings

The LVCSR system used in the experiment is based on the two-stage decoder in [7]. The 39th-order MFCC is applied for the feature set, and the acoustic model consists of two thousand tied-states, each of which is represented by a mixture of sixteen Gaussians. The acoustic model was trained on approximately 146 hours of data from a Korean sentence database sampled at 16kHz.

A trigram language model is applied for both the CSR lattice generation and the word candidate detection stages. The training data consists of 3M words extracted from the conversational text data within Sejong DB[8]. The vocabulary size is about thirty thousands.

The test database consists of two sets, each of which contains fifty distinct sentences. BASIC style sentences are guaranteed to be predicted by the n-gram language model, and FREE style sentences are the paraphrased version of the basic sentences. The perplexity of the basic style sentences is 30 and that of the free style sentences is 174. Both sets contain OOVs less than 1%. Those sentences have been recorded by ten speakers (five males and five females), and two different microphones were used to record the utterances simultaneously: one with a headset microphone (Sennheiser ME3), one with a microphone array (Acoustic Magic Voice Tracker) at 2.5m distance apart.

Three thousand candidates have been detected in the candidate detection procedure and were combined into the hybrid word lattice, and the weights of \( \omega_{LM} = 0.05 \) and \( \omega_{N} = 100 \) have been used for the scoring.

3.2. Lattice accuracy and lattice density

The performance of the hybrid word lattice is measured by the lattice accuracy. This measure is calculated by finding the maximum number of correct words within a sequence of word candidates from the lattice, and then dividing it by the number of words in the sentence. It is restricted that the word sequence should have zero margins between words, so that the lattice accuracy becomes the theoretical least upper bound for the recognition accuracy. Both word and sentence accuracies are calculated.

Another measure is the lattice density. This measure is defined as the average number of words that contains a certain frame. It means the average number of word candidate that the speech recognizer should be keeping track of during the decoding process at a certain time point. As the lattice density increases, the computational complexity in the second phase of the speech recognizer is likely to increase.

3.3. Results and discussions

3.3.1. Lattice accuracy

Table 1 compares the lattice accuracy in different evaluation sets. The baseline corresponds to the ordinary CSR-based lattice, and the hybrid stands for the proposed lattice generation method. It can be observed from the table that the hybrid method increases both word and sentence accuracies significantly. This improvement means that the words that were pruned due to OOG problems could be recognized in the word candidate generation stage.

As for the style of sentences, the free style utterances show a much higher improvement than basic utterances. This phenomenon implies that the application of the acoustic-model based word candidate detection can compensate for the errors caused by OOG problems to a certain amount. This difference in the improvement becomes less obvious in the distant microphone case, where the accuracy of the acoustic model starts to take a bigger role.

<table>
<thead>
<tr>
<th>STYLE</th>
<th>DISTANCE</th>
<th>BASELINE</th>
<th>HYBRID</th>
<th>RATIO</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASIC</td>
<td>NEAR</td>
<td>99.3 / 95.8</td>
<td>99.4 / 96.4</td>
<td>32.4</td>
</tr>
<tr>
<td></td>
<td>FAR</td>
<td>95.8 / 81.8</td>
<td>97.7 / 87.8</td>
<td>44.3</td>
</tr>
<tr>
<td>FREE</td>
<td>NEAR</td>
<td>95.67 / 76.4</td>
<td>98.1 / 88.2</td>
<td>56.7</td>
</tr>
<tr>
<td></td>
<td>FAR</td>
<td>90.9 / 57.6</td>
<td>95.2 / 74.6</td>
<td>46.7</td>
</tr>
</tbody>
</table>
3.3.2. Lattice density

It may be argued that the more words are added to the lattice, the higher the accuracy will increase. Therefore, the increase in the lattice size is also compared in order to verify that the lattice accuracy still increases under the same lattice density.

Different pruning threshold was used for changing the lattice density. Histogram pruning method was applied to the original CSR lattice generation procedure, and the rate of pruning was determined by changing the histogram ratio. The histogram ratio of 3.5% means that the lattice is pruned in a way that only 3.5% of the whole nodes in the search net are activated at a certain point in time. Therefore, as the histogram ratio goes lower, more words are pruned; resulting in a faster recognition speed while sacrificing the recognition accuracy.

Table 2 compares the lattice accuracy and the lattice density of the free style utterances recorded by the distant microphone array. Comparing the result of baseline in 3.5% threshold and the hybrid result in 0.5% threshold, it can be observed that the lattice density is lower in the hybrid case, but the lattice accuracy has increased by about 33%. Therefore, it can be concluded that combining acoustic-based word candidates into the lattice is more advantageous than applying less pruning to the original CSR word lattice generation.

3.3.3. Corrupted Signal

The experiments above have shown that the word candidate generation procedure resolves the problem arisen from OOG sentences and that this effect can be achieved without increasing the size of the word lattice. The experiment in this section will show the evidence that the use of acoustic-model based candidate generation is effective in preventing the propagation of error as well.

The test set utterances were intentionally corrupted so as to artificially simulate the case when one or more specific words are misrecognized. The length of the corrupted region was limited to 10–30% of the whole utterance, and the corrupted region was positioned in the first half of the signal so that the propagation of error can be easily observed. The corruption was induced by superimposing a white noise of −20dB as shown in Figure 5.

Table 3 compares the rate of lattice accuracy between the clean signal and noise-corrupted signal in the baseline CSR lattice and the hybrid model. The corruption rate is calculated to be the ratio of words that are not detected in the corrupted signal but are detected in the original signal. It can be observed that the corruption rate is higher in the baseline than in the hybrid case, which implies that the misrecognition at a certain point in time have propagated to other portions of the signal in the baseline system, but it does not propagate as much if we combine the acoustic-based word candidates.

4. Conclusions

This paper proposed a method for lattice generation which can avoid the side-effect of applying the language model in a LVCSR system. By combining the word candidates detected mainly from the acoustic aspect of the signal to the word lattice from the ordinary CSR’s word lattice, a hybrid lattice can be constructed.

The hybrid lattice showed 33% improvement in terms of lattice accuracy, in the condition where the lattice density is similar. In addition it has been shown that the proposed model is less sensitive to the OOG sentences and to the error propagation due to misrecognized words.

In the future, more work is required in building a CSR system that takes advantage of this hybrid word lattice.

5. References

[8] Online: http://www.sejong.or.kr