Advanced Science and Technology Letters
Vol.58 (Electrical Engineering 2014), pp.104-108
http://dx.doi.org/10.14257/astl.2014.58.22

Confusability Measure Based Lexicon Optimization for Fast LVCSR Decoding

Nam Kyun Kim, Woo Kyung Seong, and Hong Kook Kim

School of Information and Communications
Gwangju Institute of Science and Technology (GIST)
{skarbs001, wkseong, hongkook}@gist.ac.kr

Abstract. In this paper, we propose a lexicon optimization method based on confusability measure (CM) in order to reduce the decoding time for a large vocabulary continuous speech recognition (LVCSR) system. When lexicon is built or expanded for unseen words by using grapheme-to-phoneme (G2P) conversion, the lexicon size increases since G2P is generally realized by 1-to-N-best mapping. Thus, the proposed method prunes the confusable words in the lexicon by a CM that is defined a linguistic distance between two phonemic sequences. It is demonstrated from LVCSR experiments that the proposed lexicon optimization method achieves a relative real-time factor reduction of 23.13% on a task on the Wall Street Journal, compared to the 1-to-4-best G2P converted lexicon approach.

Keywords: Large vocabulary continuous speech recognition, lexicon optimization, confusability measure, grapheme-to-phoneme conversion, weighted finite-state transducer

1 Introduction

Recently, there have been many research works proposed to develop large vocabulary continuous speech recognition (LVCSR) systems, such as feature extraction, acoustic modeling, language modeling, decoding, and so on. Especially, fast decoding or search from acoustic feature vectors to word sequences plays a main role in deploying LVCSR in practice. In general, the size of a decoding network for LVCSR becomes extremely large, which results in increasing confusability and computational complexity. Thus, weighted finite-states transducer (WFST)-based approaches have been proposed for reducing decoding time [1].

There are two ways of decreasing the size of the decoding network. One is to remove unnecessary nodes of decision-trees in word-conditioned tree search [2], and the other is to decrease the number of arcs of a WFST by constructing a parallel word-end silence and short-pause structure in the WFST framework [3]. Those methods were reported to reduce decoding time, but they failed to improve speech recognition performance [2][3]. As an alternative, the size of a decoding network was reduced by using a confusability measure (CM) [4]. While this method could improve speech recognition performance, it suffered from an excessive removal of words,

ISSN: 2287-1233 ASTL
Copyright © 2014 SERSC

Table 1. Example of CM scores for the phoneme sequences of the word ‘SOONEST’ obtained by 1-to-4-best mapping.

4-best phoneme sequence	CM score
S UW N AH S T	**0.0292969**
S UW N IH S T	**0.0292969**
S UW N AH S	0.0195312
S UW N S T	0.0195312

making out-of-vocabulary problem [5].

In this paper, we propose a new CM-based decoding network reduction method by incorporating grapheme-to-phoneme (G2P) conversion. In the proposed method, a lexicon is expanded by using multiple phoneme sequences of each word in the lexicon, which is realized by 1-to-N-best alignment from G2P conversion. Then, the lexicon is optimized by pruning confusable phoneme sequences using a CM that is defined by Levenshtein distance between phoneme sequences. A detailed explanation of the proposed method is given in the following sections.

2 Lexicon Construction Using a G2P Model

G2P conversion tries to predict phoneme sequences by aligning graphemes of a word or a sentence with phonemes [6]. Among many approaches to realizing such alignment, the simplest G2P conversion is done by a dictionary look-up [6]. That is, for a given input grapheme sequence, a possible phoneme sequence is obtained by a look-up table. Therefore, the dictionary look-up approach is a time-consuming and tedious work. Moreover, it is hard to find the pronunciation of unseen words because the dictionary used for the look-up is finite.

To overcome the limitation of such a finite dictionary, a data-driven approach is used for G2P conversion [6]. This is usually performed by mapping 1 to N-best after designing a joint-sequence model from a training corpus [6].

3 Proposed CM-Based Lexicon Optimization Method

This section describes how to optimize an N-best lexicon that is extended from the G2P conversion. This is because we need to reduce the decoding time of LVCSR with the N-best lexicon. This is achieved by removing phoneme sequences in the lexicon using a CM.

Assuming that the *i*-th word, W_i, in the original lexicon has N-best phoneme sequences, $s_{i,j}\,(j=1,\cdots,N)$, by the 1-to-N-best mapping of G2P. Then, the CM of $s_{i,j}$ is defined as $CM(s_{i,j}) = L(s_{i,j}) \cdot (\min_{1\le k\le N_w,\, k\ne i, 1\le l\le N}\ (D(s_{i,j}, s_{k,l}) \cdot L(s_{k,l})))$, where N_w is the number of words in the original lexicon and $D(x, y)$ is the Levenshtein distance between *x* and *y* [4]. In addition, $L(s_{k,j})$ is defined as the length normalized by $l_{\max}$ that is the maximum length among all the phoneme sequences in the G2P converted lexicon. That is, $L(s_{k,j}) = \#(s_{k,j}) / l_{\max}$, where $\#(s_{i,j})$ is the number of phonemes in $s_{i,j}$ and $l_{\max} = \max_{1 \le i \le N_w, 1 \le j \le N} \#(s_{i,j})$.

Table 2. Performance evaluation of an LVCSR system.

Method			WER (%)	Memory (MB)	Decoding network		RTF
					No. of States	No. of Arcs	
Baseline			12.35	1,279	26,030k	642,67k	0.187
1-to-4-best G2P converted lexicon			13.93	2,471	47,910k	125,992k	0.268
The proposed method	Thres hold	0.02	13.06	1,766	35,177k	89,377k	0.209
		0.03	12.74	1,749	34,844k	88,523k	0.207
		0.04	12.23	1,737	34,651k	87,897k	0.206
		0.05	12.67	1,723	34,302k	87,256k	0.206

The proposed CM-based lexicon optimization method finds the phoneme sequences whose CM scores are above a pre-defined threshold. In this case, in order to have the optimized lexicon include at least one phoneme sequence for each word in the original lexicon, the phoneme sequence that has the highest CM score is first maintained in the optimized lexicon. Next, the phoneme sequences having CM score lower than the threshold are assumed to be confusable words and will not appear in the pruned lexicon.

Table 1 shows an example of the phoneme sequences obtained by the 1-to-4-best G2P conversion for the word 'SOONEST' and their CM scores. In this case, the most probable phoneme sequence is /S UW N AH S T/. If the threshold is 0.02, two phoneme sequences, /S UW N AH S T/ and /SUW N IH S T/, will remain in the lexicon.

For LVCSR, a WFST-based decoder is composed of four different WFSTs: an N-gram language model, G, a lexicon, L, a context dependency expansion, C, and a hidden Markov model (HMM) state level topology, H, which results in $H \circ C \circ L \circ G$ [7]. Therefore, the proposed lexicon optimization method transforms the lexicon, L, into the optimized lexicon, L'. Thus, we obtain the WFST-based decoder that is composed as $H \circ C \circ L' \circ G$.

4 Speech Recognition Experiment

In this section, we evaluated the performance of a WFST-based ASR system employing the proposed lexicon optimization method. To this end, a baseline ASR system was constructed by using the Kaldi speech recognition toolkit [8], where 7,138 utterances of the Wall Street Journal (WSJ0) [9] were used as the training database. Note that the CMU dictionary [10] was used by dictionary look-up for generating a baseline lexicon. The acoustic models were composed of triphones with a 3-state left-to-right HMM. Each state of the triphone models was clustered by a decision tree, resulting in 1,882 tied states that consisted of the total number of 15,046 Gaussian mixture models (GMMs). As a speech recognition feature, 12 mel-frequency cepstral coefficients (MFCCs) with a logarithmic energy were extracted, and their first and second

Copyright © 2014 SERSC

derivatives were concatenated to obtain a 39-demensional feature vector. In addition, cepstral mean normalization (CMN) was applied to the feature vectors. A tri-gram language model was constructed with 20k words. As a test database, 333 utterances (Nov ’92) of WSJ0, which was composed of 5,643 words, were used.

Table 2 compares the performance of a baseline ASR system with a lexicon obtained by the dictionary look-up, an ASR system with an extended lexicon by using 1-to-4-best G2P conversion, and that with an optimized lexicon using the proposed lexicon optimization method. The comparison was done in terms of word error rate (WER), memory usage for lexicon, and real-time factor (RTF). Note that we evaluated the performance of the proposed method by changing the threshold from 0.02 to 0.05 in 0.01 increments. As shown in the table, as the threshold increased, the decoding network size decreased and average WER was lowered. However, as the threshold went higher, the average WER of the proposed method also went higher. This was because phoneme sequences were pruned excessively. Consequently, by applying the proposed method, we could achieve relative WER reduction of 12.23% and relative RTF reduction of 26.13% when the threshold was set to 0.04.

5 Conclusion

In this paper, a CM-based lexicon optimization method was proposed to reduce the size of a lexicon that was generated by using G2P conversion. It was shown from ASR experiments that an ASR system employing a lexicon optimized by the proposed method provided relative WER reduction and RTF reduction of 12.23% and 23.13%, respectively, compared to that with a lexicon by 1-to-4-best G2P conversion.

Acknowledgments. This work was supported in part by the ICT R&D program of MSIP/IITP [10035252, Development of dialog-based spontaneous speech interface technology on mobile platform] and the National Research Foundation of Korea (NRF) grant funded by the government of Korea (MSIP) (No. 2012-010636).

References

1. Mohri, M., Pereira, F., Riley, M.: Speech recognition with weighted finite-state transducers. In: Handbook on Speech Processing and Speech Communication, Springer, (2008).
2. Neukirchen, C., Willett D., Rigoll, G.: Reduced lexicon trees for decoding in a MMI-Connectionist/HMM speech recognition system. In: Proceedings of Eurospeech, Rhodes, Greece, pp. 2639-2642 (1997).
3. Guo, Y., Li, T., Si, Y., Pan, J., Yan, Y.: Optimized large vocabulary WFST speech recognition system. In: Proceedings of FSKD, Chongqing, China, pp. 1243-1247 (2012).
4. Kim, M. A., Oh, Y. R., Kim, H. K.: Optimizing multiple pronunciation dictionary based on a confusability measure for non-native speech recognition. In: Proceedings of IASTED, Innsbruck, Austria, pp. 215-220 (2008).
5. Jitsuhiro, T., Takahashi, S., Aikawa, K.: Rejection of out-of-vocabulary words using phoneme confidence likelihood. In: Proceedings of ICASSP, Seattle, WA, pp. 217-220 (1998).

6. Bisani, M., Ney, H.: Joint-sequence models for grapheme-to-phoneme conversion. Speech Communication, 50(5), pp. 434-451 (2008).
7. Moore, D., Dines, J., Magimai-Doss, M., Vepa, J., Cheng, O., Hain, T.: Juicer: a weighted finite-state transducer speech decoder. MLMI, Bethesda, MD, pp. 285-296 (2006).
8. Povey, D., et al.: The Kaldi speech recognition toolkit. In: Proceedings of IEEE ASRU, Honolulu, HI, pp. 1-4 (2011).
9. Paul, D. B., Baker, J. M.: The design for the Wall Street Journal-based CSR corpus. In: Proceedings of ICSLP, Stroudsburg, PA, pp. 357-362 (1992).
10. Weide, H.: The CMU Pronunciation Dictionary, Release 0.6. Carnegie Mellon University (1998).

Copyright © 2014 SERSC