Lexicon Optimization for WFST-Based Speech Recognition Using Acoustic Distance Based Confusability Measure and G2P Conversion

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

Abstract In this paper, we propose a lexicon optimization method based on a confusability measure (CM) to develop a large vocabulary continuous speech recognition (LVCSR) system with unseen words. When a 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 attempts to prune the confusable words in the lexicon by a CM defined as an acoustic model distance between two phonemic sequences. It is demonstrated by LVCSR experiments that the proposed lexicon optimization method achieves a relative word error rate (WER) reduction of 14.72% in a *Wall Street Journal task* compared to the 1-to-4-best G2P converted lexicon approach.

1 Introduction

Recently, many research works have been proposed to develop large vocabulary continuous speech recognition (LVCSR) systems, such as feature extraction, acoustic modeling, pronunciation modeling, language modeling, decoding, and so on [1]. Among them, decoding or search with acoustic feature vectors for word sequences plays a main role in the performance of LVCSR systems in which decision tree-based approaches or weighted finite-state transducer (WFST) approaches have been typically used for LVCSR decoding [2]. The decision tree-based approach requires a small amount of decoding memory. However, since on-the-fly composition must be performed with language models (LMs) during the recognition of test utterances, this approach makes decoding speed slow [2]. Conversely, a WFST for LVCSR decoding can generally be constructed by the composition of different speech recognition knowledge sources, such as a hidden Markov model (HMM) topology, a

Nam Kyun Kim,, Woo Kyeong Seong, Hong Kook Kim
School of Information and Communications, Gwangju Institute of Science and Technology (GIST), e-mail: {skarbs001, wkseong, hongkook}@gist.ac.kr

context-dependent phone model, a lexicon, and an n-gram LM, where each source is also represented by an individual WFST [3]. Thus, due to such modular representation and optimization techniques, a WFST-based decoder offers simpler realization and faster decoding than a decision-tree based decoder [3].

When the domain for an LVCSR system is dynamically changed due to new-coined words, a word lexicon must be reconstructed to accommodate unseen words by using a data-driven approach. A method to deal with this problem is grapheme-to-phoneme (G2P) conversion of such unseen words, which can be used for an expanded lexicon [4]. However, the accuracy of G2P conversion depends on how much knowledge is incorporated into the design of the G2P conversion, which is liable to be erroneous [4]. Thus, it is better to make multiple pronunciations for a given unseen word by using an N-best G2P conversion, which unfortunately results in the excessive increase of the lexicon size and a further increased size of the LVCSR decoder. Consequently, the word error rate (WER) of the LVCSR system is increased.

For a decision-tree based decoder, eliminating the unnecessary nodes of a decision tree was proposed in [5]. This approach reduced the size of the decoder, but the WER of the reduced decision-tree based decoder was similar to that of the original decision-tree based one. For WFST-based decoders, several structural optimization techniques were proposed by sharing silence and short-pause states and restructuring the beam depending on the token path [6]. While this approach efficiently optimized the WFST, it was hard to apply to the unseen word problem. In addition, a minimum classification error (MCE) model [7] and a conditional random field (CRF) model [8] were proposed to optimize the decoding network size during WFST training. However, these methods need to be applied repeatedly for retraining the WFST if unseen words are given. As an alternative, the decoding network sizXe was reduced by using a confusability measure (CM) [9]. This approach reduced the size, but it suffered from the excessive removal of words, causing an out-of-vocabulary problem [10].

In this paper, we propose a method to optimize a G2P converted lexicon that is realized by the N-best phoneme sequences of each word. To this end, a CM is first defined by an acoustic distance between two phoneme sequences and the length of the phoneme sequences. Then, a G2P model-based N-best lexicon is constructed to find the most probable phoneme sequences of unseen words. However, since the lexicon becomes oversized, the lexicon is then optimized by pruning the confusable phoneme sequences using the CM.

Following this introduction, Section 2 briefly explains a lexicon construction using a G2P model. Section 3 describes the CM using the acoustic models and the dynamic programming-based alignment between two phoneme sequences. Next, a lexicon optimization method based on the CM is proposed. Section 4 evaluates the performance of an automatic speech recognition system (ASR) system employing the proposed method in terms of computational complexity and WERs. Finally, the findings are summarized in Section 5.

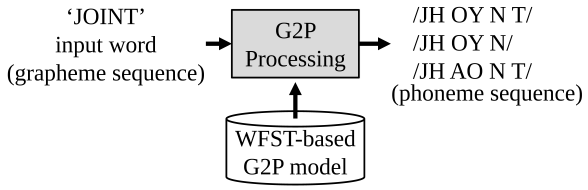

Fig. 1 Example of G2P conversion for given word JOINT.

2 G2P Model-Based Lexicon Generation

G2P conversion tries to predict phoneme sequences by aligning graphemes of words or sentences with phonemes [4]. Among many approaches for realizing such alignments, the simplest G2P conversion is achieved by a dictionary look-up [4]. 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 time- consuming and tedious. Moreover, it is hard to find the pronunciation of unseen words, because the dictionary used for the look-up is finite. In addition, it does not enable unseen words to be found that do not exist in the dictionary. To overcome the limitations of such finite dictionaries, a data-driven approach is used for the G2P conversion [4]. This is usually performed by mapping 1 to N-best after designing a joint-sequence model from a training corpus. Fig. 1 shows an example of the G2P conversion for the given word JOINT. As shown in the figure, this word can be represented by three different phoneme sequences.

3 Proposed Lexicon Optimization

A CM can be defined by the linguistic distance between two phoneme sequences in the expanded lexicon of a G2P model [4]. In this section, we propose a lexicon optimization method that is defined by an acoustic distance between two phoneme sequences using inter-phone and inter-word distances. The proposed method is explained in detail in the following subsections.

3.1 Confusability Measure Using Distance between Phoneme Sequences

First, let W_i be the i-th word in the original N-best lexicon from the G2P conversion that has phoneme sequences in which the number is N, the number of words is N_W, and $s_{i,j}(j = 1,...,N)$ are the 1-to-N-best mapped phoneme sequences. Then, the CM of $s_{i,j}$ is defined as [9]

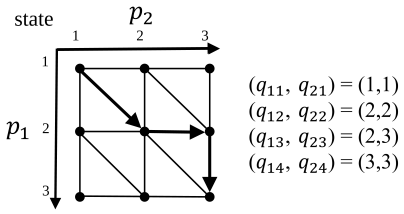

Fig. 2 Subset of alignments used to calculate inter-HMM distance, where bold lines correspond to alignment Q in Eq. (3). The values of q_{1i} and q_{2i} are the aligned states [11].

$$CM(s_{i,j}) = L(s_{i,j}) \cdot \min_{\substack{1 \le k \le N_w,\, k \neq i \\ 1 \le l \le N}} (D(s_{i,j}, s_{k,l}) \cdot L(s_{k,l})) \tag{1}$$

where $D(x,y)$ is the dynamic programming (DP)-based phoneme sequence distance that is defined by the HMM-based phoneme distance [11]. Moreover, $L(x)$ is defined as the normalized length by l_{max} That is,

$$L(x) = \frac{\#(x)}{l_{\max}} \tag{2}$$

where $\#(x)$ is the number of phonemes in x and $l_{\max} (= \max_{1 \le i \le N_W, 1 \le j \le N} \#(s_{i,j}))$ is the maximum length in the N-best G2P converted lexicon.

3.2 Phoneme Sequence Distance Measure

3.2.1 HMM-Based Phoneme Distance Measure

An acoustic distance between two phonemes can be calculated by using acoustic models [11], which is defined as

$$d_{HMM}(p_1, p_2) = \frac{\sum_Q P(Q) \frac{1}{L} \sum_{i=1}^{L} D_N(N_{q_{1i}}, N_{q_{2i}})}{\sum_Q P(Q)} \tag{3}$$

where Q is an alignment between the states of the HMMs of the phones p_1 and p_2, $P(Q)$ is the probability of Q, L is the length of the alignment, q_{1i} and q_{2i} are the states of the models that are aligned according to Q, $N_{q_{1i}}$ and $N_{q_{2i}}$ are the Gaussian distributions associated with the states $N_{q_{1i}}$ and $N_{q_{2i}}$, $D_N(\cdot)$ is the distance between the two Gaussian distributions. In Eq. (3), $P(Q)$ is calculated by multiplying the transition probabilities of both phoneme state sequences. Fig. 2 shows an example of possible $P(Q)$'s that are represented as $(q_{11} \to q_{12}, q_{21} \to q_{22})$, $(q_{12} \to q_{12}, q_{22} \to q_{23})$ and $(q_{12} \to q_{13}, q_{33} \to q_{23})$.

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

4-best Phoneme Sequence	CM score
S T AE CH UW	0.0497
S T AE CH Y UW	0.0499
S T AE CH UW EH	0.019
S T AE CH UW AH	0.0181

The acoustic model for calculating the distance between the two phonemes can be represented as one Gaussian distribution for each state of HMM models [8]. In this paper, we calculated $D_N(\cdot)$ using each of three different distance measures such as a Euclidean (EUC) distance, a Mahalanobis (MAH) distance, and a symmetric KullbackLeibler (KL) distance [12].

3.2.2 DP-Based Phoneme Sequence Distance Measure

A dynamic time warping (DTW) technique is incorporated into the acoustic distance to determine how different the two phoneme sequences are. The DTW is defined as [11]

$$D(x,y) = d_{DTW}(s_x, s_y) \tag{4}$$

where

$$d_{DTW}(s_x, s_y) = \min_{F} \left[\frac{\sum_{k=1}^{K} d_{HMM}(p_x(k), p_y(k))\, w(k)}{\sum_{k=1}^{K} w(k)} \right] \tag{5}$$

In Eq. (5), $d_{HMM}(p_x(k), p_y(k))$ is the distance between the HMMs described in Eq. (3). The weighting function, $w(k)$ applied to the DTW distance is used to normalize for the path F and it is defined as [11]

$$w(k) = i(k) - i(k-1) + j(k) - j(k-1) \tag{6}$$

where $i(1) = j(1) = 0$. In addition, $c(k)$ in the path $F = \{c(1), c(2), \cdots, c(K)\}$ consists of the pair of coordinates $(i(k), j(k))$ in the i and j directions when K is the number of alignments of the two phoneme sequences.

The measure obtained with DTW is the minimum weighted sum of the distance between the phoneme sequences for all the possible alignments between the sequences. Therefore, the DTW technique forces an alignment that minimizes the accumulated distance and forces the two sequences to consider the similarity.

Table 2 Performance evaluation of LVCSR system.

ASR System	Baseline	4-best G2P Converted Lexicon	Proposed Method		
			EUC	KL	MAH
WER (%)	12.19	13.93	12.58	12.42	11.88
RTF	0.239	0.476	0.364	0.381	0.380

3.3 Lexicon Optimization Using CM

In this subsection, we describe how to optimize the lexicon using CM. The proposed method selects the phoneme sequences with CM scores above a pre-defined threshold, except one phoneme sequence for each word in the original lexicon that has the highest CM score first maintained in the optimized lexicon. Next, the phoneme sequences having CM scores lower than the threshold are assumed to be confusable words and will not appear in the pruned lexicon.

Table 1 provides an example of the phoneme sequences obtained by the 1-to-4-best G2P conversion for the word STATUE and their CM scores. In this case, the most probable phoneme sequence is /S T AE CH Y UW/. If the threshold is 0.02, two phoneme sequences, /S T AE CH UW/ and /S T AE CH Y UW/, will remain in the lexicon.

3.4 Decoding Network Generation

A WFST-based decoder for LVCSR is fully composed as $H \circ C \circ L \circ G$ where four different WFSTs H, C, L and Grepresent the HMM state level topology, the context dependency expansion, the lexicon, and the n-gram LM, respectively [3]. 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

To evaluate the performance of the lexicon optimization method, we constructed the following ASR systems: a baseline ASR system (Baseline), an ASR system of a 1-to-4 best G2P converted lexicon and three ASR systems based on lexicons that were pruned by the proposed lexicon optimization method using different acoustic distances. The baseline system was constructed by the Kaldi speech recognition toolkit [13] with 7,138 utterances of the Wall Street Journal (WSJ0) [14]. In addition, for the baseline lexicon, a 1-best G2P lexicon was used. As a feature of the system, 39-

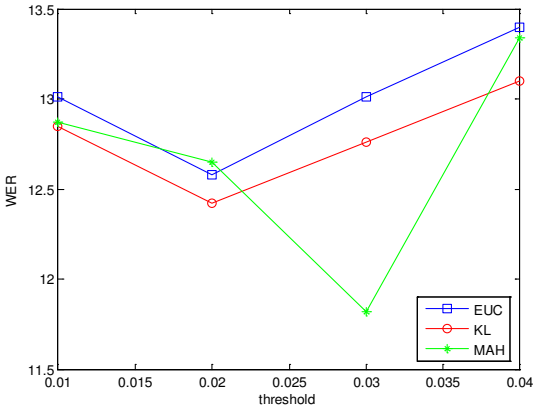

Fig. 3 Performance of the proposed method by changing a threshold

dimensional mel-frequency cepstral coefficients (MFCCs) were used, and cepstral mean normalization (CMN) was applied to the feature vector. The acoustic model was constructed by means of concatenating context-dependent HMMs, and a trigram LM was constructed from a set of sentences from WSJ0 with a vocabulary of 20k different words. The test sub-corpus was also extracted from WSJ0 and was composed by 333 utterances containing 5,643 different words.

Table 2 compares the WER and real time factor (RTF) for each ASR system using a lexicon obtained from the 1-best G2P converted lexicon, a 1-to-4 best G2P converted lexicon, and a pruned lexicon based on the proposed method with different phoneme distances, using EUC, KL, and MAH distances [12]. As shown in the table, with the different phoneme distances, the RTF and WER were lowered. In Fig. 3, we evaluated the performance of the proposed method by changing the threshold from 0.01 to 0.04 at a step of 0.01. As shown in the figure, average word error rate (WER) was lowered. However, as the threshold became greater than 0.02, average WER of the proposed method also went higher. This was because phoneme sequences were pruned excessively. Consequently, by applying the proposed method with MAH, we could achieve a relative WER reduction of 14.72% compared to that achieved with a lexicon of a 1-to-4-best G2P conversion.

5 Conclusion

In this paper, we proposed a lexicon optimization method based on CM to reduce the decoding network of lexicons constructed by the G2P model. When the lexicon was built to find the phoneme sequences of unseen words, the lexicon often became oversized, causing an increase in the size of the LVCSR decoder. As a result, the performance of the LVCSR was lowered. The proposed lexicon optimization method

was used for reducing the decoding network by pruning phoneme sequences that were much more confusable than other phoneme sequences. It was shown from ASR experiments that an ASR system employing a lexicon optimized by the proposed method provided a relative WER reduction of 14.72% compared to that of a lexicon from a 1-to-4-best G2P conversion.

Acknowledgements 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 platforms].

References

1. Saon, G., Chien, J.-T.: Large-vocabulary continuous speech recognition systems - a look at some recent advances. IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 18-33, (2012)
2. Kanthak, S., Ney, H. Rily, M., Mohri, M.: A comparison of two LVR search optimization techniques. In: Proceedings of Interspeech, Denver, CO, pp. 1309-1312 (2002)
3. Mohri, M., Pereira, F., Riley, M.: Speech recognition with weighted nite-state transducers. In: Handbook on Speech Processing and Speech Communication, Springer, (2008)
4. Bisani, M., Ney, H.: Joint-sequence models for grapheme-to-phoneme conversion. Speech Communication, vol. 50, no. 5, pp. 434-451 (2008)
5. 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)
6. 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)
7. Lin, S.-S., Yvon, F.: Optimization on decoding graphs by discriminative training. In: Proceedings of Interspeech, Antwerp, Belgium, pp. 1737-1740 (2007)
8. Kubo, Y., Watanabe, S., Nakamura A.: Decoding network optimization using minimum transition error training. In: Proceedings of ICASSP, Kyoto, Japan, pp. 4197-4200 (2012)
9. 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)
10. 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)
11. Anguita, J., Hernando, J., Peillon, S., Bramoulle, A.: Detection of confusable words in automatic speech recognition. In: IEEE Signal Process. Lett. 12 (8), 585588 (2005)
12. Sooful, J.J., Botha, E.C.: An acoustic distance measure for automatic cross-language phoneme mapping. In: Proceedings of PRASA, Franschhoek, South Africa, pp. 99-102 (2001)
13. Povey, D., et al.: The Kaldi speech recognition toolkit. In: Proceedings of IEEE ASRU, Honolulu, HI, pp. 1-4 (2011)
14. 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)