Two-Stage Lexicon Optimization of G2P-Converted Pronunciation Dictionary Based on Statistical Acoustic Confusability Measure

Nam Kyun Kim, Woo Kyeong Seong, Hun Kyu Ha, and Hong Kook Kim

School of Information and Communications, Gwangju Institute of Science and Technology, Gwangju 61005, Korea
E-mail: {skarbs001, wkseong, hnine, hongkook}@gist.ac.kr Tel: +82-62-715-3121

***Abstract*— In this paper, we propose a two-stage lexicon optimization method based on a statistical acoustic confusability measure to generate an optimized lexicon for automatic speech recognition (ASR). It is usual to build a lexicon by using grapheme-to-phoneme (G2P) conversion. However, G2P is often realized by 1-to-N best mapping, which results in the increase of lexicon size. To mitigate this problem, the proposed method attempts to prune the confusable words in the lexicon by using a confusability measure (CM) defined as an acoustic model (AM) based distance between two pronunciation variants. In particular, the first stage of the proposed method coarsely prunes the lexicon by a CM defined from monophone-based hidden Markov models (HMMs), and the second stage prunes it further by a CM defined from triphone-based HMMs. It is demonstrated from ASR experiments that an ASR system employing the proposed lexicon optimization method achieves a relative word error rate reduction of 18.88% on a task of Wall Street Journal, compared to that using a G2P-converted pronunciation dictionary without any optimization.**

I. INTRODUCTION

Recently, a number of automatic speech recognition (ASR) systems have been proposed to handle large vocabularies of words by means of acoustic modeling, language modeling, pronunciation modeling and hybrid modeling [1]. A pronunciation lexicon plays a main role in linking the acoustic model and language model [2]. The lexicon is typically handcrafted by experts, which is a time-consuming and tedious process. Thus, several methods have been developed to automatically generate the words in a lexicon. One of them is a weighted finite-state transducer (WFST)-based grapheme to phoneme (G2P) conversion technique [3]. This technique has a larger size of lexicon and can achieve better word accuracy than a hand-labeled lexicon. However, the confusability between words in such an expanded lexicon is increased because when the pronunciation variants of each word are generated, the pronunciation dictionary incurs more confusability due to an increase in overlapped pronunciation variants [2]. Thus, the proper pronunciation variants should be selected in lexicon modeling for the further improvement of ASR performance.

There has been a multitude of approaches regarding lexicon modeling. For example, a G2P conversion method based on hidden conditional random fields (HCRFs) was proposed for a large vocabulary continuous speech recognition (LVCSR) system [4]. In addition, a data-driven method using pronunciation mixture model (PMM) and letter to sound model (L2S) was proposed for generating a weighted lexicon [5]. However, both methods did not consider the overlapped pronunciation variants in the expanded lexicon.

In this paper, we focus on a lexicon optimization approach that uses a statistical acoustic CM and a lexicon expansion technique. While there was an approach to reduce lexicon size using a CM, this approach suffered from the excessive removal of words, causing an out-of-vocabulary problem [7]. To remedy such an excessive removal problem, we had proposed a lexicon optimization method using monophone-based acoustic distance between two pronunciation variants [8]. That is, we applied the technique to an ASR system based on triphone HMMs but the lexicon was optimized using monophone HMMs. Therefore, there might be a room for further optimization suitable for triphone-based HMMs. Thus, the proposed method in this paper tries to extend our previous one for triphone-based ASR. It is usual to train hidden Markov models (HMMs) first using monophone followed by extending the monophone HMMs into triphone HMMs. Similarly, the proposed lexicon optimization method is realized by a two-stage approach that prunes the confusable pronunciation variants using a monophone-based CM for monophone HMMs, which is identical to the approach in [8], and a triphone-based CM for triphone HMMs.

II. LEXICON EXPANSION USING A G2P MODEL

G2P conversion is used to predict pronunciation variants by aligning the graphemes of sentences or words with phonemes [9]. The simplest G2P conversion is achieved by a dictionary look-up [9]. That is, for a given input grapheme sequence, a possible pronunciation variant is obtained using a look-up table. However, this look-up table approach can cause an out-of-vocabulary since the look-up table size is in general finite.

In order to overcome the limitation of such finite dictionary, a data-driven approach can be used for the G2P conversion [9]. This is usually performed by mapping 1 to N-best after designing a joint-sequence model from a training corpus. Fig. 1 illustrates the G2P conversion result for the word “APPLE.” As shown in the figure, this word can be represented by three different pronunciation variants.

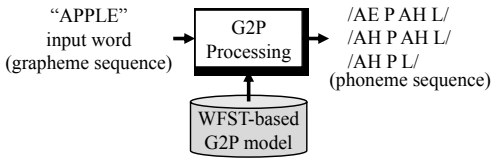

Fig. 1. Example of the G2P conversion result for a given word, "APPLE."

III. Proposed Two-Stage Lexicon Optimization Using Confusability Measure

The proposed method uses an acoustic model-based CM and a lexicon expansion approach to model the lexicon for an ASR system. Fig. 2 shows the procedure for the lexicon optimization method using Gaussian mixture acoustic models. First, the lexicon is generated for input words using a 1-to-N-best G2P model. In order to reduce pronunciation variants, the first stage of the proposed method applies a CM to prune the confusable pronunciation variants in a view of monophone acoustic models. Next, the second stage of the proposed method optimizes the output lexicon of the first-stage lexicon optimization by using the CM applied to triphone-based acoustic models. Finally, the optimized lexicon is used for constructing an ASR system employing acoustic models.

A. Confusability Measure Between Pronunciation Variants

A CM can be defined by the linguistic distance between two pronunciation variants in the expanded lexicon of a G2P model [6]. In this subsection, we define a CM as the acoustic distance between two pronunciation variants using inter-phone and inter-word distances.

Let W_i be the i-th word in the original lexicon, and let $s_{i,j}(j=1,\cdots,N)$ be the 1-to-N-best mapped phoneme sequences for W_i. Then, the CM of $s_{i,j}$ is defined as [10]

$$CM(s_{i,j}) = L(s_{i,j}) \min_{\substack{1\le k\le N_W,\, k\ne l \\ 1\le l\le N}} (D(s_{i,j}, s_{k,l})) \cdot L(s_{k,l}) \quad (1)$$

where N is the number of phoneme sequences, N_W is the total number of words of the original lexicon, and $D(x,y)$ is the dynamic programming (DP)-based pronunciation variants distance between x and y. In addition, $L(s_{k,j}) = \#(s_{k,j})/l_{\max}$, where $\#(s_{i,j})$ is the number of phonemes in $s_{i,j}$ and $l_{\max}$ is the maximum length among all the phoneme sequences in the G2P converted lexicon defined as $l_{\max} = \max_{1\le i\le N_W, 1\le j\le N}(\#(s_{i,j}))$.

B. Phoneme Distance Measure

The acoustic distance between two phonemes can be calculated by using acoustic models [10], which is defined as

$$d_{HMM}(p_1,p_2) = \frac{\sum_{Q} P(Q) \frac{1}{L} \sum_{i=1}^{L} D_N(\mathbf{N}_{q_{1i}}, \mathbf{N}_{q_{2i}})}{\sum_{Q} P(Q)} \quad (2)$$

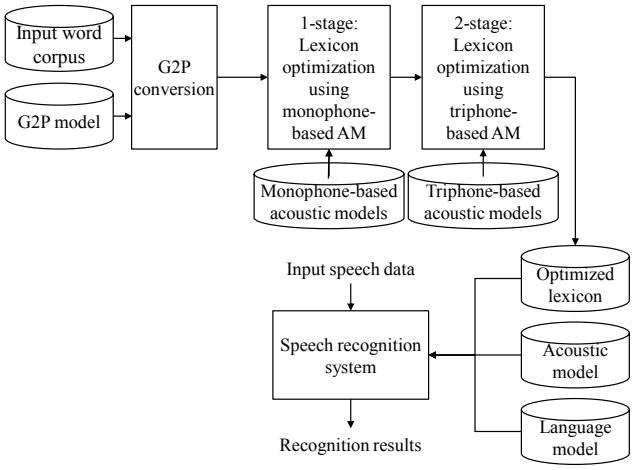

Fig. 2. Procedure of the proposed two-stage lexicon optimization method.

where Q is the alignment between the HMM states of the phonemes 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. Also, $\mathbf{N}_{q_{1i}}$ and $\mathbf{N}_{q_{2i}}$ are the Gaussian mixtures associated with the states q_{1i} and q_{2i}, respectively. In Eq. (2), $P(Q)$ is calculated by multiplying the transition probabilities of both phoneme state sequences.

In order to calculate the distance between two Gaussian components, $D_N(\cdot,\cdot)$, we apply a weighted approximation method using the Kullback-Leibler (KL) divergence between two Gaussian mixtures [11], such as,

$$D_N(\mathbf{N}_{q_{2i}}, \mathbf{N}_{q_{2i}}) = KL(\mathbf{N}_{q_{2i}} \,||\mathbf{N}_{q_{2i}}) \quad (3)$$

where

$$KL(\mathbf{N}_{q_{1i}} \,||\mathbf{N}_{q_{2i}}) \approx \sum_{n=1}^{N} \alpha_n \min_m \left[KL(N_{q_{1i}}^n \,||N_{q_{2i}}^m) + \log\frac{\alpha_n}{\beta_m} \right]. \quad (4)$$

In Eq. (4), $KL(N_{q_{1i}}^n \,||N_{q_{2i}}^m)$ is the approximation for the KL divergence between two Gaussian mixtures, $N_{q_{1i}}^n$ and $N_{q_{2i}}^m$, when $\mathbf{N}_{q_{1i}} = \sum_n \alpha_n N_{q_{1i}}^n$ and $\mathbf{N}_{q_{2i}} = \sum_m \beta_m N_{q_{2i}}^m$. Note that the KL divergence between two Gaussian distributions, $N(\mu_1,\Sigma_1)$ and $N(\mu_2,\Sigma_2)$, is defined as $\frac{1}{2}(\log|\Sigma_2| - \log|\Sigma_1| + Tr(\Sigma_2^{-1}\Sigma_1) + (\mu_1-\mu_2)^T\Sigma_2^{-1}(\mu_1-\mu_2))$.

C. Pronunciation Variant Distance Measure

By using the distance between two phonemes as described in the previous subsection, we compute the acoustic distance between two phoneme sequences, s_x and s_y. To this end, the DTW technique is also used as [10]

TABLE I
EXAMPLE OF CM SCORES BETWEEN ONE- AND TWO-STAGE LEXICON OPTIMIZATION METHODS FOR EACH PRONUNCIATION VARIANT OF THE WORD "DENTIST" OBTAINED BY 1-TO-4 BEST MAPPING

4-best Pronunciation Variant	One-stage CM score	Two-stage CM score
D EH N T IH S T	0.1246	0.06618
D EH N IH S T	0.0452	-
D EH N T AH S T	0.1201	0.06618
D EY N T IH S T	0.0866	0.04741

$$D(s_x, s_y) = \min_F \left[\frac{\sum_{k=1}^{K} d_{HMM}(p_x(k), p_y(k)) w(k)}{\sum_{k=1}^{K} w(k)} \right] \quad (5)$$

where $d_{HNMM}(p_x(k), p_y(k))$, the distance between the HMMs is defined in Eq. (2) and $w(k)$ is a weighting function used to normalize the path, F. That is, $w(k)$ is defined as

$$w(k) = i(k) - i(k-1) + j(k) - j(k-1) \quad (6)$$

where $i(1) = j(1) = 0$, and $c(k)$ 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 pronunciation variants.

D. Two-stage Lexicon Optimization

As described in Fig. 2, the first stage of the proposed method computes the acoustic distance of Eq. (5) when $p_x(k)$ and $p_y(k)$ in Eq. (5) are estimated from the monophone HMMs. After that, the G2P converted lexicon is pruned by applying Eq. (1). Next, the pronunciation variants are selected if the CM score defined in Eq. (1) is higher than a pre-defined threshold, except the pronunciation variants for each word in the original lexicon. Subsequently, the pronunciation variants with CM scores lower than the threshold are assumed to be confusable words, thus they should not be included in the pruned lexicon.

In the second stage, the triphone-based acoustic distance is used to optimize the output lexicon from the first stage. Here, $d_{HMM}(\cdot,\cdot)$, in Eq. (5) is defined when $p_x(k)$ and $p_y(k)$ are estimated from the triphone HMMs. In other words, the second stage repeats the optimization procedure for the first stage by estimating the probabilities from the triphone HMMs.

Table I provides an example of the pronunciation variants obtained by the 1-to-4-best G2P conversion for the word "DENTIST" and their CM scores. In this case, the most probable pronunciation variant is /D EH N T IH S T/. In the first stage, three pronunciation variants, /D EH N T IH S T/, /D EH N T AH S T/, and /D EY N T IH S T/, are included in the pruned lexicon if the threshold of the first stage is set to 0.05. After that, CMs of the pronunciation variants passed from the first stage are calculated using triphone-based HMMs in the second stage. If the threshold of second stage is set to 0.05, two pronunciation variants, /D EH N T IH S T/ and /D EH N T AH S T/, will remain in the lexicon.

IV. PERFORMANCE EVALUATION

A. ASR System

To evaluate the performance of the lexicon optimization method, we constructed a baseline ASR system (Baseline), an ASR system using a 1-to-4-best G2P-converted pronunciation dictionary, where a WFST-based G2P model with 4 pronunciation variants per word was generated, and ASR systems based on lexicons pruned by the proposed lexicon optimization method using acoustic distance based on KL divergence. The baseline system was constructed using the Kaldi speech recognition toolkit [12] with 7,138 utterances from Wall Street Journal (WSJ0) [13]. In addition, a CMU dictionary was used for the baseline lexicon [14].

As an ASR feature, 39-dimensional mel-frequency cepstral coefficients (MFCCs) were used, and the cepstral mean normalization (CMN) was applied to the feature vector. The acoustic model was constructed by means of concatenating context-dependent HMMs. A trigram language model (LM) was constructed from a set of sentences from the WSJ0 with a vocabulary of 20k different words.

The evaluation test dataset (Eval set) was also extracted from the WSJ0 and was composed of 333 utterances containing 5,643 different words (Nov' 92). In addition, in order to find the best thresholds for the proposed methods, 403 utterances containing 6,722 different words in the si_dt_20 development test set of WSJ0 were used and referred to as Dev set.

B. Comparison of ASR Performance

We evaluated the performance of an ASR system using the lexicon pruned by the proposed CM and compared it with the baseline lexicon (CMU dictionary), 1-to-4 best G2P converted pronunciation dictionary, the optimized lexicon using a) the Levenshtein (LEV) distance, and b) two-stage optimized lexicon which used acoustic distance based on the KL divergence for monophone and triphone acoustic models.

Next, in order to investigate the effect of the threshold for the two-stage lexicon optimization method, we evaluated the performance of the proposed method on Dev set by changing the threshold from 0.01 to 0.15 at a step of 0.01. As shown in Fig. 3, word error rates (WERs) were lowered by applying the first stage of the proposed method by increasing the threshold. Especially, it was found that the lowest WER was achieved when the threshold was set 0.06, but the WER was not changed even when the threshold was increased. In addition, the second stage of the proposed method provided the lowest WER when the threshold was set to 0.07, but the WER was not lowered any more even when the threshold was increased. Consequently, we set the thresholds for the first and second stage as 0.06 and 0.07 for the evaluation of the Eval set with the optimized lexicon.

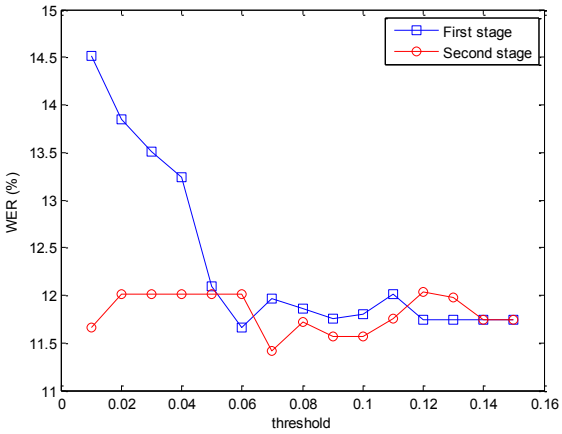

Fig. 3. Comparison of average word error rates of an ASR system employing lexicons optimized by acoustic distance using KL divergence according to the threshold in the first stage and the second stage.

Table II shows WERs for the different methods described above on the task of Dev and Eval set. As shown in the table, the WERs were lowered with the proposed two-stage optimized lexicon. Consequently, we achieved relative WER reduction of 18.88% and 3.50% with the two-stage optimized lexicon, compared to that using 1-to-4-best G2P conversion and one-stage optimized lexicon, respectively.

V. Conclusion

In this paper, we proposed a two-stage lexicon optimization method based on a statistical acoustic CM in order to reduce the confusable pronunciation variants of lexicons constructed by the G2P model. In particular, the first stage and the second stage of the proposed method were applied to monophone-based HMMs and triphone-based ones, respectively. By doing this, the proposed method could achieve more optimized lexicon than a conventional method applied to only monophone-based HMMs. It was demonstrated from ASR experiments that an ASR system employing a lexicon optimized by the proposed method provided a relative WER reduction of 18.88% and 3.50%, compared to those by a 1-to-4-best G2P conversion and the conventional method applied to monophone-based HMMs, respectively.

Acknowledgment

This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the government of Korea (MSIP) (No. 2015R1A2A1A05001687), and by the MSIP, Korea, under the ITRC support program (IITP-2015-H8501-15-1016) supervised by the IITP.

References

[1] G. Saon and J.-T. Chien, "Large-vocabulary continuous speech recognition systems – a look at some recent advances," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 18–33, 2012.

Table II
Performance Comparison of ASR Systems Using Lexicons Optimized by Different Methods

Lexicon	WER (%)	
	Dev set	Eval set
- Baseline (CMU dictionary)	11.93	12.53
- 4-best pronunciation dictionary	15.01	13.93
Optimized lexicon using different inter-pronunciation variant distance		
a) Optimized lexicon LEV-based distance	14.43	12.23
b) Optimized lexicon using KL divergence		
- Monophone single Gaussian	12.90	12.42
- One-stage monophone Gaussian Mixtures	11.66	11.70
- Two-stage triphone Gaussian Mixtures	11.41	11.29

[2] K. Livescu, E. Fosler-Lussier, and F. Metze, "Subword modeling for automatic speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 44–57, 2012.
[3] J. Novak, N. Minematsu, and K. Hirose, "WFST-based grapheme-to-phoneme conversion: open source tools for alignment, model-building and decoding," in *Proc. FSMNLP*, San Sebastian, Spain, 2012, pp. 45–49.
[4] S. Hahn, P. Lehnen, S. Wiesler, R. Schlüter, and H. Ney, "Improving LVCSR with hidden conditional random fields for grapheme-to-phoneme conversion," in *Proc. Interspeech*, Lyon, France, 2013, pp. 495–499.
[5] D. Harwath and J. Glass, "Speech recognition without a lexicon – bridging the gap between grapheme and phonetic systems," in *Proc. Interspeech*, Singapore, 2014, pp. 2655–2659.
[6] M. A. Kim, Y. R. Oh, and H. K. Kim, "Optimizing multiple pronunciation dictionary based on a confusability measure for non-native speech recognition," in *Proc. IASTED*, Innsbruck, Austria, 2008, pp. 215–220.
[7] T. Jitsuhiro, S. Takahashi, and K. Aikawa, "Rejection of out-of-vocabulary words using phoneme confidence likelihood," in *Proc. ICASSP*, Seattle, WA, 1998, pp. 217–220.
[8] N. K. Kim, W. K. Seong, and H. K. Kim, "Lexicon optimization for WFST-based speech recognition using acoustic distance based confusability measure and G2P conversion," in *Proc. IWSDS*, Busan, Korea, 2015, paper 12.
[9] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech Communication*, vol. 50, no. 5, pp. 434–451, 2008.
[10] J. Anguita, J. Hernando, S. Peillon, and A. Bramoulle, "Detection of confusable words in automatic speech recognition," *IEEE Signal Processing Letters*, vol. 12, no. 8, pp. 585–588, 2005.
[11] J. Goldberger, S. Gordon, and H. Greenspan, "An efficient image similarity measure based on approximations of KL-divergence between two Gaussian mixtures," in *Proc. ICCV*, Nice, France, 2003, pp. 487–493.
[12] D. Povey, *et al.*, "The Kaldi speech recognition toolkit," in *Proc. IEEE ASRU*, Honolulu, HI, 2011, pp. 1–4.
[13] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," in *Proc. ICSLP*, Alberta, Canada, 1992, pp. 899–902.
[14] CMU Pronouncing Dictionary, [Online] Available: http://www.speech.cs.cmu.edu/cgi-bin/cmudict, Carnegie Mellon University.