Performance Improvement of Dysarthric Speech Recognition Using Context-Dependent Pronunciation Variation Modeling Based on Kullback-Leibler Distance

Woo Kyeong Seong, Ji Hun Park, and Hong Kook Kim

School of Information and Communications
Gwangju Institute of Science and Technology (GIST)
1 Oryong-dong, Buk-gu, Gwangju 500-712, Korea
{wkseong, jh_park, hongkook}@gist.ac.kr

Abstract. In this paper, we propose context-dependent pronunciation variation modeling based on the Kullback-Leibler (KL) distance for improving the performance of dysarthric automatic speech recognition (ASR). To this end, we construct a triphone confusion matrix based on KL distances between triphone models, and build a weighted finite state transducer (WFST) from the triphone confusion matrix. Then, dysarthric speech is recognized by a baseline ASR system. The corresponding phoneme sequence of the recognized sentence is then passed through the WFST to correct recognition errors. It is shown from dysarthric ASR experiments that average word error rate of an ASR system employing an error correction based on the proposed method is relatively reduced by 16.54% and 3.34%, compared to those of an ASR system without any error correction and using an error correction based on a conventional context-dependent phoneme confusion matrix, respectively.

Keywords: Dysarthric speech recognition, Kullback-Leibler distance, context-dependent pronunciation variation model, weighted finite state transducer, speech recognition error correction

1 Introduction

Dysarthria refers to impairment of speech resulting from damaged control of the oral, pharyngeal, or laryngeal articulators [1]. The impairment of a dysarthric speech results in pronunciation variations such as deletion, substitution, insertion, and distortion of particular phonemes. These pronunciation variations can cause degradation in dysarthric automatic speech recognition (ASR) performance [2]. In order to compensate for the pronunciation variations, there have been reported several works which automatically extracted pronunciation variation rules based on a data-driven approach [3][4].

However, in the case of dysarthric speech, the pronunciation variation rules extracted by the data-driven approach are not reliable because of the insufficiency of the database [5]. For this reason, a model-based approach can be considered as an alterna-

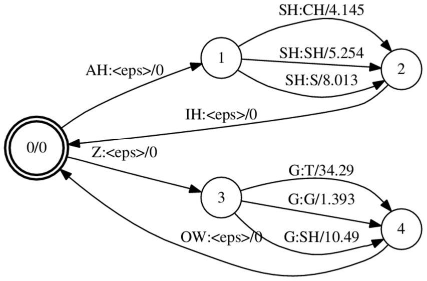

Fig. 1. An example of the weighted finite state transducer obtained from a triphone confusion matrix.

tive for extracting pronunciation variation rules for dysarthric speech. Therefore, a context-dependent (CD) pronunciation modeling method is proposed in order to correct speech recognition errors. Here, the proposed method is based on Kullback-Leibler (KL) distance between acoustic models, and it is incorporated into the construction of a weighted finite state transducer for error correction.

2 Proposed KL Distance Based Context-Dependent Pronunciation Variation Modeling Method

In order to obtain pronunciation variation rules, we first calculate the KL distances between acoustic models. In this paper, acoustic models are composed of triphones, where each triphone is represented as a three-state, left-to-right hidden Markov model (HMM) with four Gaussian mixtures. Thus, a KL distance is defined as [6]

$$KL(L-C_i+\mathrm{R}\,\|\,L-C_j+\mathrm{R}) = \sum_{k=1}^{4}\omega_{i,k}KL(b_{i,k}\,\|\,b_{j,k}) + \sum_{k=1}^{4}\omega_{i,k}\log\frac{\omega_{i,k}}{\omega_{j,k}} \qquad (1)$$

where $L-C_i+\mathrm{R}$ means a triphone whose center phoneme is C_i while its left and right context are defined as phoneme L and phoneme R, and $\omega_{i,k}$ is the k-th Gaussian mixture weight of the triphone of $L-C_i+\mathrm{R}$. In addition, $KL(b_{i,k}\,\|\,b_{j,k})$ is the KL distance between the k-th Gaussian mixtures of $L-C_i+\mathrm{R}$ and $L-C_j+\mathrm{R}$, defined as

$$KL(b_{i,k}\,\|\,b_{j,k}) = \frac{1}{2}(\ln(\frac{|\Sigma_{j,k}|}{|\Sigma_{i,k}|}) + \mathrm{tr}(\Sigma_{j,k}^{-1}\Sigma_{i,k}) + (\mu_{j,k}-\mu_{i,k})^{\mathrm{T}}\Sigma_{j,k}^{-1}(\mu_{j,k}-\mu_{i,k}) - K) \qquad (2)$$

where $\mu_{i,k}$ and $\Sigma_{i,k}$ represent a mean vector and a covariance matrix of $b_{i,k}$, respectively, and K indicates the feature vector dimension that is 39 in this paper. Therefore, $KL(L-C_i+\mathrm{R}\,||\,L-C_j+\mathrm{R})$ in Eq. (1) corresponds to the degree of pronunciation variation from phoneme C_i to phoneme C_j with the same left and right context as L and R. Next, we obtain a triphone confusion matrix whose *ij*-th element is $KL(L-C_i+\mathrm{R}\,||\,L-C_j+\mathrm{R})$. Finally, the triphone confusion matrix is transformed to a weighted finite state transducer (WFST). Each element in the triphone confusion matrix is represented as a self-loop, in which the WFST outputs a phoneme C_j with $KL(L-C_i+\mathrm{R}\,||\,L-C_j+\mathrm{R})$ when a phoneme C_i with $L-C_i+\mathrm{R}$ comes into the WFST.

Fig. 1 illustrates an example of WFST obtained from the triphone confusion matrix. If a phoneme /SH/ with /AH/-/SH/+/IH/ passes through the WFST, the WFST outputs the phoneme /CH/, /SH/, or /S/ with the KL distance values of 4.145, 5.254, or 8.013, respectively. Here, a higher value indicates more confusability.

3 Performance Evaluation

To evaluate the performance of the proposed method, we constructed the following three ASR systems: a baseline ASR system (Baseline), an ASR system with error correction based on a data-driven CD phoneme confusion matrix (EC-DD) [4], and an ASR system with error correction based on the proposed CD phoneme confusion matrix (EC-KL). For the baseline ASR system, 10,000 utterances of British English sentences from 92 non-dysarthric speakers were used as a training database [7]. As a recognition feature, 39-dimensional mel-frequency cepstral coefficient vectors were used. The acoustic models were composed of 22,433 triphones, where each triphone was represented as a three-state, left-to-right hidden Markov model (HMM) with four Gaussian mixtures. In addition, triphone models in the baseline ASR system were adapted to each dysarthric speaker with 34 utterances from the Nemours database [8]. The lexicon size was 112 words, and a back-off bigram model was employed as a language model. For EC-DD and EC-KL, CD phoneme confusion matrices were constructed by using 34 utterances that were identical to those used for the adaptation of the triphone models. As a test database, we used 400 utterances from ten dysarthric speakers [8].

Table 1 compares average word error rates (WERs) of Baseline, EC-DD, and EC-KL. As shown in the table, EC-KL provided the lowest WER and relatively reduced average WER by 3.34%, compared to EC-DD.

4 Conclusion

In this paper, we proposed a context-dependent pronunciation variation modeling method based on KL distance for improving the performance of dysarthric speech recognition. First, a triphone confusion matrix was constructed based on KL distances

Table 1. Comparison of average word error rates (%) of the baseline ASR system (Baseline) and ASR systems with error correction based on a conventional data-driven CD phoneme confusion matrix (EC-DD) and proposed CD phoneme confusion matrix based on KL distances (EC-KL).

System	Baseline	EC-DD	EC-KL
Word error rate (%)	36.75	31.73	30.67

between triphone models, and a weighted finite state transducer (WFST) was built from the triphone confusion matrix. Then, the WFST was used to correct speech recognition errors by passing through a phoneme sequence of a recognized sentence. It was shown from speech recognition experiments that a dysarthric ASR system employing the proposed pronunciation variation modeling method provided a relative WER reduction by by 3.34%, compared to that employing a data-driven pronunciation variation modeling method.

Acknowledgments. This research was supported in part by the R&D Program of MKE/KEIT [1003641, Development of an embedded key-word spotting speech recognition system individually customized for disabled persons with dysarthria], and by the basic research project through a grant provided by the Gwangju Institute of Science and Technology (GIST) in 2012.

References

1. Haines, D.: Neuroanatomy: an Atlas of Structures, Sections, and Systems. Lippingcott Williams and Wilkins, Hagerstown, MD (2004)
2. Hasegawa-Johnson, M., Gunderson, J., Perlman, A., Huang, T.: HMM-based and SVM-based recognition of the speech of talkers with spastic dysarthria. In: Proceedings of ICASSP, Toulouse, France, pp. 1060-1063 (2006)
3. Morales, S. O. C., Cox, S. J.: Modeling errors in automatic speech recognition for dysarthric speakers. EURASHIP Journal on Advances in Signal Processing, Article ID 308340, 14 pages (2009)
4. Seong, W. K., Park, J. H., Kim, H. K.: Dysarthric speech recognition error correction using weighted finite state transducers based on context-dependent pronunciation variation. In: Proceedings of International Conference on Computers Helping People with Special Needs, Linz, Austria, pp. 475-482 (2012)
5. You, H., Alwan, A.: A statistical acoustic confusability metric between hidden Markov models. In: Proceedings of ICASSP, Los Angeles, CA, pp. 745-748 (2007)
6. Do, M.: Fast approximation of Kullback-Leibler distance for dependence trees and hidden Markov models. IEEE Signal Processing Letters, 10(4), 115-118 (2003)
7. Fransen, T. J., Pye, D., Foote, J., Renals, S.: WSJCAM0: a British English speech corpus for large vocabulary continuous speech recognition. In: Proceedings of ICASSP, Detroit, MI, pp. 81-84 (1995)
8. Menendez-Pidal, X., Polikoff, J. B., Peters, S. M., Leonzio, J. E., Bunnell, H. T.: The Nemours database of dysarthric speech. In: Proceedings of ICSLP, Philadelphia, PA, pp. 1962-1965 (1996)