Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 67 (2015) 154–166

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers

Wenping Hu [a,b], Yao Qian [b,*], Frank K. Soong [b], Yong Wang [a]

[a] University of Science and Technology of China, Hefei 230026, China
[b] Microsoft Research Asia, Beijing 100080, China

Received 28 July 2014; received in revised form 19 December 2014; accepted 27 December 2014
Available online 8 January 2015

Abstract

Mispronunciation detection is an important part in a Computer-Aided Language Learning (CALL) system. By automatically pointing out where mispronunciations occur in an utterance, a language learner can receive informative and to-the-point feedbacks. In this paper, we improve mispronunciation detection performance with a Deep Neural Network (DNN) trained acoustic model and transfer learning based Logistic Regression (LR) classifiers. The acoustic model trained by the conventional GMM-HMM based approach is refined by the DNN training with enhanced discrimination. The corresponding Goodness Of Pronunciation (GOP) scores are revised to evaluate pronunciation quality of non-native language learners robustly. A Neural Network (NN) based, Logistic Regression (LR) classifier, where a general neural network with shared hidden layers for extracting useful speech features is pre-trained firstly with pooled, training data in the sense of transfer learning, and then phone-dependent, 2-class logistic regression classifiers are trained as phone specific output layer nodes, is proposed to mispronunciation detection. The new LR classifier streamlines training multiple individual classifiers separately by learning the common feature representation via the shared hidden layer. Experimental results on an isolated English word corpus recorded by non-native (L2) English learners show that the proposed GOP measure can improve the performance of GOP based mispronunciation detection approach, i.e., 7.4% of the precision and recall rate are both improved, compared with the conventional GOP estimated from GMM-HMM. The NN-based LR classifier improves the equal precision–recall rate by 25% over the best GOP based approach. It also outperforms the state-of-art Support Vector Machine (SVM) based classifier by 2.2% of equal precision–recall rate improvement. Our approaches also achieve similar results on a continuous read, L2 Mandarin language learning corpus.
© 2014 Elsevier B.V. All rights reserved.

Keywords: Computer-aided language learning; Mispronunciation detection; Deep neural network; Logistic regression; Transfer learning

1. Introduction

The current globalization of people in regions of different languages has accelerated the demand of foreign language proficiency. For non-native language learners, the one-to-one teacher–student interaction and communication is the most effective way, but it can be too pricey and unaffordable for many learners. Computer Aided Language Learning (CALL) systems, powered by the advancement of speech technology, can bridge the gap between disproportional supply and demand in language learners and teachers and have become ubiquitous learning tools with handy smart phones, tablets, laptop computers, etc. However, as an indispensable component of CALL system, phone level mispronunciation detection, which aims at

* Corresponding author. Tel.: +86 10 59174292.
E-mail addresses: hwping@mail.ustc.edu.cn (W. Hu), yaoqian@microsoft.com (Y. Qian), frankkps@microsoft.com (F.K. Soong), yongwang@ustc.edu.cn (Y. Wang).

http://dx.doi.org/10.1016/j.specom.2014.12.008
0167-6393/© 2014 Elsevier B.V. All rights reserved.

detecting or identifying pronunciation errors or deficiency at the phone level in a high precision, is still challenging.

There are a great deal of research work on mispronunciation detection. An in-depth review of automatic error detection in pronunciation training is given by Witt (2012), in which pronunciation errors are divided into two types, i.e., phonemic error and prosodic error, and the phonemic error is further categorized into many sub error types. Features used for detecting the pronunciation errors are mostly extracted from the output of an HMM based speech recognizer. Kim et al. (1997) compare three HMM based scores, e.g., log-likelihood score, log-posterior probability score and segment duration score, in pronunciation quality evaluation for some specific phones and find log-posterior probability scores have the highest correlation with human ratings. Besides this HMM based log-posterior probability based method, Franco et al. (1999) further adopt the Log-Likelihood Ratio (LLR) between native-like and non-native models as the measure for mispronunciation detection. The results show that LLR based method has better overall performance than the posterior based method, but it needs to be trained with specific examples of the target non-native user population. Witt and Young (2000) introduce a Goodness Of Pronunciation (GOP) method, a variation of the posterior probability, for phone level pronunciation scoring. This GOP measure is later widely used in pronunciation evaluation and mispronunciation detection tasks. Some variations of GOP measure are also proposed in the last decade. Zhang et al. (2008) propose a Scaling Log-Posterior Probability (SLPP) method for Mandarin mispronunciation detection and achieve considerable performance improvement. Wang and Lee (2012) also combine the GOP based method with error pattern detectors for phone mispronunciation diagnosis with a serial and parallel structure and find that the serial structure can reduce the average error rate and improve diagnosis feedback. To improve the scores generated by the traditional GMM-HMM based speech recognizer, some discriminative training algorithms have been applied, e.g. Maximum Mutual Information Estimation (MMIE) (Bahl et al., 1986), Minimum Classification Error (MCE) (Juang et al., 1997) and Minimum Phone Error (MPE) and Minimum Word Error (MWE) (Povey and Woodland, 2002). Yan and Gong (2011) introduce the discriminatively refined acoustic models by MPE into pronunciation proficiency evaluation. Qian et al. (2010) also investigate using MWE-trained HMM models to minimize mispronunciation detection errors for L2 English learners.

Mispronunciation detection can be formulated as a 2-class classification task. Many classifier based approaches are applied to mispronunciation detection to improve its performance. Ito et al. (2005) use a decision tree based method to set thresholds for different kinds of mispronunciations and achieve a significant improvement, compared with a universal threshold. Truong (2004) uses decision tree and Linear Discriminant Analysis (LDA) to distinguish different pronunciation errors of L2 learners of Dutch based on some discriminative features, e.g., formants and durations. A specific classifier is built for each mispronunciation pattern. Experimental results show that the classifier based approach has a good performance of detecting vowel pronunciation errors but a poor performance of detecting consonant pronunciation errors. It also shows that LDA yields a better detection performance than decision tree. Strik et al. (2009) draw a comparison of four different approaches, i.e., GOP score, decision tree and LDA with two kinds of features, acoustic–phonetic features and Mel-Frequency Cepstrum Coefficients (MFCCs), in distinguishing two Dutch phones: the velar fricative/x/ and the velar plosive/k/. The result shows that LDA based methods outperform the GOP and decision tree based methods. Doremalen et al. (2009) build a set of Support Vector Machine (SVM) classifiers to detect substitution errors for Dutch vowels and improve the performance by combining different features: confidence measures (Jiang, 2005), phonetic features and MFCCs. Wei et al. (2009) also apply SVM to classify the correct and incorrect pronunciations of Mandarin syllables and improve the performance by enhancing the discrimination of pronunciation variation with Pronunciation Space Models (PSMs). Other research work, which applies SVM as the underlying classifier for mispronunciation detection in different scenarios, can be found in (Jie and Xu, 2009; Xu et al., 2009; Yoon et al., 2010; Hirabayashi and Nakagawa, 2010). Additionally, to improve the accuracy of pronunciation error detection, the knowledge of L1 (the native language), e.g. the set of common pronunciation errors made by non-native speakers, can be used to contrast L1–L2 phone confusion pairs to build a more custom-made mispronunciation detection system.

Recently, Deep Neural Network (DNN), which attempts to model high-level abstractions in data, has significantly improve the discrimination of acoustic models in speech recognition (Hinton et al., 2012a). Transfer learning (Bengio et al., 2013), which can exploit commonalities between different learning tasks in order to share statistical strength, is successfully employed to multilingual speech recognition (Huang et al., 2013). Application of using Deep Belief Nets (DBN) to mispronunciation detection and diagnosis in L2 English has been tried by Qian et al. (2012), and a significant improvement on word pronunciation relative error rate was obtained on L1 (Cantonese)-dependent English learning corpus. We have used DNN trained acoustic models for English pronunciation quality scoring. We find the GOP scores estimated from DNN outputs correlate better with human expert's evaluations than conventional GOP scores obtained from a conventional GMM based system (Hu et al., 2013). We have also investigated different pitch embedding methods in DNN acoustic model training to further enhance the discrimination of acoustic models and detect the pronunciation errors caused by misusing lexical stress or lexical tone for L2 language learners (Hu et al., 2014a). In this study, we focus on the detection of phonemic error, i.e., phone-level mispronunciation. We

Download English Version:

https://daneshyari.com/en/article/6961164

Download Persian Version:

https://daneshyari.com/article/6961164

Daneshyari.com