THE IMPORTANCE OF MISPRONUNCIATION DETECTION IN NON-NATIVE LANGUAGE

Khidirova Sarvinoz
The Student Of Uzswlu

Abstract: This article is based on providing basic information on mispronunciation detection in non-native language. In linguistics, mispronunciation is the act of pronouncing a word incorrectly. The matter of what is or is not mispronunciation is a contentious one, and there is disagreement about the extent to which the term is even meaningful. Languages are pronounced in different ways by different people, depending on such factors as the area they grew up in, their level of education, and their social class. Even within groups of the same area and class, different people can have different ways of pronouncing certain words.

Key words: Mispronunciation detection and diagnosis, Acoustic, Phonetic and Linguistic (APL) embedding.

Mispronunciation detection and diagnosis (MDD) is designed to identify pronunciation errors and provide instructive feedback to guide non-native language learners, which is a core component in computer-assisted pronunciation training (CAPT) systems. However, MDD often suffers from the data-sparsity problem due to that collecting non-native data and the associated annotations is time-consuming and labor-intensive. To address this issue, we explore a fully end-to-end (E2E) neural model for MDD, which processes learners' speech directly based on raw waveforms. Compared to conventional hand-crafted acoustic features, raw waveforms retain more acoustic phenomena and potentially can help neural networks discover better and more customized representations. To this end, our MDD model adopts a co-called SincNet module to take input a raw waveform and covert it to a suitable vector representation sequence. SincNet employs the cardinal sine (sinc) function to implement learnable band pass filters, drawing inspiration from the convolutional neural network (CNN). By comparison to CNN, SincNet has fewer parameters and is more amenable to human interpretation. Extensive experiments are conducted on the L2-ARCTIC dataset, which is a publicly-available non-native English speech corpus compiled for research on CAPT. We find that the sinc filters of SincNet can be adapted quickly for non-native language learners of different nationalities. Furthermore, our model can achieve comparable mispronunciation detection performance in relation to state-of-the-art E2E MDD models that take input the standard handcrafted acoustic features. Besides that, our model also provides considerable improvements on phone error rate (PER) and diagnosis accuracy.

Many mispronunciation detection and diagnosis (MD&D) research approaches try to exploit both the acoustic and linguistic features

as input. Yet the improvement of the performance is limited, partially due to the shortage of large amount annotated training data at the phoneme level. Phonetic embeddings, extracted from ASR models trained with huge amount of word level annotations, can serve as a good representation of the content of input speech, in a noise-robust and speaker-independent manner.

These embeddings when used as implicit phonetic supplementary information, can alleviate the data shortage of explicit phoneme annotations. We propose to utilize Acoustic, Phonetic and Linguistic (APL) embedding features jointly for building a more powerful MD&D system.. Index Terms — Computer aided Pronunciation Training, Mispronunciation Detection and Diagnosis, Phoneme Recognition, Acoustic-phonetic-linguistic Embeddings.

The development of Computer-aided Pronunciation Training(CAPT) system empowers language learners a convenient way to practice their pronunciations[1, 2, 3], especially for those who have little access to professional teachers.

Mispronunciation Detection and Diagnosis (MD&D) is a key part of CAPT and several methods have been proposed to tackle it. Goodness of Pronunciation (GOP)[4], developed by Witt and Young, computes scores based on log-posterior probability from acoustic models and then detects mispronunciation with phone-dependent thresholds. Even though these kinds of approaches provide scores for mispronunciation detection[5, 6, 7], they cannot provide sufficient diagnosis information for pronunciation correction. To better obtain diagnosis information, Extended Recognition Network (ERN)[8, 9, 10] extends the decoding stage of Automatic Speech Recognition (ASR) by modeling pre-defined context-dependent phonological rules. However, ERN fails to deal with the mispronunciation patterns which are absent in training data or manual rules. Additionally, when too many phonological rules are included in ERN, recognition accuracy may be affected, thus leading to unreliable MD&D feedbacks.

Work performed as intern in Microsoft

Actually, Computer-aided pronunciation training (CAPT) systems provide feedback to second language learners on their pronunciation quality, with positive impacts on learning and motivation [1]. One family of CAPT systems frames the problem as a phone recognition task, using non-native data during training. These systems identify pronunciation errors by comparing the phonetic transcription of a student's speech to a native target sequence using dynamic programming algorithms. Another family of CAPT systems frames the problem as detection of mispronunciations, generating scores that are then thresholded for the final decision. These systems can

be classified into two groups. Those that do not use non-native data during training rely on automatic speech recognition (ASR) systems trained with native speakers, and generate pronunciation scores using the acoustic model's outputs. The most widely used approach in this family is called Goodness of Pronunciation (GOP). The second group uses non-native data to directly train the system to distinguish correctly- from incorrectly-pronounced segments using a variety of input features and classifiers. Recently, transfer learning techniques have been used to mitigate the problem of data scarcity that is the norm in the task. In these approaches, deep neural networks (DNNs) models trained for ASR or on a self-supervised fashion are fine-tuned to detect mispronunciations.

With the growing population of second language learners, there is a strong need for additional language learning resources. Kachru estimates there are 533 million English learners in India and China alone a number greater than the total population of the USA, UK, and Canada combined. With such a huge demand, there is an acute shortage of qualified teachers. Computer-assisted language learning (CALL) applications can supplement existing learning resources and provide unique benefits to learner in terms of accessibility, reduced anxiety, and individualized instruction.

Effective language learning tools, and particularly pronunciation training, needs to provide learners with detailed corrective feedback. The automatic pronunciation scores at the word-level or sentence-level correlate highly with human raters but fail to lead to measurable improvement in learner's overall pronunciation.

However, locating mispronunciations at the phone-level to learners has been shown to lead to statistically significant improvement for the production of those targeted phones. Speech recognition systems must be specially designed for computer-assisted pronunciation training (CAPT) in order to support detailed corrective feedback while still obtaining satisfactory performance.

Basically, using computers to help students learn and practice a new language has long been seen as a promising area for the use of automatic speech recognition (ASR) technology. It could allow spoken language to be used in many ways in language-learning activities, for example by supporting different types of oral practice and enabling feedback on various dimensions of language proficiency, including language use and pronunciation quality. A desirable feature of the use of speech technology for computeraided language learning (CALL) is the ability to provide meaningful feedback on pronunciation quality. In this area of pronunciation scoring, the smaller the unit to be scored, the higher the uncertainty in the associated score. Currently, the most reliable estimates of pronunciation quality are overall levels obtained from a paragraph composed of several sentences that can be used to characterize the speaker s overall pronunciation proficiency. At this level, it has been shown that automatic scoring performs as well as human scoring. For many CALL applications we would like to score smaller units, to allow the student to focus on

specific aspects of his or her speech production. For instance, overall pronunciation scoring can be obtained at the sentence level with a level of accuracy that, while lower than that of human scoring, can nonetheless provide valuable feedback for language learning. More detailed feedback, at the level of individual phones, can direct attention to specific phones that are mispronounced.

REFERENCES:

[1] Keelan Evanini and Xinhao Wang, "Automated speech scoring for non-native middle school students with multiple task types," in Proc. Interspeech 2013, 2013, pp. 2435–2439.

[2] Yanlu Xie, Xiaoli Feng, Boxue Li, Jinsong Zhang, and Yujia Jin, "A mandarin l2 learning app with mispronunciation detection and feedback," in Proc. Interspeech 2020, 2020, pp. 1015–1016.

[3] Ke Shi, Kye Min Tan, Richeng Duan, Siti Umairah Md. Salleh, Nur Farah Ain Suhaimi, Rajan Vellu, Ngoc Thuy Huong Helen Thai, and Nancy F. Chen, "Computer-assisted language learning system: Automatic speech evaluation for children learning malay and tamil," in Proc. Interspeech 2020, 2020, pp. 1019–1020.

[4] S.M.Witt and S.J.Young, "Phone-level pronunciation scoring and assessment for interactive language learning," Speech communication, vol. 30, no. 2-3, pp. 95–108, 2000.

[5] Joost van Doremalen, Catia Cucchiarini, and Helmer Strik, "Using non-native error patterns to improve pronunciation verification," in Proc. Interspeech 2010, 2010, pp. 590–593.

[6] Wenping Hu, Yao Qian, Frank K.Soong, and Yong Wang, "Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers," Speech Communication, vol. 67, pp. 154–166, 2015.