Developing Speech Recognition and Synthesis Technologies to Support Computer-Aided Pronunciation Training for Chinese Learners of English[*]

Helen Meng

Human-Computer Communications Laboratory
The Chinese University of Hong Kong
hmmeng@se.cuhk.edu.hk

Abstract. We describe ongoing research in the development of speech technologies that strives to raise the efficacy of computer-aided pronunciation training, especially for Chinese learners of English. Our approach is grounded on the theory of language transfer and involves a systematic phonological comparison between the primary language (L1 being Chinese) and secondary language (L2 being English) to predict possible segmental and suprasegmental realizations that constitute mispronunciations in L2 English. The predictions are validated based on a specially designed corpus that consists of several hundred hours of L2 English speech. The speech data supports the development of automatic speech recognition technologies that can detect and diagnose mispronunciations. The diagnosis aims to support the design of pedagogical and remedial instructions, which involves text-to-speech synthesis technologies in audiovisual forms.

1 Introduction

English is the lingua franca of our world. It is the second language (L2) most actively studied across Asia, as well as an official language or working language used in education, government, media, etc. in many regions. Hence, acquiring communicative competence in English is of prime importance. It has been estimated that by 2010 there will be 2 billion English learners worldwide, and the proportion in Asia alone will exceed the number of native speakers (Asia Economic News, 2006). Second language learning, specifically pronunciation learning, involves correct perception and production of sounds in the target language. The learning process tends to be influenced by well established perceptions of sounds and articulatory motions in the primary language (L1). This cross-linguistic influence is often referred as language transfer. Negative transfer of L1 features causes inaccuracies and errors in L2 speech productions, which impede intelligibility. Consequently, the study of L2 speech productions (i.e. the "interlanguage" of learners who have not acquired native-like proficiency) is of great interest to phoneticians, linguists, language educators, as well as technologists engaged in the development of CAPT (computer-aided pronunciation training) systems. These applications can complement classroom teaching and provide unique benefits to the learner in terms of accessibility, reduced anxiety and individualized instructions.

Effective pronunciation training tools need to provide learners with detailed mispronunciation detection, diagnosis and corrective feedback (Ehsani and Knodt, 1998). Previous work has shown that 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 (Precoda *et al.*, 2000). However, locating mispronunciations at the phone-level to learners has been shown to lead to statistically significant improvement for the

[*] Joint Work with Alissa Harrison, Pauline Lee, Wai-Kit Lo, Lan Wang and Virginia Yip

production of those targeted phones (Neri *et al.*, 2006). Moreover, diagnostic feedback to learners (e.g. "you inserted a vowel at the end of the word") has also been shown to lead to significant improvements in pronunciation training (Kim *et al.*, 2004).

2 Speech Technologies for CAPT

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 (Ehsani and Knodt, 1998). Although large vocabulary continuous speech recognition (LVCSR) systems are widely available in the commercial market, they are not necessarily appropriate for pronunciation training for non-native speakers. Generally, LVCSR systems are designed to *accommodate* a wide variety of accents and non-standard pronunciations. They are not intended to be used as a tool for discriminating phonetically similar pronunciations of a given word. Free phone recognition, in principle, could support a CAPT tool in providing detailed phone-level feedback to a learner. However, highly accurate free phone recognition is still not possible for native-speech, and is only expected to be even more difficult for the interlanguage of second language learners. Furthermore, non-native mispronunciations may be due to a diversity of factors, such as imperfect understanding of semantics, syntax, morphology, phonology, coarticulatory effects and letter-to-sound rules. As an initial step, we focus on the use of comparative phonological analysis between L1 (Cantonese) and L2 (English) to derive a set of 51 context-dependent phonological rules for mispronunciation prediction. The predicted deviations are validated against observations in the CU-CHLOE (Chinese University Chinese Learners of English) corpus. This consists of Chinese-accented English speech recordings from 100 speakers reading the AESOP fable, "The North Wind and the Sun", that we have collected to support our investigation. The phonological rules are incorporated in the automatic speech recognizer as finite-state transducers (FSTs) that form an extended recognition network (ERN) (Harrison *et al.*, 2009). The ERN enables us to model the target pronunciations and the related non-native mispronunciations in a unified framework. Evaluation based on a disjoint test set shows that the ERN achieves a phone-level recognition accuracy of 73.0% and models 58.4% of the mispronounced words in the test set. Automatic alignment between the recognized and target phone strings with a phonetically-sensitive alignment algorithm enables us to diagnose specific phonetic confusions and generate corrective feedback. Insufficient coverage of the mispronounced words in a disjoint test set is due to sparsity of training data and insensitive acoustic models. To raise the performance of mispronunciation detection, we have also devised a decision fusion methodology that augments the linguistically-motivated approach described above with pronunciation scoring and phone-dependent thresholding (Lo *et al.*, 2008). Such augmentation brought 30% relative improvement over the linguistically-motivated approach, in terms of phoneme decision errors (i.e. sum of the false acceptance and false rejection errors).

We have also begun to develop text-to-speech synthesis technologies that can automatically synthesize emphasis on localized, problematic speech segments, with the aim of enhancing correct feedback to the learner. The approach to synthesis is based on a perturbation model, where a set of acoustic features relating to energy, fundamental frequency and duration are modulated to present contrast between neutral and emphatic speech segments. This constitutes pedagogical and remedial instructions to facilitate perceptual training for the learner. This should, in turn, assist productive training that is supported by automatic speech recognition technologies. The invited talk will also showcase both the automatic speech recognition and text-to-speech synthesis technologies that are being developed under this research effort (Harrison *et al.*, 2008; Meng *et al.*, 2007; Wang *et al.*, 2008a; Wang *et al.*, 2008b).

References

Asia Economic News, 20 Feb., 2006 http://findarticles.com/p/articles/mi_m0WDP/is_2006_Feb_20/ai_n16086425/

Ehsani, F. and E. Knodt. 1998. Speech technology in computer-aided language learning. *Language Learning and Technology*, vol. 2, no. 1, pp. 45–60.

Harrison, A., W.K. Lo, X.J. Qian and H. Meng. 2009. Implementation of an Extended Recognition Network for Mispronunciation Detection and Diagnosis in Computer-Aided Pronunciation Training. In *Proceedings of the 2nd ISCA Workshop on Speech and Language Technology in Education*, Warrickshire.

Harrison, A., W.Y. Lau, H. Meng and L. Wang. 2008. Improving Mispronunciation Detection and Diagnosis of Learners' Speech with Context-sensitive Phonological Rules based on Language Transfer. In *Proceedings of INTERSPEECH.*

Kim, J.-M., C. Wang, M. Peabody and S. Seneff. 2004. An interactive English pronunciation dictionary for Korean learners. In *Proceedings of INTERSPEECH.*

Lo, W.K., A. Harrison, H. Meng and L. Wang. 2008. Decision Fusion for Improving Mispronunciation Detection using Language Transfer Knowledge and Phoneme-dependent Pronunciation Scoring. In *Proceedings of the 6th International Symposium on Chinese Spoken Language Processing*, Kuming, China.

Meng, H., Y.Y. Lo, L. Wang and W.Y. Lau. 2007. Deriving Salient Learners' Mispronunciations from Cross-Language Phonological Comparisons. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*.

Neri, A., C. Cucchiarini and H. Strik. 2006. ASR-based corrective feedback on pronunciation: does it really work? In *Proceedings of INTERSPEECH*, Pittsburg, USA, pp. 1982–1985.

Precoda, K., C.A. Halverson and H. Franco. 2000. Effects of speech recognition-based pronunciation feedback on second-language pronunciation ability. In *Proceedings of InSTILL*.

Wang, L., X. Feng and H. Meng. 2008a. Automatic Generation and Pruning of Phonetic Mispronunciations to Support Computer-Aided Pronuinciation Training. In *Proceedings of INTERSPEECH.*

Wang, L., X. Feng and H. Meng. 2008b. Mispronunciation Detection Based on Cross-Language Phonological Comparisons. In *Proceedings of the IEEE IET International Conference on Audio, Language and Image Processing*.