Available online at www.sciencedirect.com

SciVerse ScienceDirect

Speech Communication 55 (2013) 691–706

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

On mispronunciation analysis of individual foreign speakers using auditory periphery models

Christos Koniaris *, Giampiero Salvi, Olov Engwall

Centre for Speech Technology, School of Computer Science & Communication, KTH – Royal Institute of Technology, Lindstedtsvägen 24, SE-100 44 Stockholm, Sweden

Received 17 March 2012; received in revised form 2 January 2013; accepted 15 January 2013
Available online 29 January 2013

Abstract

In second language (L2) learning, a major difficulty is to discriminate between the acoustic diversity within an L2 phoneme category and that between different categories. We propose a general method for automatic diagnostic assessment of the pronunciation of non-native speakers based on models of the human auditory periphery. Considering each phoneme class separately, the geometric shape similarity between the native auditory domain and the non-native speech domain is measured. The phonemes that deviate the most from the native pronunciation for a set of L2 speakers are detected by comparing the geometric shape similarity measure with that calculated for native speakers on the same phonemes. To evaluate the system, we have tested it with different non-native speaker groups from various language backgrounds. The experimental results are in accordance with linguistic findings and human listeners' ratings, particularly when both the spectral and temporal cues of the speech signal are utilized in the pronunciation analysis.

Keywords: Second language learning; Auditory model; Distortion measure; Perceptual assessment; Pronunciation error detection; Phoneme

1. Introduction

L2 speakers often have difficulties attaining a native-like pronunciation (Flege, 1995; Guion et al., 2000), especially when the foreign sounds are of a different phonological origin. This is partly the result of a higher-level process[1] by which humans develop the ability to harmonize their hearing (and thereby, their production) system to the sounds of their native language (Werker and Tees, 1984; Kuhl, 1993). Sometimes, L2 learners transfer some speech sounds from their first language (L1), produce the L2 phonemes with inconsistent variations or even discard unfamiliar ones (Piske et al., 2001). In addition, phenomena such as reduced rate in oral communication contributing to an abnormal duration expansion, or articulatory inconvenience, causing unfamiliar expressional elements, are common among L2 speakers, especially during their first contact period with the target language. Several theories and experiments have tried to explain the inability of some L2 speakers to produce a target phoneme correctly. One theory (Diehl and Kluender, 1989) suggests that the learner's auditory perception of the L2 phonemes is not sufficiently accurate and, according to the dispersion principle and the auditory enhancement hypothesis, this is manifested in the learner's production. Additionally, physiological causes such as vocal tract differences in connection with various ethnic backgrounds may lead to voice quality divergences (Andrianopoulos et al., 2001). As a result, native speakers may have difficulties decoding speech sounds produced by L2 speakers (Munro and Derwing,

* Corresponding author. Tel.: +46 8 790 7567; fax: +46 8 790 7854.
E-mail addresses: koniaris@kth.se (C. Koniaris), giampi@kth.se (G. Salvi), engwall@kth.se (O. Engwall).

[1] Central auditory processing is out of the scope of this work, as we only focus on the perception of speech sounds and therefore apply models of auditory periphery. Unless otherwise stated, all the auditory models mentioned in this article concern the periphery.

0167-6393/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2013.01.004

1995; Schmid and Yeni-Komshian, 1999). To overcome these problems, special computer programs can help L2 learners practise, by detecting and analyzing the pronunciation errors. In this article, we propose an automatic diagnostic evaluation of the phonemes that require additional practicing by the L2 speaker. At this stage, we are interested in making an offline L2 pronunciation evaluation by using a precollection of data recordings in which the learner has produced the L2 phonemes several times in different settings. In particular, we want our approach to be *language independent*, *automatic* and *perceptually-motivated*.

In the literature (e.g., Kawai and Hirose, 1998; Moustroufas and Digalakis, 2007), one may find a series of pronunciation evaluation approaches designed for a specific L1 and a certain L2. The obvious advantage of fixing the L1–L2 pair is the possibility to tailor the system with the appropriate information on the differences between the L1 and L2 utilizing theoretical linguistic propositions and experimental data from individual speakers. The drawback is naturally that these systems can only be used for the L1–L2 pair that they were constructed for.

Further studies (Anderson-Hsieh et al., 1992; Neumeyer et al., 1996; Franco et al., 1997) have shown that native speakers' judgements of L2 pronunciation can be influenced by various features of speech, such as intonation, fluency, syllable structure, word stress, etc. This means that the process behind the judgement may be complex and difficult to explain. Consequently, subjective ratings of L2 pronunciation may differ between listeners. More importantly, human listener judgement is not always available, in particular in self-practice. For these reasons, general and objective automatic methods are of interest.

In the last decades, computer-assisted pronunciation training (CAPT) programs have gained popularity among learners wishing to practise their language skills either in the classroom or at home. Some examples of such systems are FLUENCY (Eskenazi and Hansma, 1998), ISLE (Menzel et al., 2000), and EduSpeak® (Franco et al., 2010). Two important features of a CAPT software should be firstly that it provides functional and profitable feedback to improve the learner's performance and secondly that it adapts the training to practice on the difficulties that the learner has. We are focusing on the latter in this article by proposing a way to diagnose the pronunciation weaknesses and provide an ordered list of problematic phonemes.

Classification techniques are common in pronunciation error detection algorithms. In (Witt and Young, 2000) for example, the goodness of pronunciation (GOP) algorithm was presented to calculate the likelihood ratio of a phoneme realization by an L2 speaker to its canonical pronunciation. Alternatively, articulatory information has been used to improve automatic detection of typical phoneme-level errors made by non-native speakers (Tepperman and Narayanan, 2008). For this, a new version of the Hidden-articulator Markov Model (Richardson et al., 2003), adapted for pronunciation evaluation, was presented. Strik et al. (2009) examined four different classifiers to account for mispronunciation detection: a GOP-based, one combining cepstral coefficients and linear discriminant analysis, and two acoustic-phonetic classifiers. Wei et al. (2009) addressed the problem with a support vector machine framework, with pronunciation space models to improve performance.

We believe that models of the human auditory system could be beneficial for detection of mispronunciations. When two native speakers produce the same word, the two speech signals differ, even if both are natively produced. It is the common perceptual characteristics of the speech signal that make the listener classify the utterances from the two speakers as native, despite their acoustic differences. A method that takes some perceptual characteristics into account may therefore be of relevance. In a previous study (Koniaris and Engwall, 2011a), inspired by the hypothesis presented by Diehl and Kluender (1989), we introduced a method to estimate the perceptually relevant characteristics of the native speech in one target language and the perceptually relevant characteristics of the non-native speech in the same target language. We focused on the perceptual affinity between native and non-native speakers, and suggested that the difference could explain why many foreign speakers lack precision in producing L2 phonemes. In a second study (Koniaris and Engwall, 2011b), we proposed a method to be used in diagnosing the pronunciation of L2 learners. Motivated by the ability of the human hearing system to perform relatively well in sound class separation, and the assumption that little information relevant for phoneme distinction is lost in the mapping from the acoustic domain to the perceptual domain, the fundamental principle of our approach is built upon measuring the similarity of the Euclidean geometry of the data (Koniaris et al., 2010b). We only considered the static properties of the speech signal and therefore a spectral auditory model was used. We compared, for each phoneme, the distortion measure between the auditory representation for a group of native speakers and the speech signal's power spectrum for, on the one hand, the same group of native speakers, and, on the other, a group of non-native speakers with the same L1. By comparing the measures for the non-native and the native speakers, we found, quantitatively, the phonemes that deviated the most from the native norm for each group of L2 speakers. In this article, we suggest a second approach, which in addition considers the dynamic aspects of the speech signal. The procedure is similar to that described above, except that (i) the auditory model is of spectro-temporal nature and (ii) the comparison is now with the non-native speakers' acoustic – static and dynamic – feature domain.

This article describes both the static and the dynamic approaches. It is organized as follows: Section 2 explains the underlying idea and basic assumptions of our approach. Section 3 discusses our proposed method to mis-