Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language 50 (2018) 126–140

www.elsevier.com/locate/csl

Computer based speech prosody teaching system☆

Dávid Sztahó*, Gábor Kiss, Klára Vicsi

Budapest University of Technology and Economics, Department of Telecommunication and Media informatics, Laboratory of Speech Acoustics, 1117-H, Budapest, Magyar tudósok körútja 2, Hungary

Received 25 February 2016; received in revised form 28 November 2017; accepted 28 December 2017
Available online 5 January 2018

Abstract

Children who are born with a profound hearing loss have no or only distorted acoustic speech target to imitate and compare their own production with. Computer based visual feedback, visual presentation of speech on screen has shown to be an effective supplement of incomplete or distorted auditory feedback in the case of children with grave hearing-impairment. In this paper, we introduce a novel prosody teaching system where intensity (accent), intonation and rhythm are presented visually for the students (in both separate and combined display mode) as visual feedback and automatic assessment scores are given jointly and separately for the goodness of intonation and rhythm. Evaluation of the automatic assessment was done with cooperation of experts in the field of treatment of hard of hearing children. The results showed that the automatic assessment scores correspond to the subjective evaluations given by the teachers. The evaluation of the whole system was done in a school for hard of hearing children, by comparing the development of a group of students using our prosody teaching system with the development of a control group. The speaking ability of students were compared by a subjective listening experiment after a 3 months teaching course. The students who used the computer based prosody teaching software could produce nicer prosody than the students in the control group.

Keywords: Speech prosody; Intonation; Speech recognition; Speech aid; CAPT

1. Introduction

Children, who are born with a profound hearing loss, do not have or have only distorted acoustic speech target to imitate and compare their own production with. Thus, their speech production is very poor. Computer based visual feedback, visual presentation of speech on screen has shown to be an effective supplement of incomplete or distorted auditory feedback in the case of children with grave hearing-impairment (Osberger et al., 1981; Watson et al., 1989; Yamada and Murata, 1991; Levitt, 1993; Javkin, 1994; Rooney, et al., 1994; Öster, 1996; Vicsi, 2006). It has been shown to be a valuable supplement tool in audio and verbal feedback in speech training for moderately and severely hearing impaired and normally hearing children with speech deviations, as well as for second language (L2) learners. Adult second language learners have difficulties in perceiving the phonetics and prosody of a second language through listening, not because of a hearing loss but because they are not able to hear new sound contrasts due to

☆ This paper has been recommended for acceptance by Prof. R. K. Moore.

* Corresponding author.
E-mail address: sztaho@tmit.bme.hu (D. Sztahó), vicsi@tmit.bme.hu (K. Vicsi).

http://dx.doi.org/10.1016/j.csl.2017.12.010
0885-2308/

interference with their native language. Moreover, experiments have shown that visual display of supra-segmental features combined with audio feedback is more effective than audio feedback alone (James, 1976; De Bot, 1983), especially if the pitch contour of the student is displayed along with a reference model.

Computer-based speech pronunciation teaching systems generally include measuring and displaying the dynamic characteristics of speech parameters, using auditive, visual and automatic feedback. A good review of such systems is given by Levis (2007).

Having established the development of computer based pronunciation-teaching tools, the first step is to determine, which components of pronunciation to address. Roughly speaking, pronunciation quality is defined by its phonetic (segmental) and prosodic features (supra-segmental).

For beginners phonetic characteristics are the most important factors, because the mistakes here cause mispronunciations. Many product and process oriented pronunciation teaching systems have been developed on phonetic level, mostly for L2 learning. Good reviews for L2 learning are given by Donaldson (2009) and Hismanoglu (2011), but only some of them are specialized for hearing impaired children (Crepy, et al., 1983; Youdelman, 1994; Kewley-Port and Watson, 1995; Massaro and Light, 2004). However, the speech processing techniques in a system can be similar for L2 learning and for hearing impaired children, but the didactic construction of a pronunciation teaching system must be different. Auditive feedback is the most important for language learning of students with normal hearing, but for hard of hearing children the basic information must be given by the visual feedback. Of course, an automatic scoring method of the automatic feedback should also be very different. Thus, the individual steps of education differ preferably at the two cases.

Earlier, in the framework of the European SPECO project (Contract no. 977126), a product-oriented multilingual CAPT (Computer Aided Pronunciation Training) system was developed for hearing impaired or normally hearing children with speech deviations, utilizing visually displayed acoustic properties of speech mostly on phonetic level. The visual display is correct from an acoustic-phonetic point of view, but are easy to understand and interesting for children (Vicsi et al., 2000; Vicsi et al., 2001; Vicsi, 2006). The system uses mainly the segmental description of speech in a clear pedagogical basis. The user-friendly tools of this basic system are still being used actively for three languages.

With an already increasing fluency (due to development) of a student's speech, more emphasis should be placed on other aspects, such as teaching prosody: intonation, stress and rhythm. Thus the development of a good prosody teaching system, which uses the latest speech signal processing methods, is very timely. Up to now most of researchers have concentrated only on assessment of the prosodic parameters (Hönig et al., 2012; Bertinetto and Bertini, 2008; Dellwo, 2010) or used a very simple visual feedback to display the intonation contours (Warren et al., 2009). We aim to provide both automatic assessment scores and multiple visual feedback ways that are easy-to-use, but also informative to the users (and the educator staff).

Some years ago due to our promising results in the automatic sentence modality recognition (Vicsi and Szaszák, 2010), we adopted the method for children modality recognition, and looked for the possibility, how it can be used as an automatic feedback in an audio–visual pronunciation teaching and training system. Our goal was to develop a sentence intonation teaching and training system for speech handicapped children, helping them to learn the correct prosodic pronunciation of a sentence. HMM models of modality types were built by training the recognizer with a correctly speaking children spontaneous speech database (Sztahó et al., 2010). During this work, a large database was collected from speech impaired children. Subjective tests were carried out using the speech material, in order to examine, how human listeners are able to categorize the heard recordings of sentence modalities. Automatic sentence modality recognition experiments were done with HMM models trained with different sentence modality types. By the result of the subjective tests, the probability of acceptance of the sentence modality recognizer can be adjusted. In these subjective tests, human listeners had to categorize the heard sentences. The results showed that the automatic recognizer classified the recordings more strictly, but not worse. However, there are variances in the prosody of normal speech that causes incorrect classification in such a system. The variation of the place of stress inside a sentence, which is not necessary considered as incorrect, causes misclassification or lower likelihood of an HMM model, to which the sentence originally belongs to. Thus scoring of a speech sample can be false. Contrary to the language of read speech, in case of the spontaneous language, a wide variety of intonation is acceptable. However, for intonation teaching purposes, in case of children, pronunciation of read speech shall be utilized in a didactical point of view. Therefore, the assessment of a given sentence may not be preferred based on an average intonation model computed from a large spontaneous language dataset. Moreover, the exact locations of the possible pronunciation