Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 37 (2016) 67–81

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Capitalizing on musical rhythm for prosodic training in computer-aided language learning☆

Hao Wang [a], Peggy Mok [b], Helen Meng [a,*]

[a] *Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong, China*

[b] *Department of Linguistics and Modern Languages, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong, China*

Received 4 June 2015; received in revised form 8 October 2015; accepted 21 October 2015
Available online 4 November 2015

Abstract

Language transfer creates a challenge for Chinese (L1) speakers in acquiring English (L2) rhythm. This appears to be a widely encountered difficulty among foreign learners of English, and is a major obstacle in acquiring a near-native oral proficiency. This paper presents a system named MusicSpeak, which strives to capitalize on musical rhythm for prosodic training in second language acquisition. This is one of the first efforts that develop an automatic procedure which can be applied to arbitrary English sentences, to cast rhythmic patterns in speech into rhythmic patterns in music. Learners can practice by speaking in synchrony with the musical rhythm. Evaluation results suggest that after practice, the learners' speech generally achieves higher durational variability and better approximates stress-timed rhythm.

Keywords: Musical rhythm generation; Prosodic training; CALL

1. Introduction

The use of information and communication technologies (ICT) to support computer-aided language learning (CALL) is gaining momentum. Existing work predominantly address phonetic deviances in L2 (second language) speech compared with native speech. Major thrusts lie in applying automatic speech recognition to the learner's speech for automatic scoring and mispronunciation detection. In contrast, there is a paucity of research in developing technologies to support L2 acquisition of suprasegmental phonology. As a suprasegmental feature, rhythm plays a very important role in communication, because it reflects the structure of information in the spoken message (Avery and Ehrlich, 1992). Native-speaking listeners can be frustrated by learners who use incorrect rhythm; and if the stress and rhythm patterns deviate too much from proper native productions, the L2 speakers may not be understood well (Celce-Murcia et al., 1996). Adams (1979), who studied the influence of rhythm on intelligibility, held the same view and found that many learners produce an anomalous rhythm which seriously hampers the total intelligibility of the speech. Pennington

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

* Corresponding author. Tel.: +852 2609 8327; fax: +852 2603 5505.
E-mail addresses: hwang@se.cuhk.edu.hk (H. Wang), peggymok@cuhk.edu.hk (P. Mok), hmmeng@se.cuhk.edu.hk (H. Meng).

http://dx.doi.org/10.1016/j.csl.2015.10.002
0885-2308/

(1994), Yun (2000) etc. also argued that faulty stress and rhythm patterns may cause greater difficulty in intelligibility, compared with inaccurate pronunciations of individual sounds (Gong, 2002). On the other hand, a learner's speech may sound much less foreign when they use the appropriate rhythm and intonation patterns, even though they may have other faults of pronunciations (Rivers and Temperley, 1978).

The acquisition of speech rhythm is essential. However, it is difficult for both teaching and learning. A conventional view of speech rhythm often categorizes languages into syllable-timed and stress-timed; the former has quasi-isochronous duration in syllables, while the latter has quasi-isochronous inter-stress intervals (Abercrombie, 1967). In spite of instrumental studies that shows the lack of systematicity in isochronous units of speech timing across syllable-timed or stress-timed languages (see review in Dauer, 1983), some empirical results demonstrated that syllable-timing and stress-timing may be perceptually distinguishable (Nazzi et al., 1998; Ramus et al., 1999, 2003).

This work focuses on the Chinese (L1) and English (L2) language pair. Chinese is a syllable-timed language while English is usually regarded as stress-timed (Grabe and Low, 2002; Mok and Dellwo, 2008; Mok, 2009). In terms of rhythmic features, negative language transfer easily occurs. Since stress timing is intrinsically more difficult to master (Allen and Hawkins, 1980; Vihman et al., 2006), Chinese learners tend to impose syllable-timed rhythmic pattern on English: typically, giving all syllables relatively regular durations. Stress-timing appears to be the most widely encountered difficulty among (Chinese) learners of English (Chela-Flores, 1993; Faber, 1991; Low et al., 2000; Setter, 2006; Taylor, 1991) and is a major obstacle in acquiring a near-native pronunciation (Adams and Munro, 1978; Gimson, 2001).

In teaching English rhythm, language teachers face the challenges posed by the shortage of teaching materials, difficulties in the design of teaching rhythm, etc. (Gong, 2002). To address those issues, we attempt to leverage commonalities between speech and music. While music may be considered to exhibit a higher structural rigidity than speech, both have melodic, rhythmic and linguistically communicative characteristics. An empirical comparison between speech and music in terms of rhythm has shown some cross-domain similarities, in terms of rhythmic grouping and the statistical patterning of event duration (Patel, 2003). The above motivates us to develop techniques of automatic musical rhythm generation for the purpose of L2 prosodic training which is realized as a system named MusicSpeak. This system follows a procedure that can automatically generate musical rhythm based on arbitrary English text inputs. Users can follow the generated musical rhythms to practice reading the target sentences. We believe that music can enhance learners' engagement in audio-lingual practices. Based on our previous work (Wang et al., 2010), this paper presents improvements by incorporating new rules for rhythm generation which better capture the alternating patterns between stressed and unstressed syllables. We also collect speech data from a larger number of users to evaluate our system. Results suggest that the speech produced when users speak in synchrony with the generated musical rhythm has clearer stress-timed characteristics.

The paper is structured as follows: In Section 2, related previous work involving musical rhythm for English language teaching is stated. Section 3 discusses the similarities between speech and music in terms of rhythmic features. Automatic rhythm generation procedures for generating musical rhythm based on arbitrary English texts are provided and the system interface is shown in Section 4. Section 5 conducts a comparison of collected contrastive recordings between naturally spoken L2 English utterances and their counterparts that are recorded alongside the MusicSpeak rhythm. Finally, conclusions and future work are given in Section 6.

2. Related work

Previous work related to musical rhythm for English prosody training includes:

Graham (1978) created "Jazz Chants", which connects spoken American English to the beat of Jazz. The technique of jazz chanting uses upbeat chants and poems through jazz rhythms to illustrate the natural stress and intonation patterns of conversational American English.

Nakata (2002) developed the KenMc method that connects spoken English rhythm to the beat of Bossa Nova (a style of Brazilian music). "Jazz Chants" and KenMc method are similar. They are both materials of audio recordings, providing English utterances in synchrony with the pieces of matched musical rhythm. The limitation of both materials is that learners can only choose from the pre-set recordings of lessons rather than practice their own sentences.

Fischler (2005) designed and organized a four-week intensive summer pronunciation course. Six L2 English learners from various L1 backgrounds voluntarily participate in the course, in which English word and sentence stress patterns are taught through recitation of rap music and related activities. Comparison of the learners' production before and after