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

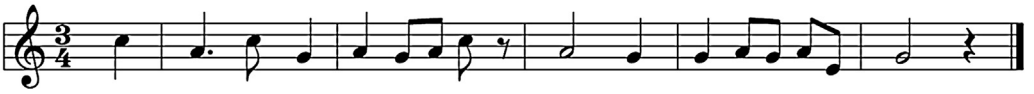

Fig. 1. Example of a piece of music beginning with an incomplete bar.

completion of the course indicates improvement in stress placement. Most of the previous work is applied to previously known text. In this work, we have developed a methodology that can automatically generate musical rhythm for arbitrary texts that learners may wish to follow and practice.

3. Cognition of rhythm

According to the Oxford English Dictionary, rhythm[1] is generally defined as a "movement marked by the regulated succession of strong and weak elements, or of opposite or different conditions". It "is frequently used for any kind of repetition or periodicity in the physical world, and, generally, for practically anything connected with verse experience as long as it is not clearly defined" (de Groot, 1968, cited in Gong, 2002).

3.1. Rhythm in speech

Speech rhythm, although descriptively elusive, is perceptually quite salient. The strong and weak elements in the above dictionary definition generally relate to syllables and stress in speech. According to Lado (1964), rhythm includes stress, time and junctures. Dauer (1983) and Roach (1982) pointed out three important phonological features that differentiate stress-timed and syllable-timed languages: syllable structure, vowel reduction and stress. Stress-timed languages have greater variation in syllable length and structure; more reduced unstressed syllables, more variation in the phonetic realization of stress and more stress-related rules than syllable-timed languages. The coalescence of these phonological differences results in a perceptual distinction between stress-timing and syllable-timing. Wong (1987) also suggested that the rhythm of a language is characterized by the timing pattern of successive syllables (see also Low et al., 2000). That is, it depends on the typical variation of syllable length at the sentence level (Tuan and An, 2010). Recent studies show that other properties like pitch also contribute to the perception of speech rhythm (e.g., Cumming, 2009), but a full understanding of the relevance of these properties to speech rhythm is still lacking. Therefore, in this study we consider stress and syllable length as the keys to rhythm.

A sentence consists of content words and function words. Content words are typically nouns, verbs, adjectives and adverbs. These words are normally carrying high information loads. Function words are articles, conjunctions and pronouns, which have little lexical meaning and mainly serve to express grammatical relationships among words and concepts in a sentence. In English, stress usually falls on content words, and function words are often unstressed. We acknowledge that this is a simplifying assumption and exceptions often arise. Speakers accent content words by uttering the stressed syllables with higher intensity, pitch and duration. On the contrary, unstressed syllables are acoustically reduced (Roach, 2000).

3.2. Rhythm in music and comparison with English rhythm

Alternating stressed and unstressed syllables in English forms the rhythm of the language (Allen, 1972). Musical rhythm is manifested in terms of durations and accents of sounds that produce regular patterns in time, constituting the musical beat (Hawes, 2003). The duration and accent of a musical beat may correspond well with those of an English syllable. Musical rhythm has a rigid structure, where each musical bar is of the same duration and the first beat of each bar is usually accented. We impose this structure onto English rhythm, by forming groups of syllables that begin with a syllable carrying primary stress and optionally followed by one or more unstressed syllables. Each group of syllables constitutes a musical bar. An English sentence may also begin with an "incomplete bar" that does not begin with a stressed syllable (or an accented beat) and has shorter duration than a complete bar. This is illustrated in Fig. 1.

[1] Works cited: "Rhythm", The Compact Edition of the Oxford English Dictionary. II, Oxford University Press, 1971, pp. 2537.

Table 1
Examples of sentences with syllable stress sequences (obtained from dictionary lookup) that do not follow the structure of stress alternation. In the column of sentence, the symbols "\" and "′" indicate a primary stress and a secondary stress, respectively. For the symbolic syllable stress sequences, "1" denotes a syllable carrying a primary stress, "2" represents a syllable with a secondary stress, "0" is an unstressed syllable and "#" indicates the start of a sentence.

Case	Sentence	Syllable stress sequence
1	\big \blue \eye	1 1 1
2	He should have \done it him\self	#0 0 0 1 0 0 1
3	′six\teen \men	2 1 1
4	′Japa\nese \student	2 0 1 1 0

4. Automatic rhythm generation

Based on the analysis in Section 3, we develop the following procedures to generate rhythm for an arbitrary English sentence.

4.1. Stress identification

All polysyllabic words have stressed and unstressed syllables. According to the degree of emphasis, syllable stresses are categorized into three groups: primary stress, secondary stress and no stress. In the rest of the paper, the phonetic transcriptions to be shown are based on the CMU Pronunciation Dictionary (Weide, 2008) where the pronunciations are encoded using a modified form of the ARPAbet system.[2] An example is given as follows: the word "pronunciation/P R OW0 N AH2 N S IY0 EY1 SH AH0 N/" as a total of five syllables where the syllable *a* carries primary stress that is stronger than the secondary stress carried by the syllable *nun*.

To identify the level of stress for each syllable in a sentence, first of all, we get the phonetic transcription (with stress information) of the sentence by means of dictionary lookup, based on the CMU Pronunciation Dictionary. By referring to a function word list, content words and function words of the sentence can be identified. According to Section 3.1, all stressed syllables in content words are accented, while other syllables including unstressed syllables in content words and all syllables in function words are treated as unstressed. Then, a syllable stress sequence in English follows the principle of rhythmic alternation; that is, weak and strong syllables alternate with one another (Sabater, 1991; Gimson, 2001). There exists some syllable stress sequences obtained from the dictionary having structures which do not observe the stress alternation principle (e.g., too many primary stresses coming next to each other, or a long sequence of unstressed syllables). Examples are illustrated in Table 1.

We modify such structures using several rhythm rules (Sabater, 1991) such that the generated rhythm can better exhibit rhythmic alternation.

(1) *Stress deletion*: If a series of content words appear next to each other, some stresses are dropped, e.g., the intermediate stress tends to be dropped for a succession of three primarily stressed syllables in order to achieve a more regular alternation.
(2) *Stress addition*: If an utterance consists of a succession of unstressed function words, stresses are added to produce a more regular rhythm.
(3) *Stress shift*: If two stresses are next to each other in a phrase with no intervening beats, there is a "stress clash", and stress shifts toward the preceding strong syllable.

Examples are given in Table 2 to illustrate the use of the above rhythm rules for achieving the structure of stress alternation.

[2] ARPAbet is a phonetic transcription code, representing each phoneme of General American English with a distinct sequence of ASCII characters. More details about ARPAbet can be found in Wikipedia: http://en.wikipedia.org/wiki/Arpabet. The difference between the form used in CMU Pronunciation Dictionary and that in ARPAbet system is stress marks on vowels with level 0 (no stress), 1 (primary stress) and 2 (secondary stress).

Table 2
Examples of modification of the structures of sentences given in Table 1 using the above rhythm rules.

Case	Rhythm rule	Symbolic expression	Sentence after modification	Final stress sequence
1	Stress deletion	1 1 1 → 1 0 1	\big blue \eye	1 0 1
2	Stress addition	# 0 0 0 → # 1 0 0	\He should have \done it him\self	# 1 0 0 1 0 0 1
3	Stress shift	2 1 1 → 1 0 1	\sixteen \men	1 0 1
4	Stress shift	2 0 1 1 → 1 0 0 1	\Japanese \student	1 0 0 1 0

4.2. Musical bar placement

Having the final syllable stress sequence of the target English sentence obtained in Section 4.1, we proceed to generate the musical rhythm. First of all, we let every syllable occupy a beat. Then, we organize the syllables into musical bars by following the heuristic that syllables carrying primary stress are always placed as the first beat. In musical notation, there are two types of musical bars – if a bar begins with an accented beat, it is regarded as a "complete" bar; otherwise, it is "incomplete". We draw analogies with English rhythm, relating specifically to whether the sentence starts with a primarily stressed syllable. Therefore, musical bars in a piece of English rhythm can also be categorized into two types. If we consider further about the detailed structure in a musical bar, we sum up a total of five basic cases for musical bar placement, which are listed as follows:

Incomplete bars have two different structures that are illustrated as the following two cases:

Case 1: <u> |
where <u> denotes an arbitrary number of unstressed syllables and '|' is the boundary of a (incomplete) musical bar.

Case 2: <u> 2 <u> |
where "2" denotes a syllable carrying secondary stress. This case shows that an incomplete bar can also include a secondarily stressed syllable.

A complete bar begins with a primarily stressed syllable. The next three cases represent different structures in complete bars.

Case 3: | 1 <u> |
where "1" denotes a syllable with primary stress. It is the case that a primarily stressed syllable followed by any number of unstressed syllables form a complete bar.

Case 4: | 1 <u> 2 <u> |
This is a complete bar formed by one primary stressed syllable, one secondarily stressed syllable and an arbitrary number of unstressed syllables.

Case 5: | 1 <u> 2 <u> 2 <u> |
This case illustrates a complete bar formed by one primary stressed syllable, two secondarily stressed syllables and an arbitrary number of unstressed syllables.

4.3. Duration assignment

Based on the cases in the previous section, the rhythm can be generated by assigning appropriate durations to syllables carrying different levels of stress in each musical bar. The assigned durations are calculated based on the following heuristics: Syllables carrying primary stress consume the longest durations and unstressed syllables consume the shortest ones. The syllables carrying the same level of stress in a musical bar are assigned the same duration. Should a musical bar contain only two syllables with different levels of stress, we impose a duration ratio of 3:2 between the two syllables.

4.3.1. Notations

Before we calculate the duration of each syllable in every musical bar, we first define some notations. We use D_p, D_s, D_u, D_b and D_i to represent the durations of a primary stressed syllable, a secondary stressed syllable, an unstressed syllable, a complete bar and an incomplete bar, respectively. We roughly consider the average rate of speech to be

five syllables per second (Kendall, 2009), i.e., 0.2 s per syllable; thus, for an English sentence, we use the following equations to assign the default values of D_b, and D_i, respectively:

$$D_b = D_a \cdot N_b, \tag{1}$$

where D_a is the average syllable duration that is 0.2 s; N_b denotes the number of syllables in the musical bar which contains the most syllables.

$$D_i = D_a \cdot N_i, \tag{2}$$

where N_i denotes the number of syllables in the incomplete bar.

We also define N_p, N_s and N_u as the numbers of primarily stressed, secondarily stressed and unstressed syllables in a musical bar, respectively. All the above notations are used in the subsequent calculations.

4.3.2. Calculations

We present the calculation procedures for each case of musical bar structure in Section 4.2, as follows:

Case 1: <u> |

The duration of an incomplete bar is distributed equally across the number of unstressed syllables, as shown in Eq. (3):

$$D_u = \frac{D_i}{N_u}, \quad \text{if } N_p = 0,\ N_s = 0 \text{ and } N_u \geq 1. \tag{3}$$

Case 2: <u> 2 <u> |

In this case, the incomplete bar can contain any number of unstressed syllables including 0. If the number of unstressed syllable is 0, then the secondarily stressed syllable consumes the entire duration of this incomplete bar; the corresponding calculation is shown in Eq. (4a):

$$D_s = D_i, \quad \text{if } N_p = 0,\ N_s = 1 \text{ and } N_u = 0. \tag{4a}$$

If the musical bar contains one secondarily stressed syllable and one unstressed syllable, we impose a duration ratio 3:2, as shown in Eq. (4b):

$$D_s = \frac{3}{5} \cdot D_i, \quad D_u = \frac{2}{5} \cdot D_i, \qquad \text{if } N_p = 0,\ N_s = 1 \text{ and } N_u = 1. \tag{4b}$$

Equations in (4c) handle a general scenario where proportionate duration is assigned to the secondarily stressed syllable and the remaining duration is distributed equally among the unstressed syllables.

$$D_s = \frac{D_i}{N_u}, \quad D_u = \frac{D_i - D_s}{N_u}, \qquad \text{if } N_p = 0,\ N_s = 1 \text{ and } N_u > 1. \tag{4c}$$

Case 3: | 1 <u> |

This case involves two levels of stress as in Case 2 with the main difference that the primarily stressed syllable must occupy the first beat in the musical bar. Equations in (5a)–(5c) are similar to equations in (4a)–(4c) in their rationale.

$$D_p = D_b, \quad \text{if } N_p = 1,\ N_s = 0 \text{ and } N_u = 0. \tag{5a}$$

$$D_p = \frac{3}{5} \cdot D_b, \quad D_u = \frac{2}{5} \cdot D_b, \qquad \text{if } N_p = 1,\ N_s = 0 \text{ and } N_u = 1. \tag{5b}$$

$$D_p = \frac{D_b}{N_u}, \quad D_u = \frac{D_b - D_p}{N_u}, \qquad \text{if } N_p = 1,\ N_s = 0 \text{ and } N_u > 1. \tag{5c}$$

Case 4: | 1 <u> 2 <u> |

In this case, when the number of unstressed syllables is 0, there are only two syllables carrying different levels of stress contained in the musical bar. We impose a duration ratio of 3:2 as shown in Eq. (6a):

$$D_p = \frac{3}{5} \cdot D_b, \quad D_s = \frac{2}{5} \cdot D_b, \qquad \text{if } N_p = 1,\ N_s = 1 \text{ and } N_u = 0. \tag{6a}$$

Table 3
Transcriptions with stress information obtained from CMU Pronunciation Dictionary for all words in the sentence "She opened a big blue bag and a box of tools". In the transcriptions, "1" indicates primary stress and "0" denotes no stress.

Word	Phonetic transcriptions with stress information
She	/ SH IY1 /
opened	/ OW1 P AH0 N D /
a	/ AH0 /
big	/ B IH1 G /
blue	/ B L UW1 /
bag	/ B AE1 G /
and	/ AH0 N D /
a	/ AH0 /
box	/ B AA1 K S /
of	/ AH1 V /
tools	/ T UW1 L Z /

A general scenario is handled by equations in (6b), where proportionate durations are assigned to the primarily and secondarily stressed syllables; and the remaining duration is distributed equally among the unstressed syllables.

$$D_p = \frac{D_b}{N_s + N_u},\ D_s = \frac{D_b}{N_s + N_u + 1},\ D_u = \frac{D_b - (D_p + N_s \cdot D_s)}{N_u}, \quad \text{if } N_p = 1,\ N_s = 1 \text{ and } N_u > 0. \tag{6b}$$

Case 5: | 1 <u> 2 <u> 2 <u> |

The musical bar contains one primarily stressed syllable, two secondarily stressed syllables and no unstressed syllables. Thus, the syllable carrying primary stress is assigned a proportionate duration; the remaining duration is shared equally between the two secondarily stressed syllables as shown in Eq. (7a):

$$D_p = \frac{D_b}{N_s},\ D_u = \frac{D_b - D_p}{N_s}, \quad \text{if } N_p = 1,\ N_s = 2 \text{ and } N_u = 0. \tag{7a}$$

When the musical bar includes unstressed syllables, a proportionate duration is assigned to the syllable carrying primary stress; a shorter proportionate duration is assigned to each of the two secondarily stressed syllables; the remaining duration is distributed equally among the unstressed syllables. The calculations are laid out in Eq. (7b):

$$D_p = \frac{D_b}{N_s + N_u},\ D_s = \frac{D_b}{N_s + N_u + 1},\ D_u = \frac{D_b - (D_p + N_s \cdot D_s)}{N_u}, \quad \text{if } N_p = 1,\ N_s = 2 \text{ and } N_u > 0. \tag{7b}$$

4.4. Beat strength assignment

We aim to develop a system that can generate an audio output. To make the generated rhythm expressive and instructive, different beat strengths are assigned to the syllables according to the levels of stress. Syllables carrying primary stress get the strongest beats and unstressed syllables get the weakest beats.

4.5. Example

This section presents an illustrative example of the end-to-end automatic rhythm generation procedure, based on the input sentence "She opened a big blue bag and a box of tools." A step-by-step walkthrough is as follows:

First of all, we refer to the function word list to identify the function words and content words in the sentence. By looking up the CMU Pronunciation Dictionary (Weide, 2008), we obtain the phonetic transcriptions of all the words, together with information about the stressed vowels (see Table 3).

Table 4
An example illustrating the musical bar placement process: the first row is the target sentence with the content words boldfaced (the word "blue" is italic since its stress is dropped by applying the stress deletion rule in Section 4.1); the second row is the final sequence of syllable stress; and the third row illustrates the result of musical bar placement.

Input Sentence	She **opened** a **big** ***blue*** **bag** and a **box** of **tools**.
Syllable Stress Sequence	0 1 0 0 1 0 1 0 0 1 0 1
Generated Musical Bars	0 | 1 0 0 1 | 1 0 | 1 0 0 | 1 0 | 1 |

We apply the three rules for rhythm generation as described in Section 4.1. In this example, we find that there exists a series of content words next to each other in the sentence –"big blue bag". Based on the above transcriptions, we drop the stress of the word "blue" by applying the stress deletion rule according to Section 4.1. Then, we obtain a final syllable stress sequence that conforms to the structure of the rhythmic alternation. After that, we use the technique described in Section 4.2 to organize all the syllables into musical bars. Table 4 illustrates the brief process of musical bar placement.

Based on the result of the musical bar placement, we identify the numbers of syllables in the incomplete bar and the musical bar containing the most syllables in this sentence; the numbers are 1 and 3, respectively. Thus, according to Eqs. (1) and (2) in Section 4.3.1, the default values of D_b, and D_i are assigned as follows:

$$D_b = D_a \cdot N_b = 0.2 \times 3 = 0.6\,(\mathrm{s})$$

$$D_i = D_a \cdot N_i = 0.2 \times 1 = 0.2\,(\mathrm{s})$$

Finally, we calculate the duration of each syllable in every musical bar of the sentence using the corresponding equations presented in Section 4.3.2, e.g., the second musical bar has one primarily stressed syllable followed by two unstressed syllables. Equations in (5c) are applied, i.e.:

$$D_p = \frac{D_b}{N_u} = \frac{D_b}{2} = 0.3\,(\mathrm{s})$$

$$D_u = \frac{D_b - D_p}{N_u} = \frac{D_b - D_b/2}{2} = \frac{D_b}{4} = 0.15\,(\mathrm{s})$$

4.6. System interface

We build a system named "MusicSpeak" to implement the automatic rhythm generation mechanism. The system is implemented in Java. Fig. 2 shows the screenshots of the MusicSpeak user interface. Users can input an arbitrary English sentence, and the system generates a musical rhythm according to the input text and displays the output on "Result Panel" tab (see Fig. 2). The user can then click the *Play Now* button to listen to the generated rhythm, while the corresponding words are highlighted with the beat in a time-synchronous manner. The user interface also color codes content words differently from function words (the former in red and the latter in green). The area on the upper left side of the result panel has three sliders that are designed to control the beat strength of each level of stress; a higher value means a higher beat strength (volume) of the corresponding syllable. In the current version of the system, we fix the pitch of each level of stress according to the principle that the pitch of primary stress is highest, that of no stress is lowest and that of secondary stress is in between. On the upper right side of the panel are two sliders that are used for adjusting the durations (the values are in milliseconds) of the incomplete and complete bars. The lower part of the panel displays different kinds of instruments. Users can choose among them to change the sound effect of the generated rhythm. The default sound effect is the drum (see circled part in Fig. 2).

5. Evaluation

We invite several Chinese learners of English to try out our system and collect their speech recordings. We also recruit a few native American English speakers to record them reading the English sentences.

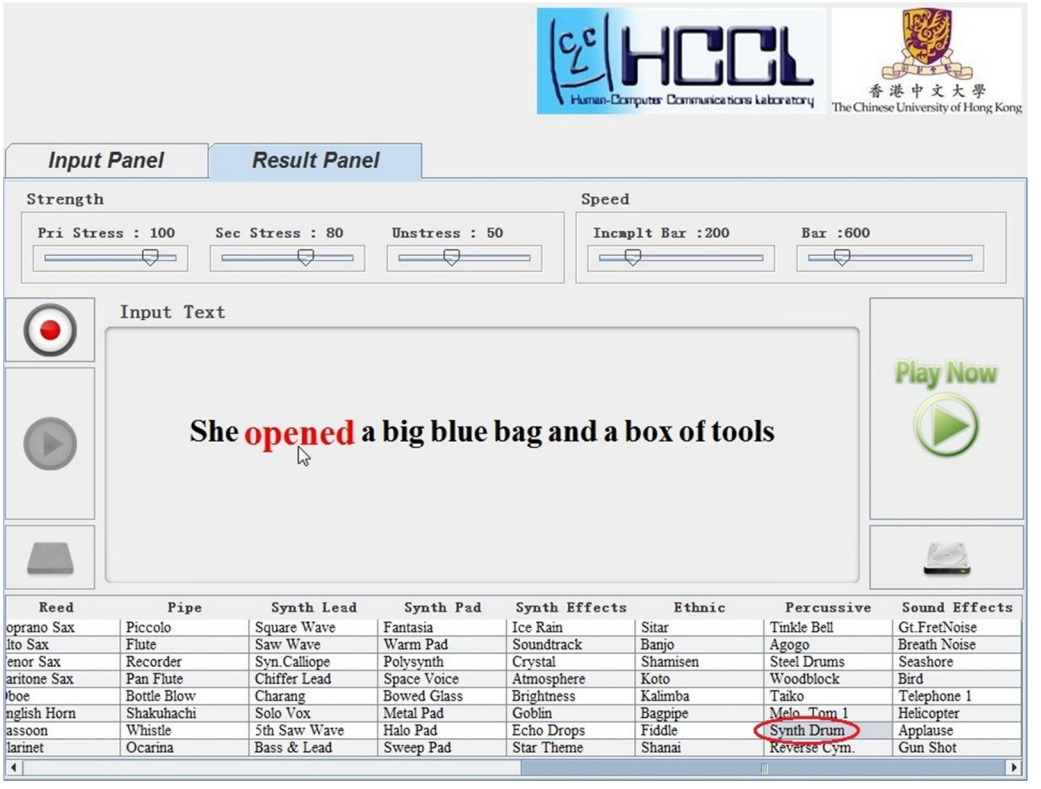

Fig. 2. Illustration of the MusicSpeak interface.

Table 5
Examples of text prompts used in recording.

Sticks and stones are never gonna shake me.
She wants you to be a part of the future.
She opened a book and a box of tools.
It's the end of the world in my mind.
All the stars are coming out tonight.
etc.

5.1. Corpus

In order to investigate the effectiveness of the model, we randomly sample 15 English sentences from song lyrics. The number of words per sentence range from 7 to 12. Examples are shown in Table 5. We invite 20 subjects (9 female and 11 male Chinese learners of English) to record each sentence in two speaking styles – first in their natural way of speaking English (natural style) and then speaking alongside the generated rhythm from MusicSpeak (rhythmic style). All our volunteers are university students (both undergraduates and postgraduates). For each sentence, the subjects are asked to record in natural style before they are given the corresponding generated rhythm to record in rhythmic style. Each subject is allowed to practice reading the sentences in each style as many times as they like before the actual recording. Six native American English speakers (4 females and 2 males) are also recruited to record the same 15 English sentences in their natural way of speaking; these data are used as references. We collect 600 non-native English utterances (15 sentences $\times$ 20 non-native speakers $\times$ 2 styles) and 90 native English utterances (15 sentences $\times$ 6 native speakers). Each recording is digitized at 16 kHz sampling rate and stored at 16 bits per sample, mono, in .wav format.

5.2. Data analysis

We obtain phonetic boundaries for all recordings by means of forced alignment with an automatic speech recognizer (Meng et al., 2007). The phone segmentations are then mapped automatically into consonantal and vocalic intervals and thereafter syllabic intervals. Segmentation criteria follow those in Grabe and Low (2002). Phonotactic constraints and the maximal onset principle are used in deciding syllable boundaries (Deterding, 2001). Durations (ms) of syllabic, consonantal and vocalic intervals are extracted. Any silent pause within an utterance is excluded from further analysis.

5.2.1. The pairwise variability index

The pairwise variability index (PVI) (Grabe and Low, 2002) is used to compare the rhythmic difference between the non-native utterances in natural style (UTT_n) and the utterances in rhythmic style (UTT_r). It is noted that the use of rhythmic metrics to quantify speech rhythm is under much criticism recently (e.g., Arvaniti, 2009), but it is justified in our case because we compare the same sentences spoken by the same speakers. The PVI expresses the level of durational variability in successive intervals. There are two versions of the PVI, raw (see Eq. (8)) and normalized (see Eq. (9)):

$$rPVI = \sum_{k=1}^{m-1} \left[\left| \frac{d_k - d_{k+1}}{(m-1)} \right| \right] \tag{8}$$

$$nPVI = 100 \times \sum_{k=1}^{m-1} \left[\left| \frac{d_k - d_{k+1}}{(d_k + d_{k+1})/2} \right| /(m-1) \right] \tag{9}$$

where m = number of units; d = duration of the kth interval.

Raw PVI takes the absolute difference in duration between each pair of successive units, while normalized PVI uses the mean duration of each pair of successive units to normalize for speech rate variations. Since the utterances in our corpus are produced by multiple speakers having different speech rates, raw PVI is of less interest than normalized PVI. Since normalization for speech rate may also eliminate differences on consonant intervals due to syllable structure (Grabe and Low, 2002), normalized PVI is only calculated for vocalic (nPVIV) and syllabic (nPVIS) durations.

The higher the PVI value, the greater the durational variability exhibited which is a characteristic of stress-timing. For each speaker, we calculate the PVI measures for each of his/her utterances and then obtain the average measurement for the speaker. It is expected that speakers following the generated rhythm will exhibit a higher durational variability than when they just spoke naturally.

5.2.2. Perceptual evaluation

Higher PVI values alone do not always indicate a better approximation of stress-timed rhythm of English. In order to better evaluate the usefulness of our system, we also design a perceptual evaluation questionnaire to collect subjective ratings of the non-native utterances as perceived by native English speakers. In fact, we find that among UTT_n, several sound good in terms of rhythm. Since our system is designed to help learners practice the production of good English rhythm, it may be less useful to those who can already articulate well in terms of speech rhythm. A phonetician (the second author) divides the 20 non-native speakers into two groups (Group 1: 11 speakers, Group 2: 9 speakers) according to the well-formedness of the rhythm of their English speech in natural style. To restrict the length of the questionnaire, we randomly selected five speakers from each group, and randomly pick 4 out of 15 pairs of UTT_n and UTT_r for each selected speaker. The sampled utterances cover 14 out of 15 text prompts, which can be deemed free of biasedness. We also randomly pick five utterances produced by native English speakers as references to check the quality of the subjective evaluation data. All the selected utterances are included in the questionnaire in random order.

We recruit ten native American/Canadian English speakers with backgrounds in linguistics as the subjects of this evaluation task. The raters are asked to assess each utterance on a 7-point scale in terms of the nativeness of the speech rhythm. A higher rating means the rhythm is more native-like. It is expected that, for the speaker group that can produce good rhythm in their natural English speech (Group A, a subset of Group 1), the ratings of UTT_n are not lower than the ratings of the corresponding UTT_r. On the other hand, for the speaker group that speakers did not perform well on English speech rhythm in natural style (Group B, a subset of Group 2), the ratings of UTT_n are lower than the ratings of the corresponding UTT_r.

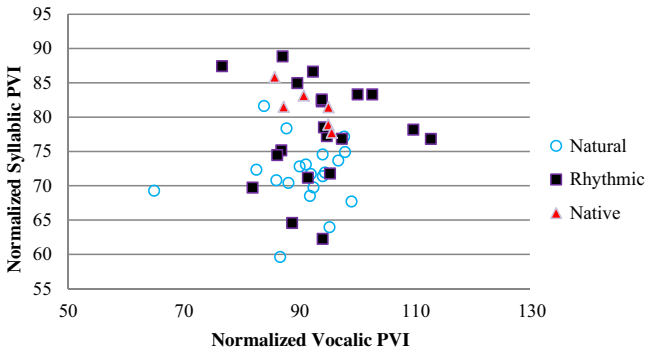

Fig. 3. Normalized syllabic PVI vs. normalized vocalic PVI.

5.3. Results and discussion

This section presents and discusses the results of both objective and subjective tests conducted to evaluate the effectiveness of our system.

5.3.1. Objective test

Fig. 3 shows the average PVI values (across 15 sentences) of 20 individual non-native speakers for each of the two styles (i.e., natural versus rhythmic) and also the average PVI values of 6 individual native speakers. We divide these average PVI values into groups of "Natural", "Rhythmic" and "Native" and conduct student t-tests to see if the two speaking styles of the non-native speakers are separated and how close each of them is to the native speakers in terms of their PVI values. The null hypothesis we test is that the populations from which the two testing groups of PVI values are taken are the same. Paired t-tests are conducted for both normalized syllabic and vocalic PVIs between groups "Natural" and "Rhythmic", while independent t-tests are conducted for normalized syllabic and vocalic PVIs between groups "Native" and "Natural" and between groups "Native" and "Rhythmic", respectively. We fix the level of significance at 0.01, i.e., the null hypothesis is rejected if the p-value is smaller than 0.01.

Paired t-tests confirm that $\mathrm{UTT_r}$ have higher values than $\mathrm{UTT_n}$ for normalized syllabic PVI [$t(19) = 3.5692, p \leq 0.01$]. The results of independent t-tests [$t(24) = 4.6659$, $p \leq 0.01$] for groups of "Native" and "Natural" and [$t(24) = 1.1699$, $p = 0.1267$] for groups of "Native" and "Rhythmic" suggest that the $\mathrm{UTT_r}$ are closer to native utterances than $\mathrm{UTT_n}$ in terms of variability in syllabic durations. These results confirm that speakers do have more variable speech timing when they follow the generated rhythm.

However, the measures for normalized vocalic PVIs do not show statistically significant differences between the two speaking styles [$t(19) = 1.6016, p = 0.0629$]. Independent t-tests between native and non-native utterances also give similar results: [$t(24) = 0.3757$, $p = 0.3552$] for the groups of "Native" and "Natural" and [$t(24) = 0.5255$, $p = 0.3020$] for the groups of "Native" and "Rhythmic". This is probably because speakers tend to produce more pauses in $\mathrm{UTT_r}$ than $\mathrm{UTT_n}$. Sometimes, as only one or two long beats occupy a musical bar, speakers naturally slow down and lengthen the target syllables in order to follow the generated rhythms closely. This results in much vowel lengthening for these syllables, which reduce the durational variability between vocalic intervals.

5.3.2. Subjective test

The subjective ratings can clearly indicate whether our system can help learners improve the rhythm of their English speech. A higher rating on a 7-point scale means the rhythm of the utterance is more native-like. All the raters give a rating of 6 or 7 to all the native English utterances in the questionnaire. Thus, we can regard the ratings from all the 10 raters to be reliable. For each pair of the $\mathrm{UTT_n}$ and $\mathrm{UTT_r}$, there are ten pairs of the ratings. As mentioned in Section 5.2.2, the non-native speakers are split into two groups according to a phonetician's perception on the nativeness of the rhythm of $\mathrm{UTT_n}$ (the rhythm of $\mathrm{UTT_n}$ in Group A is better than that in Group B). If we take a closer look at the pairs of ratings, we can find that the two groups quite distinguishable (see Figs. 4 and 5).

Fig. 4a and b illustrates the distribution of pairs of non-native utterances according to the comparison of average ratings and the preferences of raters in each pair for Group A. $\mathrm{UTT_n}$ in more than 17 pairs (out of a total of 20 pairs;

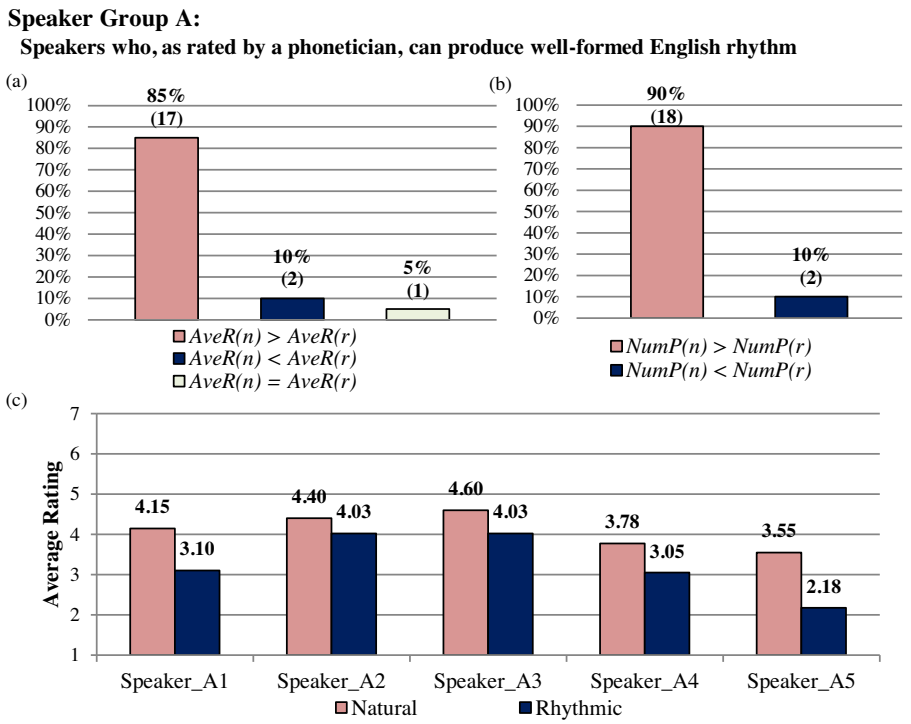

Fig. 4. (a, upper left) Distribution of pairs of non-native utterances according to the comparison of average ratings in each pair for Speaker Group A. (b, upper right) Distribution of pairs of non-native utterances according to raters' preferences of styles in each pair for Speaker Group A. *AveR(n)* denotes the average rating (across the ratings from the 10 raters) of a UTT_n. *AveR(r)* denotes the average rating (from the seven raters) of a UTT_r. *NumP(n)* denotes the number of raters who give higher ratings to UTT_n in the pair. *NumP(r)* denotes the number of raters who give higher ratings to UTT_r in the pair. (c, lower) Average per speaker ratings of UTT_n and UTT_r for Group A.

85%) from Group A have higher average ratings and are given higher preferences by the raters. Fig. 4c shows the average ratings of UTT_n and UTT_r per speaker in Group A.

All the figures indicate that our system is not useful for these speakers who originally produce appropriate English rhythm. On the contrary, perhaps due to its structural rigidity, following the generated musical rhythm can reduce the naturalness of the speakers' speech and make their UTT_r sound worse. This interpretation is also supported by a one-tailed paired *t*-test on the average ratings of UTT_n and UTT_r [$t(17) = 8.2121$, $p \leq 0.01$] for the (18) pairs (refer to "*NumP(n) > NumP(r)*" in Fig. 4b) in each of which UTT_n get more raters' preferences. We do not conduct a *t*-test for the pairs in each of which more raters prefer UTT_r to UTT_n because the sample size of 2 (refer to "*NumP(n) < NumP(r)*" in Fig. 4b) is too small to allow a reliable calculation of the *t* statistic.

From Fig. 5a and b we can observe that compared with Group A, a larger number of UTT_r in the pairs from Group B have higher average ratings and higher preferences given by the raters than the corresponding UTT_n. Fig. 5c shows the average ratings of UTT_n and UTT_r per speaker in Group B.

The observations meet our expectations well except for Speaker B1 (see Fig. 5c). We also conduct the following one-tailed paired *t*-tests on the average ratings of UTT_n and UTT_r for verification. The statistically significant difference [$t(11) = 4.2378$, $p \leq 0.01$] is found for the (12) pairs in each of which UTT_r get at least the same number of raters' preferences as UTT_n (refer to "*NumP(n) < NumP(r)*" and "*NumP(n) = NumP(r)*" in Fig. 5b). It confirms that there are indeed improvements. For the (8) pairs in each of which more raters prefer UTT_n to UTT_r (refer to "*NumP(n) > NumP(r)*" in Fig. 5b), the difference is not statistically significant at the 0.01 level [$t(7) = 2.4052$, $p = 0.1418$]. It shows that although UTT_n in some of the pairs get more raters' preferences than the corresponding UTT_r, the average ratings of UTT_n and UTT_r in each of those pairs do not seem to be very much different. All the above results suggest that for the speakers who do not perform well on rhythm when they speak English naturally, our system can help them improve the nativeness of their English speech rhythm.

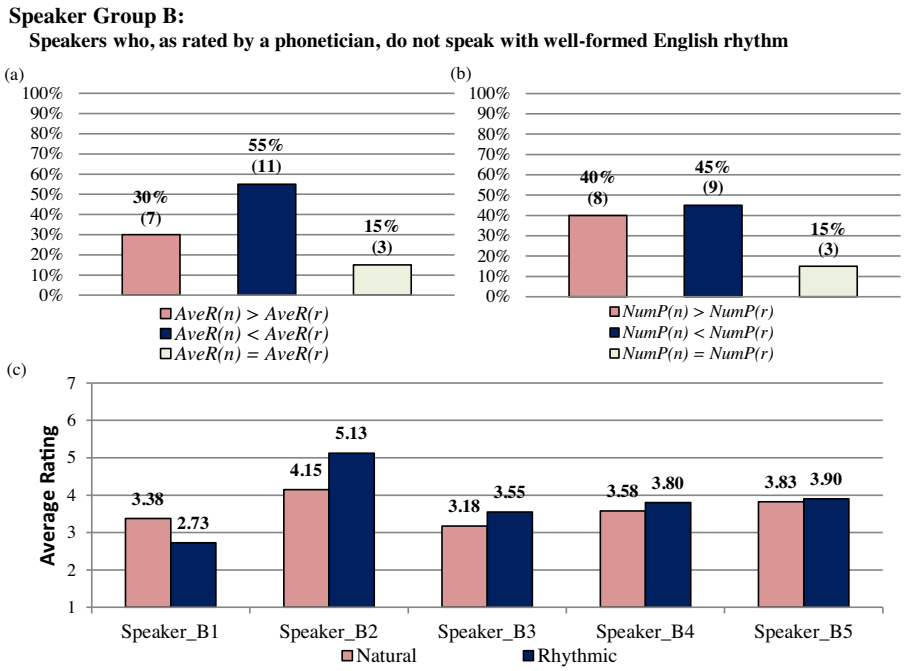

Fig. 5. (a, upper left) Distribution of pairs of non-native utterances according to the comparison of average ratings in each pair for Speaker Group B. (b, upper right). Distribution of pairs of non-native utterances according to raters' preferences of styles in each pair for Speaker Group B. (c, lower) Average per speaker ratings of UTT_n and UTT_r for Group B.

The average rating across all (20) UTT_n from Group A is 4.1 (on a 7-point scale) while that from Group B is 3.6. The result of an independent t-test shows that the difference between the two groups is significant at the 0.05 level [$t(38) = 1.6466$, $p = 0.05$]. It is in reasonable agreement with our grouping criterion. However, even for Group B, an average rating of 3.6 does not seem to be too bad. It is worth pointing out that all the non-native speakers involved in this experiment are at the undergraduate level of education or above; they have been learning English for more than 10 years. In addition, we only have 10 native English speakers for perceptual rating. Still, from our results we can infer that, for the learners who do not perform well in producing English rhythm, especially for beginner-level Chinese learners of English, our system would be particularly helpful to improve the nativeness of the rhythm of their English speech.

6. Conclusions and future work

This paper presents the MusicSpeak system, which incorporates an automatic procedure that casts rhythmic patterns in speech (based on alternating stressed and unstressed syllables) into rhythmic patterns in music (based on musical bars and beats). This procedure can be applied to arbitrary English sentence inputs, where rhythmic generation considers the discrimination between content and function words in the sentence, as well as the locations of stressed syllables. We collect speech recordings from 20 non-native speakers uttering 15 English sentences each, first in natural style and then in synchrony with the generated musical rhythm. We also record speech made by native American English speakers as reference. Comparison between the two styles of speech based on rhythm metrics suggests that the latter style has higher variability in rhythm, which may better approximate stress-timed rhythm in English. Subjective evaluations for the selected pairs of non-native utterances (in the two styles) also demonstrate that the use of musical rhythm in suprasegmental training for second language acquisition is a promising approach.

In the future, we will extend our work beyond the syllable level, e.g., to the phrase level, to seek a more precise duration assignment of a beat. We will also develop a data-driven approach to learn synthesis rules that can generate musical rhythm that is more natural and native. Furthermore, employing the generated rhythm in "text-to-rhythmic-speech" synthesis is a very interesting and useful research direction since synthesized rhythmic speech would be a

better reference than just musical beats for learners to follow. We also plan to conduct evaluations based on perception by a large population to assess the effectiveness of our system.

Acknowledgement

The work is partially supported by the grant from the Hong Kong SAR Government's Research Grants Council General Research Fund (Project No. 415511).

References

Abercrombie, D., 1967. Elements of General Phonetics. Edinburgh University Press, Edinburgh.
Adams, C., 1979. English Speech Rhythm and the Foreign Learner. Mouton Publishers, The Hague.
Adams, C., Munro, R., 1978. In search of the acoustic correlates of stress: fundamental frequency, amplitude and duration in the connected utterance of some native and non-native speakers of English. Phonetica 35, 125–156.
Allen, G.D., 1972. The location of rhythmic stress beats in English: an experimental study I. Langu. Speech 15, 72–100.
Allen, G., Hawkins, S., 1980. Phonological rhythm: definition and development. In: Yeni-Komshian, G., Kavanagh, J., Ferguson, C. (Eds.), Child Phonology: Volume 1 Production. Academic Press, New York, pp. 227–256.
Arvaniti, A., 2009. Rhythm, timing and the timing of rhythm. Phonetica 66, 46–63.
Avery, P., Ehrlich, S., 1992. Teaching American English: Oxford Handbook for Language Teachers. Oxford University Press, Hong Kong.
Celce-Murcia, M., Briton, D.M., Goodwin, J.M., 1996. Teaching Pronunciation, A Reference for Teachers of English to Speakers of Other Languages. Cambridge University Press, pp. 131–174.
Chela-Flores, B., 1993. On the acquisition of English rhythm: theoretical and practical issues. Leng. Mod. 20, 151–164.
Cumming, R., 2009. The interdependence of tonal and durational cues in the perception of rhythmic groups. Phonetica 67, 219–242.
Dauer, R.M., 1983. Stress-timing and syllable-timing reanalyzed. J. Phon. 11, 51–62.
Deterding, D., 2001. The measurement of rhythm: a comparison of Singapore and British English. J. Phon. 29, 217–230.
Faber, D., 1991. In: Brown, A. (Ed.), Teaching the rhythms of English: a new theoretical base. , pp. 245–258, First Published in 1986 in International Review of Applied Linguistics 24, 205–216.
Fischler, J., (Thesis) 2005. The rap on stress: teaching stress patterns to English language learners through rap music. Hamline University, Saint Paul, Minesota.
Gimson, A.C., 2001. Gimson's Pronunciation of English (6th edition revised by Cruttenden, A.). Oxford University Press, London.
Gong, J., 2002. Introducing English rhythm in Chinese EFL classroom: a literature review. Post-Script 3 (1), 26–42.
Grabe, E., Low, E.L., 2002. Durational variability in speech and the rhythm class hypothesis. In: Gussenhoven, C., Warner, N. (Eds.), Laboratory Phonology VII. Mouton de Gruyter, Berlin, pp. 515–546.
Graham, C., 1978. Jazz Chants, 1st ed. Oxford University Press, New York.
Hawes, N.V., 2003. Basic Music Theory, Online, URL: http://neilhawes.com/sstheory/theory16.htm (accessed 19.10.13).
Kendall, T.S., (Doctoral dissertation) 2009. Speech rate, pause and linguistic variation: an examination through the sociolinguistic archive and analysis project.
Lado, R., 1964. Language Teaching: A Scientific Approach. Mc Graw-Hill, Inc., pp. 79–84.
Low, E.L., Grabe, E., Nolan, F., 2000. Quantitative characterisations of speech rhythm: syllable-timing in Singapore English. Lang. Speech 43, 377–401.
Meng, H., Lo, Y.Y., Wang, L., Lau, W.Y., 2007. Deriving salient learners' mispronunciations from cross-language phonological comparisons. In: Proceedings of ASRU.
Mok, P., 2009. On the syllable-timing of Cantonese and Beijing Mandarin. Chin. J. Phon. 2, 148–154.
Mok, P., Dellwo, V., 2008. Comparing native and nonnative speech rhythm using acoustic rhythmic measures: Cantonese, Beijing Mandarin and English. In: Proceedings of Speech Prosody 4, Campinas, Brazil.
Nakata, K., 2002. Eigo No Atama Nikawaru Hon. Chukei Publishing Company, Tokyo.
Nazzi, T., Bertoncini, J., Mehler, J., 1998. Language discrimination by newborns: towards an understanding of the role of rhythm. J. Exp. Psychol.: Hum. Percept. Perform. 24, 756–766.
Patel, A.D., 2003. Rhythm in language and music parallels and differences. Ann. N. Y. Acad. Sci. 999, 140–143.
Pennington, M.C., 1994. Recent research in L2 phonology: implications for practice. In: Morley, J. (Ed.), Pronunciation Pedagogy and Theory: New Views, New Directions. Alexandria, VA, TESOL, Inc., pp. 94–108.
Ramus, F., Nespor, M., Mehler, J., 1999. Correlates of linguistic rhythm in the speech signal. Cognition 73, 265–292.
Ramus, F., Dupoux, E., Mehler, J., 2003. The psychological reality of rhythm classes: perceptual studies. In: Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona, pp. 337–342.
Rivers, W.M., Temperley, M.S., 1978. A Practical Guide to the Teaching of English as a Second or Foreign language. Oxford University Press, New York.
Roach, P., 1982. On the distinction between stress-timed languages and syllable-timed languages. In: Crystal, D. (Ed.), Linguistic Controversies: Essays in Honour of Palmer, F.R. Arnold, London, pp. 73–79.
Roach, P., 2000. English Phonetics and Phonology: A Practical Course. Cambridge University Press, Cambridge.
Sabater, M.S., 1991. Stress and Rhythm in English. Rev. Alicant. Estud. Ingl. 4, 145–162.

Setter, J., 2006. Speech rhythm in world Englishes: the case of Hong Kong. TESOL Q. 40, 763–782.
Taylor, D., 1991. In: Brown, A. (Ed.), Non-native speakers and the rhythm of English. , pp. 235–244, First Published in 1981 in International Review of Applied Linguistics 19, 219–226.
Tuan, L.T., An, P.T.V., 2010. Teaching English rhythm by using songs. Stud. Lit. Lang. 1 (2), 13–29.
Vihman, M., Nakai, S., DePaolis, R., 2006. Getting the rhythm right: a cross-linguistic study of segmental duration in babbling and first words. In: Goldstein, L., Whalen, D.H., Best, C.T. (Eds.), Laboratory Phonology 8. Mouton de Gruyter, New York, pp. 343–368.
Wang, H., Mok, P., Meng, H., 2010. MusicSpeak: capitalizing on musical rhythm for prosodic training in computer-aided language learning. In: Proceedings of the Second Language Studies: Acquisition, Learning, Education and Technology.
Weide, R., 2008. The Carnegie Mellon Pronouncing Dictionary, v. 0.7a, http://www.speech.cs.cmu.edu/cgi-bin/cmudict.
Wong, R., 1987. Teaching Pronunciation: Focus on English Rhythm and Intonation. Prentice Hall Regent, Englewood Cliffs, NJ, pp. 1–53.
Yun, H., 2000. On the content and the methods of English phonetics teaching. J. Guizhou Norm. Univ. (Soc. Sci.) (3), 129–130.