MusicSpeak: Capitalizing on Musical Rhythm for Prosodic Training in Computer-Aided Language Learning

Hao Wang [1], Peggy Mok[2] and Helen Meng [1]*

[1] Human-Computer Communications Laboratory,
Department of Systems Engineering and Engineering Management
[2] Department of Linguistics and Modern Languages
The Chinese University of Hong Kong

Abstract

This paper presents a system named MusicSpeak, which strives to capitalize on musical rhythm for prosodic training in second language acquisition. The system targets for Chinese (L1) speakers learning English (L2). Their speech rhythms are considered to be syllable-timed and stress-timed respectively. Hence, language transfer creates a challenge for Chinese learners in acquiring English rhythm. We develop an automatic procedure that can be applied to any English sentence, to cast rhythmic patterns in speech (based on alternating stressed and unstressed syllables) into rhythmic patterns in music (based on musical bars and beats). We collected speech recordings from 9 speakers uttering 15 English sentences, first in natural style and then in synchrony with the generated musical rhythm. Comparison between the two styles based on rhythm metrics suggests that the latter has higher variability and better approximates stress-timed rhythm.
Index Terms: musical rhythm generation, suprasegmental pronunciation training, prosodic training, CALL

1. Introduction

The use of information and communication technologies (ICT) to support computer-aided language learning (CALL) is gaining increasing momentum. Existing work predominantly address phonetic deviances in L2 (second language) speech viz-a-viz 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. This work is our first attempt to capitalize on musical rhythm for L2 prosodic training. We focus on the Chinese (L1) and English (L2) language pair. Chinese and English have stark contrasts linguistically. A classic view of speech rhythm often categorizes Chinese as a syllable-timed language and English as stress-timed [1,2], an impression which is created by such elements as syllable structure, vowel reduction and stress. Language transfer creates a challenge for Chinese learners in acquiring English rhythm. This appears to be the most widely encountered difficulty among foreign learners of English, and is a major obstacle in acquiring a near-native oral proficiency [3–6].

To address this issue, we attempt to leverage commonalities between speech and music. While both have melodic, rhythmic and linguistically communicative characteristics, music may be considered to exhibit a higher structural rigidity than speech. 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" [7]. Hence, this study attempts to cast English rhythm into musical rhythm for the purpose of prosodic training. We believe that music can enhance learners' engagement in audio-lingual practices.

Previous work that involved musical rhythm for English language teaching include "Jazz Chants" [8] by Graham, which used upbeat chants and poems through jazz rhythms to illustrate the natural stress and intonation patterns of conversational American English. There is also the KenMc method by Nakata [9] that connects spoken English rhythm with the beat of Bossa Nova (a style of Brazilian music).

Both Jazz Chants and the KenMc method are based on given (i.e. fixed) examples of English sentences. Our current work aims to generalize further through the implementation of a system called MusicSpeak. We develop a technique that can automatically generate musical rhythm based on arbitrary English text input. Users are invited to speak English with the musical rhythm output by the system (akin to a karaoke system). We have collected contrastive recordings between naturally spoken L2 English utterances and their counterparts that are recorded alongside the MusicSpeak rhythm. We have also conducted a comparison between the two styles of speech based on rhythm metrics. Details are presented in the following.

2. Automatic Rhythm Generation

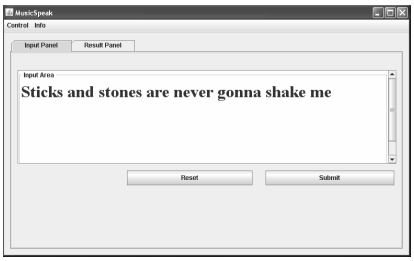
Figure 1 shows the screenshots of the MusicSpeak user interface. Users can enter an arbitrary English sentence in the text box and then click SUBMIT (see Figure 1a). The system generates a musical rhythm according to the input text and displays the output on a new tab (see Figure 1b). The user can click the PLAY button to listen to the generated rhythm, while the corresponding words are highlighted with the beat in a time-synchronous manner. As illustrated in the figure, users can also move the pointer over any word and check its phonetic transcription. Vowels in the syllable with primary stress are highlighted in red. The user interface also color-codes content words differently from function words (the former in red and the latter in green). We devised the following procedure for automatic rhythm generation in MusicSpeak.

2.1. Text Analysis

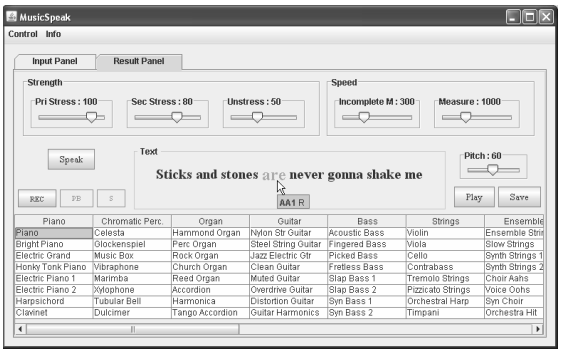
Words in the input sentence are classified as either content words or function words based on a function word list. Content words are typically nouns, verbs, adjectives and adverbs. Function words are articles, conjunctions and pronouns, which have little lexical meaning and mainly serve to express grammatical 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 adopted in our rhythm generation procedure and exceptions often arise. Speakers accent content words by uttering the stressed

* Corresponding author: Helen Meng (hmmeng@se.cuhk.edu.hk)

syllables with higher intensity, pitch and duration. On the contrary, unstressed syllables are acoustically reduced. As will be explained later, we identify the stressed/unstressed syllables in words by means of dictionary lookup, based on the CMUDict [10].



(a)



(b)

Figure 1: *The MusicSpeak interface. (a) user input and (b) system output.*

2.2. Comparing the Rhythmic Features between Speech and Music

Alternating stressed and unstressed syllables in English forms the rhythm of the language [11]. Musical rhythm is manifested in terms of durations and accents of sounds that produce regular patterns in time, constituting the musical beat [12]. The duration and accent of a musical beat may correspond well with those of an English syllable. Musical rhythm imposes a more 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" [13] that does not begin with a stressed syllable (or an accented beat) and has shorter duration than a normal bar. This is illustrated in Figure 2.

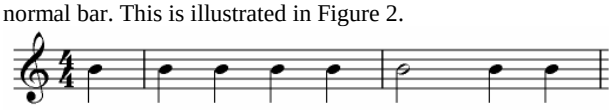


Figure 2: *A piece of music beginning with an incomplete bar.*

2.3. Musical Bar Placement

2.3.1 Heuristics

We devise a set of general heuristics placing syllables with primary, secondary or no stress in musical bars, as follows:

a) Each syllable occupies a beat in a musical bar.
b) Syllables carrying primary stress are always placed in the first beat. Hence, if a sentence does not begin with a primary stressed syllable, the generated rhythm begins with an incomplete bar.
c) Syllable durations in a musical bar are assigned according to the level of stress. Syllables with primary stress are the longest and unstressed syllables are the shortest. Should a musical bar contain only two syllables with different stress levels, we impose a duration ratio of 3:2 between the two syllables.
d) Beat strengths (see upper left corner of Figure 1b) are assigned according to the levels of stress. Syllables with primary stress get the heaviest beat and unstressed syllables get the lightest beat.

These four heuristics result in a total of 5 cases in musical bar placement, illustrated below:

Case 1: <U> |
where <U> denotes one or more unstressed syllables and '|' is the boundary of the (incomplete) musical bar.

Case 2: <u> 2 <u> |
where '2' denotes a syllable with secondary stress, and <u> denotes an arbitrary number (including 0) of unstressed syllables, which also forms an incomplete bar.

Case 3: | 1 <u> |
where '1' denotes a syllable with primary stress. This is a complete bar.

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

Case 5: | 1 <u> 2 <u> 2 <u> |
A complete bar formed by one primary stressed syllable, two secondary stressed syllables and an arbitrary number of unstressed syllables.

2.3.2 Duration assignment

Based on the heuristics and cases in musical bar placement presented above, the rhythm can be generated by assigning appropriate durations to syllables with different levels of stress in each bar. There are 8 variables involved in the calculations.

- D_p, D_s, D_u, D_b and D_i are, respectively, the durations of a primary stressed syllable, a secondary stressed syllable, an unstressed syllable, a complete bar and an incomplete bar. In MusicSpeak, the default values of D_b, and D_i are set to 1 second and 0.3 second. These can be adjusted by users.
- N_p, N_s and N_u represent, respectively, the counts of primary stressed, secondary stressed and unstressed syllables in a bar. These parameters are used in duration assignment for each musical beat, as described below.

Each case in musical bar placement corresponds to specific calculations in duration assignment, as follows:

Case 1: The duration of an incomplete bar is distributed across the number of unstressed syllables, according to Equations (1) and (2).

$$D_u = D_i, \text{if } N_p = 0, N_s = 0 \text{ and } N_u = 1 \quad (1)$$

$$D_u = \frac{D_i}{N_u}, \text{if } N_p = 0, N_s = 0 \text{ and } N_u > 1 \quad (2)$$

Case 2: Equations in (3a,b) impose heuristic (c) if the musical bar contains only one secondary stressed syllable and one unstressed syllable. Equations in (4a,b) handle the more general case where a proportionate duration is assigned to the secondary stressed syllable and the remaining duration distributed among the unstressed syllables.

$$D_s = \frac{3}{5} \cdot D_i, D_u = \frac{2}{5} \cdot D_i, \text{if } N_p = 0, N_s = 1 \text{ and } N_u = 1 \quad (3a,b)$$

$$D_s = \frac{D_i}{N_u}, D_u = \frac{D_i - D_s}{N_u}, \text{if } N_p = 0, N_s = 1 \text{ and } N_u > 1 \quad (4a,b)$$

Case 3: Equation (5) says that the single primary stressed syllable in a musical bar will consume the entire duration.

Equations in (6a,b) and (7a,b) are similar to Equations in (3a,b) and (4a,b) in their rationale.

$$D_p = D_b, \text{if } N_p = 1, N_s = 0 \text{ and } N_u = 0 \quad (5)$$

$$D_p = \frac{3}{5} \cdot D_b, D_u = \frac{2}{5} \cdot D_b, \text{if } N_p = 1, N_s = 0 \text{ and } N_u = 1 \quad (6a,b)$$

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

Case 4: Equations in (8a,b,c) enforce heuristic (c) in section 2.3.1. above

$$D_p = \frac{D_b}{N_s + N_u}, D_s = \frac{D_b}{N_s + N_u + 1}, D_u = \frac{D_b - (D_p + N_s \cdot D_s)}{N_u} \quad (8a,b,c)$$
$$\text{if } N_p = 1, N_s = 1 \text{ and } N_u > 0$$

Case 5: Equations in (9a,b,c) enforce heuristic (c) in section 2.3.1. above

$$D_p = \frac{D_b}{N_s + N_u}, D_s = \frac{D_b}{N_s + N_u + 1}, D_u = \frac{D_b - (D_p + N_s \cdot D_s)}{N_u} \quad (9a,b,c)$$
$$\text{if } N_p = 1, N_s = 2 \text{ and } N_u > 0$$

2.4. Example

This subsection presents an illustrative example of the automatic rhythm generation procedures, based on the input sentence, "*Sticks and stones are never gonna shake me*." A step-by-step walkthrough is as follows:

- MusicSpeak refers to the function word list and identifies the function words and content words in the sentence (content words are boldfaced in row 1 of Table 1).
- MusicSpeak then looks up the CMU Pronunciation Dictionary to obtain the phonetic transcription of the content words, together with information about the stressed vowels ('1' indicates primary stress and '2' secondary stress):

"sticks" → /s t ih1 k s/
"and" → /ah0 n d/
"stones" → /s t ow1 n z/
"are" → /aa1 r/
"never" → /n eh1 v er0/
"gonna" → /g aa1 n ah0/
"shake" → /sh ey1 k/
"me" → /m iy1/

The stress pattern is also shown in the second row of Table 1.

- MusicSpeak then organizes the syllables into musical bars, conforming to the heuristics laid out above. This is illustrated in the third row of Table 1.

Table 1: *An example illustrating the musical rhythm generation process.*

Sentence	**Sticks** and **stones** are **never** gonna **shake** me
Syllable Arrangement	1 0 1 0 1 0 0 0 1 0
Musical Bars	| 1 0 | 1 0 | 1 0 0 0 | 1 0 |

- Finally, MusicSpeak computes the durations for each beat. The third musical bar has 1 primary stressed syllable followed by 3 unstressed syllables. Equations in (7a,b) are applied for duration assignment, i.e.:

$$D_p = \frac{D_b}{N_s + N_u} = \frac{D_b}{0+3} = \frac{1}{3} \cdot D_b$$
$$D_u = \frac{D_b - (D_p + N_s \cdot D_s)}{N_u} = \frac{D_b - D_p - 0}{3} = \frac{2}{9} \cdot D_b$$

The procedure is similar for all the other musical bars.

3. Evaluation

3.1. Corpus

The MusicSpeak system was implemented in Java. In order to investigate the effectiveness of the model, we randomly sampled 15 English sentences from song lyrics. The number of words per sentence range from 7 to 12. Examples are shown in Table 2. We also recruited subjects to record each sentence in two speaking styles – first naturally and then alongside the generated rhythm from MusicSpeak. All our volunteers are undergraduate students from The Chinese University of Hong Kong (5 male, 4 female). Each subject is allowed to practice reading the sentences in any style as many times as they like before the actual recording. We recorded 270 utterances (15 sentences × 9 speakers × 2 styles). Each recording is digitized at 16kHz sampling rate and stored at 16 bits per sample, mono, in .wav format.

Table 2: *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.
etc.

3.2. Data Analysis

We obtain phonetic boundaries for all recordings by means of forced alignment with an automatic speech recognizer [14]. The phone segmentations are then mapped automatically into consonantal and vocalic intervals and thereafter syllabic intervals. Any anomaly is manually adjusted in Praat with reference to acoustic cues and careful listening. Segmentation criteria followed those in [1]. Phonotactic constraints and the maximal onset principle are used in deciding syllable boundaries [15]. Durations (ms) of syllabic, consonantal and vocalic intervals are extracted using a Praat script. Any silent pause within an utterance is excluded from further analysis. The Pairwise Variability Index (PVI) [1] is used to compare the rhythmic difference between the utterances spoken normally and those following the generated rhythms. The PVI expresses the level of durational variability in successive intervals. There are two versions of the PVI, raw (see Equation 10) and normalized (see Equation 11):

$$rPVI = \left[\sum_{k=1}^{m-1} |d_k - d_{k+1}| / (m-1) \right] \quad (10)$$

$$nPVI = 100 \times \left[\sum_{k=1}^{m-1} \left| \frac{d_k - d_{k+1}}{(d_k + d_{k+1})/2} \right| / (m-1) \right] \quad (11)$$

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

Raw PVI, taking the absolute difference in duration between each pair of successive units, is calculated for consonantal (rPVIC) and syllabic (rPVIS) duration. Normalized PVI uses the mean duration of each pair of successive units to normalize for speech rate variations. Normalized PVI is also calculated for vocalic (nPVIV) and syllabic (nPVIS) duration. Only raw PVI is calculated for consonant intervals because normalization for speech rate may also eliminate differences due to syllable structure (see [1]). PVIs for syllabic intervals are included based on the results in [2] and [15] which show that syllable duration can also robustly classify languages into distinct rhythmic groups.

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 normally.

3.3. Results

Figures 3 and 4 show the average PVI values of individual speakers for each of the two styles (i.e. normal versus rhythmic). The two styles appear to be separated in their PVI

values. Paired-sample t-tests confirmed that utterances spoken following the generated rhythms have higher values for raw consonantal PVI [t(8) = -3.955, p = 0.004] and raw syllabic PVI [t(8) = -4.393, p = 0.002]. The normalized syllabic PVI also shows a similar trend [t(8) = -2.165, p = 0.062]. These results confirm that speakers do have more variable speech timing when they follow the generated rhythm.

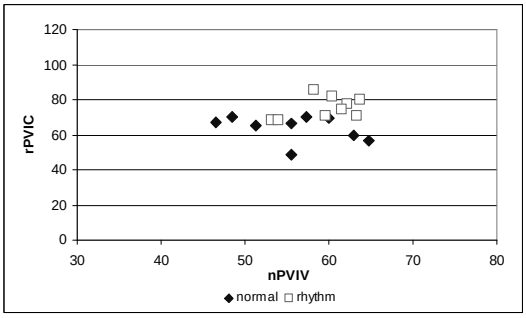


Figure 3: *Raw consonantal PVI and normalized vocalic PVI.*

In contrast, there is no significant difference in the normalized vocalic PVIs between the two speaking styles [t(8) = -1.453, p = 0.184]. This is probably because there are more pauses in the utterances following generated rhythm than those spoken normally. Sometimes only one or two long beats occupy a bar. Speakers naturally slowed down and lengthened the target syllables in order to follow the generated rhythms closely. This resulted in much vowel lengthening for these syllables, which reduced the durational variability between vocalic intervals. Since consonant duration is much less affected by pauses and lengthening in speech (see [16]), the significant result of consonantal intervals suggests that the speakers were indeed speaking with a more variable (stress-timed) rhythm.

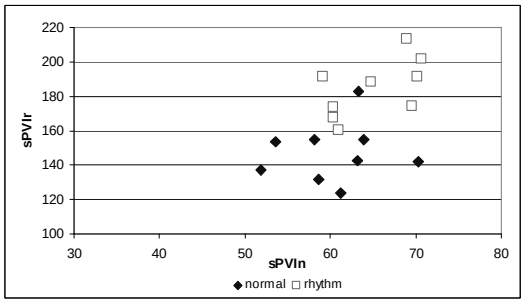


Figure 4: *Raw and normalized PVI for syllable duration.*

4. Future Developments

We have developed 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 any English sentence input, where rhythmic generation considers the discrimination between content and function words in the sentence, as well as the locations of stressed syllables. We collected speech recordings from 9 speakers uttering 15 English sentences, first in natural style and then in synchrony with the generated musical rhythm. 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. This implies that the use of musical rhythm in suprasegmental training for second language acquisition is a promising approach. Data analysis suggests that the rhythm generation procedure may be enhanced by packing more syllables into a musical bar to prevent unnecessary vowel lengthening in the rhythm-synchronized speech recordings. Previous experimental studies also suggests that it may be possible for generation considerations to extend beyond the syllable level – to seek the precise location of the beat within the syllable [11] or to consider timing at the phrasal level [17].

5. Acknowledgements

This work is partially supported by a grant from the HKSAR Government Research Grants Council General Research Fund (project number 416108). We thank Shuang Zhang for her support in labeling of the recorded speech data.

6. References

[1] Grabe, E. and Low, E.L., "Durational variability in speech and the rhythm class hypothesis", Laboratory Phonology VII, Gussenhoven C., Warner, N. (eds.). Berlin: Mouton de Gruyter, pp. 515-546, 2002.

[2] Mok, P. and Dellwo, V., "Comparing native and non-native speech rhythm using acoustic rhythmic measures: Cantonese, Beijing Mandarin and English", *Proc. of Speech Prosody 4*, Campinas, Brazil, 2008.

[3] Chela-Flores, B., "On the acquisition of English rhythm: Theoretical and practical issues", *Lenguas Modernas 20*, pp. 151-164, 1993.

[4] Faber, D. "Teaching the rhythms of English: a new theoretical base", *Intl Rev. of Applied Linguistics 24*, 1986.

[5] Taylor, D., "Non-native speakers and the rhythm of English", *Intl Rev. of Applied Linguistics 19*, 1981.

[6] Anderson, P., "Defining intelligibility parameters: The secret of sounding native.", XXVII TESOL Annual Convention, Atlanta, 1993.

[7] Patel, A. D., "Rhythm in Language and Music Parallels and Differences", *Ann. N.Y. Acad. Sci. 999*, 2003.

[8] Graham, Carolyn, *Jazz Chants*, 1st Ed. New York: Oxford University Press, 1978.

[9] Nakata, K., *Eigo No Atama Nikawaru Hon.* Tokyo: Chukei Publishing Company 2002.

[10] The Carnegie Mellon University (CMU) Pronunciation Dictionary, http://www.speech.cs.cmu.edu/cgi-in/cmudict.

[11] Allen, G. D., "The Location of Rhythmic Stress Beats in English: An Experimental Study I", *Lang. Speech* 15, 1972.

[12] Hawes, Neil V., "Basic Music Theory", 2003. [Online], Available: http://neilhawes.com/sstheory/theory16.htm.

[13] Blood, B., "Dolmetsch Online - Music Theory Online",: http://www.dolmetsch.com/musictheory4.htm.

[14] Meng, H. et al., "Deriving Salient Learners' Mis-pronunciations from Cross-Language Phonological Comparisons," *Proceedings of ASRU*, 2007.

[15] Deterding, D. The measurement of rhythm: a comparison of Singapore and British English. *J. of Phonetics*, 29, 2001.

[16] Gay, T.: "Mechanisms in the control of speech rate.," *Phonetica 38*, 1981

[17] Klatt, Dennis H., "Linguistic uses of segmental duration in English: Acoustic and perceptual evidence", *J. Acoustical Society of America*, 59(5), 1976.