The 18th Oriental COCOSDA / CASLRE

2015 International Conference Oriental COCOSDA held jointly with 2015 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)

Proceedings

Shanghai Jiao Tong University
Shanghai, Oct. 28-30, 2015
Table of Contents

1. Noise-Robust and Stress-Free Visualization of Pronunciation Diversity of World Englishes Using a Learner's Self-Centered Viewpoint ........................................... 1
2. Elicit Spoken-Style Data from Social Media Through a Style Classifier ........... 7
3. Stress Annotated Urdu Speech Corpus to Build Female Voice for TTS .......... 13
4. Toward Improving Estimation Accuracy of Emotion Dimensions in Bilingual Scenario Based on Three-Layered Model................................................................. 21
5. Development of New Speech Corpus for Elderly Japanese Speech Recognition 27
6. Corpus Design and Development of an Annotated Speech Database for Punjabi..................................................................................................................... 32
7. Chinese Traditional Opera Database for Music Genre Recognition ............. 38
8. A Study on Adaptation of Speaking Rate-Dependent Hierarchical Prosodic Model for Chinese Dialect TTS ................................................................................ 42
9. Contrastive Study of Focus Phonetic Realization Between Jinan Dialect and Taiyuan Dialect ........................................................................................................... 47
10. On Finding Word-Level Break-Type Formation Rules for Mandarin Read Speech...................................................................................................................... 53
11. The Recognition of Neutral Tone Across Acoustic Cues ............................. 58
12. Prosodic Processing in Developmental Dyslexia: A Case Study in Standard Chinese.................................................................................................................... 64
13. English Rhythm of Guangxi Zhuang EFL Learners .................................... 69
14. Information Content, Weighting and Distribution in Continuous Speech Prosody – A Cross-Genre Comparison ........................................................................... 75
15. An Open/Free Database and Benchmark for Uyghur Speaker Recognition ...... 81
16. Analysis of Intonation Patterns in Cantonese Aphasia Speech .................... 86
17. Melody of Mandarin L2 English—When L1 Transfer and L2 Planning Come Together ..................................................................................................................... 90
18. A Study of the Production of Unstressed Vowels by Japanese Speakers of English Using the J-AESOP Corpus ............................................................................. 96
19. Dual-Focus Intonation in Standard Chinese ................................................. 101
20. Acoustic Analysis of English Lexical Stress Produced by Native (L1) Bengali Speakers Compared to Native (L1) English Speakers .............................. 107
21. Cross-Linguistic Perception of Chinese Attitudes Praising and Blaming........113
22. Tonal Phoneme Based Model for Vietnamese LVCSR .....................118
23. An Articulatory Analysis of Apical Syllables in Standard Chinese.........123
24. Tonal Alignment in Shanghai Chinese........................................128
25. Objective Verification of Assamese Consonants..............................133
26. Segment Reduction in Disyllabic Words with Tones in Standard Chinese ......139
27. Time Group Types in Mandarin Syllable Annotations ........................145
28. Coda's Duration on Perception of Mandarin Syllables with Alveolar/Velar Nasal Endings by Japanese CSL Learners ...........................................150
29. Exploring Tonal Effects on the Perception of Word-Final Nasals: A Preliminary Study in Southern Min .................................................................155
30. Context-Dependent Grapheme-to-Phoneme Evaluation Corpus Using Flexible Contexts and Categorial Matrix ...........................................................160
31. Influence of Regional Dialects on Acoustic Characteristics of Hindi Vowels ..166
33. The Acoustic and Auditory Performances of the Three Thai Level Tones ......177
34. An Empirical Study of Phonetic Transfer in English Monophthong Learning by Tibetan (Lhasha) Speakers .................................................................181
35. Analysis on L2 Learners’ Perception Errors Between Geminate and Singleton of Japanese Consonants Using Loudness Related Parameters ....................186
36. An Experimental Study of the Production of English Diphthongs by Chinese College EFL Learners: An Acoustic Perspective .........................................190
37. Automatic Word Segmentation for Spoken Cantonese........................196
38. Construction and Analysis of Social-Affective Interaction Corpus in English and Indonesian .........................................................................................202
39. District Names Speech Corpus for Pakistani Languages .......................207
40. Cross-Linguistic Prosodic Comparison with OMProDat Database .............212
41. A Study of Labeling System for “CAPT Corpora of BLCU” ..................216
42. Speech Enhancement with EMD and Correlation Mode Selection .............220
43. Building Multiple Number of Sets of Phonetically Rich Sentences from Web Crawled Text .........................................................................................224
44. Thai Pseudo Syllable Segmentation Using Conditional Random Field ....230
45. Development of Marathi Language Speech Database from Marathwada Region

46. KALDI ASR-Based Speech Database Evaluation Method

47. Language Independent Automatic Pronunciation Scoring with Features Based on Dynamic Time Warping

48. The Break Annotation Based on Acoustic Cues — A Case Study on Neighboring Tone4 and Tone2

49. The Study of Correlation Between Voice Quality Measurements and F0 Parameters Based on Monosyllabic Chinese Corpus

50. Automatic Extraction of Fujisaki Model Parameters for Bangla Language

51. G2P-Free Grapheme-to-Speech Synthesis: UTF-8 Based Automatic Unit-Database Annotation and Mixed Viterbi UNIT-Selection

52. Amrita SRE Database: A Database for Evaluating Speaker Recognition Systems with Mimicked Speech

53. Development of Multiple Automatic Speech Recognition Systems in the Galaxy Framework

54. The Perception of The English Nasal Coda Contrast by Bilingual Changsha Chinese Learners of English

55. A Diachronic Research on Vowel System of Xianghua

56. Does Silent Pause Facilitate Focus Perception in a Non-PFC Language? An Experimental Study of Li Language

57. Measuring Reaction Time of Chinese Tone Identification for Finer Evaluation of L2 Learner’s Proficiency

58. Speech Rhythm of The Danish-Chinese Interlanguage Relies on Rhyme Structure

59. A Tentative Study on Mandarin Tone Perception by Japanese Children in Immersive Chinese Learning Program

60. An Acoustic Study of Tones in Kaifeng: Citation Forms and Contextual Variations

61. Analysis of Voice Quality of Emotional Speech Collected in Various Speech Situations

62. An Experimental Phonetic Study on Acquisition of English Vowel Segments by Learners of Shandong Dialect: A Case Study of Weifang, Mengyin, Qingdao and Rizhao

63. Comparison of Statement and Question Intonations Between Urumqi Chinese Dialect and Standard Mandarin Speakers
<table>
<thead>
<tr>
<th>No.</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>64.</td>
<td>Production of English Segments by Advanced Mandarin-Speaking Speakers</td>
<td>335</td>
</tr>
<tr>
<td>65.</td>
<td>An Experimental and Acoustic Analysis on the Chinese Primary Vowel Acquisition by Malay Students</td>
<td>339</td>
</tr>
<tr>
<td>66.</td>
<td>Analysis of Rhythmic Patterns Using Both Interval-Based and Amplitude-Based Measures in Native and L2 Speech</td>
<td>343</td>
</tr>
<tr>
<td>67.</td>
<td>Emotional Speech Conversion with Prosodic Cues in Mandarin Chinese</td>
<td>347</td>
</tr>
</tbody>
</table>
Analysis of Rhythmic Patterns Using both Interval-based and Amplitude-based Measures in Native and L2 Speech

Shuju Shi, Bei Peng, Jinsong Zhang, Yanlu Xie

College of Information Science, Beijing Language and Culture University

Abstract

This study investigates rhythmic patterns on Mandarin Chinese, Japanese and Japanese L2 Chinese using both interval-based and amplitude-based measures. Interval-based measures include published rhythm metrics concerning variation in duration of consonantal and vocalic intervals. Amplitude-based measures include measures deriving from Fourier analysis of envelope spectra. Results show that: 1) For interval-based and amplitude-based measures, Japanese L2 Chinese is different from the other two corpus types and that difference is statistical significant. Two possible reasons account for this. The first one is the relatively slow speech rate in Japanese L2 Chinese and the second one is that phonology difference between Chinese and Japanese might have affected the phonetic realization of Japanese L2 Chinese; 2) The influence of speech rate on interval-based measures is significant. The distribution of three corpus types is different in raw interval measures and speech-rate normalized interval measures. And measures of Japanese L2 Chinese are more easily influenced by speech rate; 3) Data in L2 speech suggests a possible L1 transfer on L2.

Index Terms: speech rhythm, L2, isochrony, amplitude-based measure, interval-based measure

1. Introduction

Speech rhythm, or isochrony, is described as the phenomenon of repetition of elements perceived having equivalent duration in speech. Pike [1] first proposed the existence of two types of rhythm based on Lloyd-James’[2] description of language as ‘machine gun’ and ‘morse-code’: syllable-timed languages where the repetition unit is syllable and stress-timed languages where the intervals between stresses are supposed to be the repetition units. Abercrombie further suggested that all languages can be classified into either rhythm type mentioned above. Later a number of studies [3] expended the rhythm categories to include mora-timed languages like Japanese.

Subsequent studies tried to interpret speech rhythm using laboratory methods based on acoustic signals, ending up with varying results. Bolinger[4] and Lea[5] measured durations of stress bars and found rhythmic feet are not isochronous in English. Dasher & Bolinger [6] and Dauer [7] suggested that the perception of equal duration-units might be the result of phonological differences which relate to syllable structure, vowel reduction and word stress. One key problem of these studies is that they equaled perceptual isochrony with the measurement of duration from acoustic signal. Lehiste [8,9] conducted perceptual experiments using both speech and non-speech stimuli and concluded that isochrony can be regarded as a perceptual phenomenon. Psycho-linguistic studies also showed that both adults and infants could discriminate between languages of their own rhythmic class and that from a different rhythmic class. Research by Nazzi [10] showed that infants, including newborns, can discriminate between sentences taken from their own language and those from a language belonging to a different rhythmic class (stress-timed or syllabled-timed). Ramus [11] got similar results on adults using synthesized stimulus where all segmental information is removed and only the rhythmic pattern is kept.

What could be concluded for now is that isochrony does exist but it is a more perceived subjective than measured objective conception. This means that the study of speech rhythm should not be limited to interval measures, parameters such as fundamental frequency, amplitude and spectral dynamics should also be incorporated [12]. Also, the validity of measures based on acoustic signals should be furthered confirmed by perceptual tests. Tilsen et al. [13-14] proposed several metrics derived from amplitude envelop and has got relative success in characterizing speech rhythmicity from a perspective of recurring patterning of amplitude. Fuchs [15] tried to quantify difference of F0 on the perception of duration and got a perceptually weighted measure of the old interval measures.

In this study, we aimed to:

- Cross-language comparison of rhythm measures on language pairs of Mandarin Chinese and Japanese;
- Comparing of speech rhythm by second learners (L2) with both that of their first language (L1) and that of the target language.

The reason why we chose language pairs of Mandarin Chinese and Japanese is attributed to the fact that Japanese is mora-timed and Chinese is syllabled-timed whereas temporal variation of the syllable level requires more explicit understanding. Former studies [16, 17] have shown that L2 speech scored differently in interval-based rhythm metrics from both their own L1 and the targeted language. We employed both interval-based metrics and amplitude-based metrics to testify whether this is the case in Japanese L2 of Chinese and how rhythm of L1 might influence the production of L2.

2. Methodology

This work tried to study rhythmic patterns from two perspectives, interval-based measures and amplitude-envelope measures. We intended to firstly testify the effectiveness of each method and also tried to interpret how these methods account for rhythm difference and in what way they may
interact with each other. The following parts include speech corpus design, specific metrics used and their calculating method, preprocessing of data, method of analysis and why the specific method is chosen.

2.1. Corpus Description

Six native Chinese speakers (all female) and nine native Japanese speakers (all female) took part in this study and all of them are students in the first author’s University. *Conversational Chinese 301* was chosen for Chinese speech and it was translated into and slightly adapted to Japanese. All recordings took place in a soundproof room. The Chinese speakers read Chinese material only, all Japanese speakers read Chinese materials and only seven of them read Japanese materials. Then speech files were digitized at the sampling rate of 44.1 KHz with the quantization precision being 16 bits and were saved as wav format. Average syllable number of all sentences is 7.7. Average speech rate for native Chinese speakers is 5.2 syllables / second, 4.3 syllables / second for by Japanese L2 learners, and 9.0 moras / second for native Japanese speakers.

Automatic segmentation was done using HTK and then the material was manually checked by six master students majoring in phonetics in the first author’s university using praat.

2.2. Interval-Based Metrics

The interval-based metrics we used here include $\Delta C$, $VacroC$, $\Delta V$, $VacroV$, rPVI-C, rPVI-V, nPVI-C, and %V [18-20]. Their meaning and calculation method list as below:

- $\Delta C$: The standard deviation of consonantal interval duration within each sentence.
  
  $VacroC$: A normalization version of $\Delta C$, calculated as,
  $$VacroC = \frac{100 \times \Delta C}{meanC}$$
  (1)

  where $meanC$ refers to the mean value of all consonantal durations within each sentence.

- $\Delta V$: The standard deviation of vocalic interval duration within each sentence.
  
  $VacroV$: A normalization version of $\Delta V$, calculated as,
  $$VacroV = \frac{100 \times \Delta V}{meanV}$$
  (2)

  where $meanV$ refers to the mean value of all vocalic durations within each sentence.

- rPVI-C: the raw Pairwise Variability Index for consonantal intervals

- rPVI-V: the raw Pairwise Variability Index for vocalic intervals

- rPVI is calculated as:
  $$rPVI = \frac{\sum_{k=1}^{m-1} |d_k - d_{k+1}|}{m-1}$$
  (3)

  where $m$ indicates the total number of vowels or consonants, $k$ means the $k$th element and $d_k$ and $d_{k+1}$ refers to the duration of the $k$th and $(k+1)$th element.

- nPVI-C: The normalized Pairwise Variability Index for consonantal intervals.

- nPVI-V: The normalized Pairwise Variability Index for vocalic intervals.

- nPVI is calculated by the formula below:
  $$nPVI = 100 \times \frac{\sum |d_k - d_{k+1}|}{(d_k + d_{k+1}) / 2}$$

- %V: The sum of vocalic interval duration divided by the total duration of vocalic and consonantal intervals and multiplied by 100.

2.3. Amplitude-Based Metrics

To characterize the recurring timing patterns of amplitude envelope, Fourier transform is adopted here. Fourier transform is a method to partition variations of time-domain sequence into differing Fourier analysis frequencies. And each frequency is assigned with certain amplitude. The resulting frequencies associated with the amplitude offers us a way to interpret temporal patterns of amplitude envelope.

Preprocessing of data and calculation of metrics derived from this method is explained in detail following the study of [14].

2.3.1. Preprocessing of Data

To limit possible influence of chunk duration on the effectiveness of metrics, only sentences between 1-2 seconds are selected (excluding pauses longer than 100ms). Then we band-pass filtered the speech sound using Butterworth with frequency cut-off being [400, 2000]. The low-frequency cutoff is kept as in [14] with a similar aim, that is to de-emphasize the contribution of F0 and thereby decrease the extent to which the presence of voicing is directly represented in the signal. We changed the high-frequency from 4000Hz to 2000Hz, for by this way we could decrease the representation of most consonants and still keep the intelligibility of vowels. Then a Tukey window ( $r = 0.1$) is applied to window the envelope.

2.3.2. Spectral Analysis

After the preprocessing procedure, we applied Fourier transform to the extracted envelope. Smoothing of the envelope is done using the method proposed in [21] and cited by [14]. An example of the power spectrum after Fourier transform and smoothing is given in Figure 1.

Figure 1. a). Band-passed waveform, b). amplitude envelope and c). power-frequency spectra
Two metrics based on the spectrum proposed by [14] are adopted here:

- The first one is spectral band power ratio (SBPr). It is calculated by firstly defining relatively low and high frequency bands, and computing the ratio of the power in the lower band to the power in the higher band.
- The second one is the power-spectral centroid (p-center), which is computed by summing all of the frequencies multiplied by their associated spectral power and then dividing by the sum of all spectral power.

The first metric is based on the assumption that lower-frequency in the envelope corresponds to higher prosodic components and higher-frequency corresponds to influence of syllable or mora on rhythm.

In our study, the cutoff between the bands is 3.25Hz for the longest syllable in our speech materials is about 300ms. And the range for the second metric is set to be 1.5-10Hz as in [14] for the average length of mora is about 100ms in our speech material.

3. Results

3.1. Interval-based Measures

Figures 2 and 3 showed the mean value of interval-based measures classified by different corpus type, that is, native Chinese speech (CC), Japanese L2 of Chinese (JC) and native Japanese speech (JJ).

ANOVA showed that corpus type (CC, JC and JJ) is a significant factor in all interval-based measures. Post hoc Tukey HSD comparisons showed that all three corpus types are statistically different from each other on the chosen interval-based metrics except for JC & JJ on $\Delta C$ ($p = 0.218$) and CC & JJ on $\Delta V$ ($p = 0.169$).

3.1.1. Chinese V.S. Japanese

Figure 1 together with ANOVA results showed that: 1) compared with Chinese, Japanese has higher variation on duration of consonants and lower variation on duration of vowels; 2) The fact that $\Delta C$ of Japanese is larger than Chinese while $rPVI_C$ is smaller than Chinese suggested that dynamic range of consonants in Japanese might be narrower than that in Chinese; 3)And the consistent result in $\Delta V$ and $rPVI_V$ showed that dynamic range of vowels in Chinese is larger than that in Japanese.

Figure 2 together with ANOVA results showed that: 1) The percentage of vocalic duration in Chinese are much longer than that in Japanese; 2) Results in $vacroC$, $nPVI_C$ together with $\Delta C$ and $rPVI_C$ in Figure 2 revealed that consonantal duration in Japanese is shorter than that in Chinese.

Results in this part confirmed the idea that mora-timing Japanese are similar in structure with syllable-timing Chinese but shorter in length.

3.1.2. Native Speech V.S. L2 Speech

Comparing speech by Japanese L2 learners with both their own L1 and native Chinese speech, we found that: 1) Japanese speakers tend to pronounce Chinese consonants with longer duration and larger dynamic range; 2) Results in $\Delta V_r$, $rPVI_V$ and $nPV_{VI}$ seemed to suggest a similar phenomenon in vowel like that in consonants. However, the fact that $nPVI_V$ of Japanese L2 is smaller than both native Japanese speech and native Chinese Speech revealed the fact the longer duration and larger dynamic range of vowels in Japanese L2 speech might due to slower speech rate. Also, lack of compound vowels in their own language may lead them to pronouncing Chinese compound finals longer than normal.

3.2. Amplitude-based Measures

Figure 3 displayed the mean value of the two amplitude-based measures. Statistical analysis (ANOVA) showed that speech corpus type (CC, JC and JJ) is a significant factor in both measures.

The considerably high SBPr value and low p-center value in L2 Japanese showed that speech by L2 Japanese are less rhythmic compared with native Chinese and native Japanese.
The relatively slow speaking rate of L2 Japanese might also lead to this lack of rhythmicity.

![Figure 5: Speech Rate for each corpus.](image)

P-center values for CC, JC and JJ are 5.06 Hz, 2.56 Hz and 5.09 Hz, which means the recurring timing period for each corpus are 200ms, 390ms and 200ms. The results for CC and JC are easy to interpret when speech rate is incorporated. However, the result for JJ is a bit weird. One possible reason for this might be the existence of long vowel in Japanese speech. The maintenance of high amplitude resulting from long vowels to some extent contributes to the longer period in Japanese.

### 4. Discussion

The two methods used in this research both showed their effectiveness in quantifying speech rhythm difference. But what they are accounting for is quite different. Results in the interval-based metrics reflect more about phonological difference of syllable structure in Mandarin and Japanese. Amplitude-based metrics deals with fluctuation in signal energy and tried to characterize recurring timing pattern. One thing to be noticed is that although the amplitude-based method is independent of phonological units, it is inevitably influenced by phonological structure of language, which can be reflected in the p-center value in JJ. Also, speech rate also plays an important role in interpreting the results by amplitude-envelope method.

### 5. Conclusions

Using both interval-based and amplitude-based metrics, this work studied the rhythmic difference between Mandarin Chinese and Japanese as well as rhythmic pattern in Japanese L2 of Chinese. The results showed that: 1) Both methods are effective in quantifying rhythm difference in Japanese and Chinese; 2) The rhythmic difference is more salient between L2 speech and native speech than that between native Chinese and native Japanese; 3) p-center and SPBr showed that L2 Japanese is less rhythmic compared with both native Chinese speech and their own L1; 4) Interval-based metrics are effective in characterizing phonological difference while amplitude-based measures are easily influenced by phonological structures and speaking style; 5) Data in interval-based metrics also suggest a possible L1 transfer on duration patterning of L2.

### 6. Acknowledgements

I’d like to thank all the lab-mates for their hard-work of manually checking the annotations. This work was supported by the National Science Foundation of China (NO. 61175019), Beijing Higher Education Young Elite Teacher Project(YETP0879) and Research Projects of Beijing Language and Culture University (Special Funds of Basic Research Costs for the National University) (13YBG48).

### 7. References