A Corpus-Based Study on Japanese English Rhythm

Takayuki Konishi

January 17, 2022

A doctoral dissertation submitted to
the Graduate School of International Culture and Communication Studies
Waseda University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Table of Contents

1. Introduction 1

1.1. Rhythms of languages 1

1.2. Acquisitions of English rhythm by native speakers of Japanese 3

1.3. Purpose of the current study 4

1.4. Outline of the following chapters 5

2. Acquisitions of English Rhythm by Native Japanese Speakers 9

2.1. Introduction 9

2.2. English rhythm 11

2.2.1. Foot rhythm in English 11

2.2.2. Isochrony in English foot rhythm 16

2.2.3. Lexical stress as rhythmic beat in English foot rhythm 19

2.2.4. Phonetic correlates of English lexical stress 24

2.3. Japanese rhythm 33

2.3.1. Units of Japanese rhythm 33

2.3.2. Isochrony in Japanese mora rhythm 36

2.4. Japanese English rhythm 38

2.4.1. Lexical stress in Japanese English 38

2.4.2. Vowel epenthesis in Japanese English 42

2.4.3. Foot rhythm in Japanese English 45

2.5. Summary 46

3. The Corpus 50

3.1. Introduction 50

3.2. Subjects 52

3.3. Methods 53

3.4. Data for the current study 54

3.5. Annotation ... 55

3.6. Summary ... 64

4. Perceived Proficiency of Japanese English ... 65

4.1. Introduction ... 65

4.2. Raters ... 67

4.3. Procedures ... 69

4.4. Inter-rater reliability ... 71

4.4.1. Segment ... 72

4.4.2. Prosody ... 74

4.4.3. Fluency ... 76

4.4.4. Nativelikeness ... 78

4.5. Constituents of nativelikeness ... 80

4.6. Proficiency score used in the remaining analyses ... 83

4.7. Summary ... 86

5. Lexical Stress in Japanese English ... 87

5.1. Introduction ... 87

5.2. Data coding and normalisations ... 91

5.2.1. Duration ... 92

5.2.2. Intensity ... 93

5.3. Analyses ... 94

5.3.1. Stress and vowel duration ... 94

5.3.2. Stress and vowel intensity ... 96

5.4. Discussion ... 98

5.5. Summary ... 101

6. Epenthetic Vowels in Japanese English ... 102

6.1. Introduction ... 102

6.2. Data coding and normalisations ... 105

6.2.1. Duration 105

6.2.2. Intensity 106

6.3. Analyses 107

6.3.1. Frequencies of epenthesis 107

6.3.2. Duration of epenthetic vowels 109

6.3.3. Intensity of epenthetic vowels 111

6.4. Discussion 113

6.5. Summary 115

7. Foot in Japanese English 116

7.1. Introduction 116

7.2. Data coding and normalisations 120

7.3. Analyses 125

7.3.1. Variances of the foot duration 125

7.3.2. Foot duration and the number of syllables 128

7.3.3. Compensatory shortening in stressed syllables 132

7.3.4. Compensatory shortening in unstressed syllables 136

7.3.5. Phonetic correlates of foot rhythm in Japanese English 140

7.4. Discussion 153

7.5. Summary 162

8. General Discussions 164

8.1. Perceived proficiency of Japanese English rhythm 164

8.2. Isochrony in English rhythm 167

8.3. Implications for teaching English rhythm to Japanese learners 172

9. Conclusions 174

References 176

Appendix: List of recordings in the *J-AESOP* corpus 188

Tables

Table 1 Abbreviations used in this dissertation 8

Table 2 Example analyses of English rhythm by Cruttenden (2014, p 272)..........................22

Table 3 A counterexample to the vowel-based analysis by Cruttenden (2014).....................23

Table 4 Three-way distinction of English stress in the latest theory of metrical phonology...25

Table 5 Previous studies on the comparison of English vowels with primary stress VS no stress 27

Table 6 Previous studies on the comparison of English vowels with primary VS secondary stress 29

Table 7 Previous studies on the comparison of English vowels with secondary stress VS no stress (or between stressed VS unstressed vowels in non-pitch-accented positions) 31

Table 8 Examples of closed syllables in Japanese 34

Table 9 Possible and impossible truncations of stems attaching to the suffix -tyan in Japanese (Poser 1990) 35

Table 10 Parameters used for phonological contrasts in NE and JE 39

Table 11 Hypothetical correlates of lexical stress in NE and JE...39

Table 12 The environments and types of epenthetic vowels in JE.......................................43

Table 13 The variants of the pronunciations of the word ‘closely’ in the pronunciation dictionary for the automatic alignment of the J-AESOP corpus (Kondo 2012) 56

Table 14 Examples of words with epenthetic vowels in the J-AESOP corpus 60

Table 15 An example of epenthetic vowels in the J AESOP corpus (within a consonant cluster) 62

Table 16 An example of epenthetic vowels in the J AESOP corpus (after a coda)...............63

Table 17 Raters of the J-AESOP corpus 67

Table 18 Instruction given to each rater of the J-AESOP corpus 69

Table 19 Three sections of the North Wind and the Sun for the rating of the J-AESOP corpus 70

Table 20 Mean and standard deviation of the segment scores of each rater, and the Cronbach's Coefficient Alpha value for the all-rater group when the rater in the row is excluded 72

Table 21 Mean and standard deviation of the prosody scores of each rater, and the Cronbach's Coefficient Alpha value for the all-rater group when the rater in the row is excluded 74

Table 22 Mean and standard deviation of the fluency scores of each rater, and the Cronbach's Coefficient Alpha value for the all-rater group when the rater in the row is excluded 76

Table 23 Mean and standard deviation of the nativelikeness scores of each rater, and the Cronbach's Coefficient Alpha value for the all-rater group when the rater in the row is excluded 78

Table 24 Standardised coefficient of the main effect of segment, prosody and fluency score on the nativelikeness score (the EN rater group) 81

Table 25 Standardised coefficient of the main effect of segment, prosody and fluency score on the nativelikeness score (the JP rater group) 82

Table 26 Degree of stress contrasts made with the phonetic parameter compared to NE ... 88

Table 27 The hypotheses and results of the analysis of the NE and JE lexical stress 99

Table 28 The hypotheses and results of the analysis of the JE vowel epenthesis 113

Table 29 List of the feet for the analysis 122

Table 30 Mean normalised duration of foot in relation to the number of syllables 131

Table 31 Mean normalised duration of unstressed syllables in relation to the number of syllables in the foot 139

Table 32 Standardised coefficients of the predictor variables on the variance of the foot duration (the JP Adv group) 149

Table 33 Standardised coefficients of the predictor variables on the variance of the foot duration (the JP Beg group) 150

Table 34 The hypotheses and results of the analysis of the foot duration 153

Table 35 Mean normalised duration of the stressed vowels in relation to the number of foot-internal syllables 158

Table 36 The hypotheses and results of the analyses 159

Table 37 The research question and the result of the analysis of the foot duration 160

Figures

Figure 1 English foot in its prosodic hierarchy 12

Figure 2 Examples of degenerate feet in English 14

Figure 3 The structure of English foot adopted in the current study 15

Figure 4 Example of the annotation in the J-AESOP corpus 57

Figure 5 An example of segment boundaries defined proportionally 58

Figure 6 Examples of mispronunciation, repetition and dysfluency 59

Figure 7 Distribution of the grand mean scores 84

Figure 8 Distributions of prosody scores of the JP Adv and JP Beg groups 85

Figure 9 Normalised duration of stressed and unstressed vowels 95

Figure 10 Normalised vowel intensity of stressed and unstressed vowels 97

Figure 11 The number of epenthetic vowels for each speaker 107

Figure 12 Normalised vowel duration in relation to the vowel type 110

Figure 13 Normalised vowel intensity in relation to the vowel type 112

Figure 14 Distributions of the mean foot duration 126

Figure 15 Distributions of standard deviations of the variance of the foot duration 127

Figure 16 Foot duration in relation to the number of syllables in the foot 129

Figure 17 Foot duration in relation to the number of syllables in the foot 130

Figure 18 Duration of the foot-internal stressed syllables in relation to the number of syllables in the foot 133

Figure 19 Duration of the foot-internal stressed syllables in relation to the speaker group . 135

Figure 20 Duration of the foot-internal unstressed syllables in relation to the speaker group 137

Figure 21 Duration of the foot-internal unstressed syllables in relation to the number of syllables in the foot 138

Figure 22 Distribution of the foot duration in relation to the number of syllables and inclusion of epenthetic vowels 142

Figure 23 The mean number of foot-internal syllables and the proficiency score 144

Figure 24 The variance of the foot duration and the mean number of syllables.................. 144

Figure 25 Variance of the foot duration and the degree of the stress contrast (the JP Adv group) 149

Figure 26 The number of epenthetic vowels and CV of the foot duration (the JP Beg group) 151

Figure 27 Variance of the foot duration and the degree of the stress contrast (the JP Beg group) 152

Figure 28 Duration of stressed vowels in relation to the number of syllables in the foot 156

Figure 29 Duration of stressed vowels in relation to the number of syllables in the foot 157

Figure 30 Variance of the foot duration and the number of epenthetic vowels in relation to the stress contrast (the JP Beg group) 161

1. Introduction

1.1. Rhythms of languages

English is a stress-accent language (Beckman 1986, Abercrombie 1967), which regulates its rhythm so that lexical stress repeats with equal intervals. The repetitive stress is metaphorically referred to as *rhythmic beats.* The locus of the beats is a prosodic unit called *foot*, which can be roughly defined to consist of one stressed and zero or more unstressed syllables (Section 2.2.1). That is, the English rhythm can be described either as the repetition of lexical stress (the rhythmic beats) or that of feet (the locus of the beats). The English rhythm is said to have mental isochrony, i.e. native speakers hear the rhythmic beats to repeat with roughly equal intervals or the feet to have roughly equal durations. Studies were conducted to show physical realisations of the isochrony, especially in the 1970s by Lehiste (e.g. Lehiste 1972, 1973, 1974, 1975, 1977). However, those studies failed to find complete isochrony in English speech, and now, there is an agreement among scholars that isochrony in the English rhythm is basically mental rather than physical (Section 2.2.2).

In contrast to stress-timed languages like English, whose rhythm is regulated by lexical stress, the rhythm of syllable-timed languages (such as French) maintains almost equal durations of syllables regardless of lexical stress or accent. Japanese is a mora-timed language (Port 1987) and, like syllable-timed languages, maintains the equal durations of mora intervals. Although syllable and mora could behave differently in prosodic phenomena, they are quite similar in terms of speech rhythm in that they are the smallest rhythmic unit in the respective languages.

Since Abercrombie (1976) distinguished between stress-timed and syllable-timed languages, studies have compared the two rhythmic systems (e.g. Beckman 1992, Dauer 1983, Roach 1982, Nespor et al. 2011, Warner & Arai 2001b). However, much less is evident about the difference in their speech timings compared to their structural differences (e.g. syllable structures and vowel reductions) (Dauer 1983). Compared to stress-timed and syllable-timed languages, not many studies have been conducted on mora-timed languages or their difference from the former two language groups.

In order to investigate crosslinguistic differences of speech rhythm, Ramus et al. (1999) proposed ∆V (standard deviation of vowel durations), ∆C (that of consonant durations) and %V (the proportion of vowel intervals). Those measures, together with ones normalised for speech rate, i.e. VarcoV (coefficient of variations[1] of vowel durations) and VarcoC (that of consonant durations) (Grabe & Low 2002), have successfully described crosslinguistic differences of speech rhythm. For example, English has greater ∆C and VarcoC than Japanese since it has complex consonant clusters (e.g. the sequence of three consonants /str/ in the word *strike* /straɪk/). English also has greater ∆V and VarcoV than Japanese because the vowel duration of English is extended when there is lexical stress on the syllable. In contrast, the vowel duration of Japanese is not affected by the presence or absence of lexical pitch accent. Hence, these measures have captured the differences between stress-timed and syllable-timed languages fairly successfully.

However, these measures are not so effective in describing English foot rhythm. Consisting of multiple syllables, the English foot is a larger unit than vowel and consonant intervals as measured by VarcoV and VarcoC. In addition, those measures, which describe the variability rather than the stability of the rhythmic units, are not so relevant in investigating isochrony in English foot, which is said to exist at least in native speakers' intuitions.

[1] Coefficient of variation is calculated by dividing the standard deviation by the mean.

1.2. Acquisitions of English rhythm by native speakers of Japanese

Compared to crosslinguistic comparisons of prosody in English and Japanese, there are much fewer studies on the acquisition of English prosody by native speakers of Japanese. The few studies include those on lexical stress (Konishi & Kondo 2015, Lee et al. 2006), intonation (Graham & Post 2018) and focus prosody (Ueyama & Jun 1996). Acquisition of English lexical stress is a well-investigated area in second language English acquisition and studies have also been conducted on native speakers of many other languages (e.g. Archibald 1997, Guion et al. 2004, Lee et al. 2006, Zhang et al. 2007). Although the accurate manifestation of lexical stress is quite important for L2 English speech to sound comprehensible and nativelike, its phonetic realisation has not been investigated in relation to the speech rhythm (Section 2.2.1).

A recent trend of the studies on the second language (L2) English rhythm is the use of measures of vowel and consonant intervals proposed by Ramus et al. (1999) and Grabe and Low (2002) such as ∆V, ∆C, VarcoV and Varco C (e.g. Grenon & White 2008, Kawase et al. 2016, Ozaki et al. 2017). Overall, the results suggest Japanese learners' English exhibits a smaller variance of both vowel and consonant intervals compared to that of native speakers. This is assumed to be a negative transfer of the first language (L1) Japanese rhythm since it has a much smaller variance than English. As learners become more proficient, their English exhibits a greater variance of consonant and vowel intervals, which approach that of native English speakers. Although the results are important in capturing the nature of Japanese English rhythm, they are not directly relevant to English foot rhythm, especially in terms of its isochrony.

Although there have been a few studies on the foot rhythm in Japanese English, none of them has directly examined its isochrony (Section 7.1). Mochizuki-Sudo and Kiritani (1991) investigated both production and perception of interstress intervals of English, which are roughly identical to feet, by native speakers and Japanese learners of English. Although the results showed the foot rhythm of native speakers was more isochronous compared to that of Japanese learners, they did not examine the statistical significance of the difference. Anderson (1993) examined acquisitions interstress intervals by Chinese, Korean and Japanese learners of English but focused on the means rather than variances of foot durations. Like Mochizuki-Sudo and Kiritani (1991), Mori et al. (2014) investigated durations of interstress intervals of English by native speakers and Japanese learners of English. However, they looked at only one sentence and examined the statistical difference between the interstress intervals rather than the variance of those intervals. Hence, none of the previous studies has clearly investigated the acquisition of isochronous rhythm in Japanese English speech.

1.3. Purpose of the current study

The main goal of the current study is to investigate the foot rhythm in English produced by native speakers of Japanese, with a special focus on the manifestation of isochrony. One of the primary motivations is the paucity of studies conducted on English foot rhythm, not to mention its acquisition by Japanese learners. Previous studies have almost failed to show physical isochrony of English foot rhythm. However, results have also been reported which are supportive of the existence of its mental isochrony (e.g. Lehiste 1975, 1977). Comparing native manifestations of English foot rhythm with those of L2 English foot, which are estimated to be less isochronous, could give important implications for the existence of physical isochrony in L1 English foot rhythm.

Among various L2 English, Japanese English is perhaps one of the most ideal in investigating the developmental change of its foot rhythm, especially in terms of its isochrony. This is because Japanese has a very different rhythmic system from English. Firstly, the mora-timed rhythm (Section 2.3.1) of Japanese is expected to have a certain negative transfer on the realisation of English foot rhythm, which is stress-timed. In contrast to mora-timing, in which one mora (which is roughly identical to one vowel) counts as one rhythmic unit, English foot by definition has no fixed number of syllables (i.e. no fixed number of vowels). That is, English feet count as one rhythmic unit whether they consist of only one stressed syllable or following unstressed syllables as well as one stressed syllable. In English speech by Japanese learners of low English proficiency, the rhythmic units are estimated to be equivalent to vowels (or syllables) rather than feet (or stressed syllables). The different phonotactics in the two languages further add to the influence of Japanese rhythm on L2 English. Since Japanese has a much simpler syllable structure than English, which has consonant clusters in both onset and coda positions, vowel epenthesis is observed in Japanese English speech (Chapter 6). Those epenthesised 'extra' vowels would further increase the duration of the relevant feet, adding further to their variance.

Another distinctive aspect of the current study is its evaluation of the speech proficiency of the learners. While prosody has considerable effects on the speech intelligibility in L2 English (e.g. Anderson-Hsieh et al. 1992, Munro & Derwing, 1999), its acquisition has not so much been investigated in relation to speech proficiency. Most proficiency scores used in the investigations of L2 English prosody were objective measures of general proficiency levels including test scores, durations of learning, length of residence and ages of arrival in English-speaking countries. These reflect not only pronunciation but also other factors such as grammar and vocabulary. In addition, the speaking sections of English proficiency tests such as TOEFL and IELTS usually evaluate spontaneous speech, including its grammar and vocabulary. To avoid such confounding influences, the current study used human-rated scores of the read speech in determining the proficiency of learners.

1.4. Outline of the following chapters

In order to provide a background of the current study and a solid basis for the interpretation of the results of the analyses, Chapter 2 will provide a comprehensive summary of previous studies on the acquisitions of English rhythms by native speakers of Japanese. It will first summarise English (Section 2.2) and Japanese rhythms (Section 2.3). Then, Section 2.4 will summarise previous studies on Japanese English rhythm, focusing on lexical stress, vowel epenthesis and foot rhythm, which will be examined in the analyses in Chapters 5-7.

Chapter 3 and Chapter 4 will outline the data of the speech corpus of Japanese English used for the analyses in Chapters 5-7. Chapter 3 will introduce subjects of the recordings of the corpus (Section 3.2), the recoding method (Section 3.3) and the speech data for the phonetic analyses in the current study (Section 3.4). Section 3.5 will detail the methods and criteria of annotation in the corpus, especially ones relevant to the analyses in the current study.

Chapter 4 will discuss the methodology of the human ratings of the speech data. After explanations about the raters in the current study (Section 4.2) and the procedures of the rating (Section 4.3), inter-rater reliability was statistically examined (Section 4.4). The analyses in the chapter will especially focus on the consistency (or inconsistency) of the human ratings, i.e. the perceived proficiency, of the Japanese English speech in relation to raters' L1s, including native speakers. By so doing, it will determine whether it is appropriate to use the rated scores of nonnative English speakers as well as those of native speakers. In addition, the chapter will investigate relationships of the scores (Section 4.5) and consider the most appropriate scores to be used for the remaining analyses in the dissertation (Section 4.6).

Using the data in Chapter 3 and Chapter 4, Chapters 5 will investigate manifestations of lexical stress, i.e. the phonetic realisation of English foot, in Japanese English. The analyses will be conducted in relation to the learners' proficiency. After the summary of previous studies (Section 5.1), the duration and intensity of vowels with and without lexical stress will be statistically compared in relation to learners' proficiency. For both duration and intensity, Section 5.2 will detail the data coding and normalisation methods, Section 5.3 will report the results of the phonetic analyses and Section 5.4 will discuss the results especially in relation to those of previous studies.

Chapter 6 will examine epenthetic vowels in Japanese English, which are expected to hinder the proper manifestation of Japanese English foot rhythm. After the explanation of the data coding and normalisation methods (Section 6.2), the chapter will investigate the frequency of vowel epenthesis (Section 6.3.1) and phonetic realisations (i.e. the duration and intensity) of actually epenthesised vowels (Sections 6.3.2 and 6.3.3). Again, both phonetic factors will be investigated in relation to speakers' proficiency. Section 6.4 will discuss the results in relation to those of previous studies.

Chapter 7 will investigate the foot rhythm in Japanese English in relation to speakers' proficiency. After the summary of previous studies (Section 7.1), Section 7.2 will detail how to define foot in the data for the analyses, as well as normalisations of the foot durations. Section 7.3.1 will compare the variance of foot durations. Following that, Section 7.3.2 will examine foot durations in relation to the number of foot-internal syllables. The next sections will investigate a well-known phenomenon to regulate the foot rhythm in L1 English, compensatory shortening. It will be investigated for foot-internal stressed syllables (Section 7.3.3) and unstressed syllables (Section 7.3.4). Lastly, Section 7.3.5 will investigate phonetic correlates of foot rhythm in Japanese English. Based on the results of the other sections, Section 7.4 will discuss isochrony in native English and Japanese English.

Chapter 8 will provide general discussions based on the results of the analyses in Chapters 5-7. Section 8.1 will summarise the results of proficiency ratings. Section 8.2 will discuss isochrony in both native-speaker English and Japanese English. Section 8.3 will discuss implications for teaching English rhythm to nonnative speakers.

Lastly, Table 1 is the list of abbreviations for the frequently used terms in the current thesis. Hereafter throughout the dissertation, the abbreviations will be in italic font.

Table 1 Abbreviations used in this dissertation

abbreviation	**Referents**
NE	Native-speaker English (L1 English)
NNE	Nonnative English (L2 English)
JE	Japanese(-accented) English
Adv/Beg JE	English spoken by native speakers of Japanese with advanced/beginner proficiency level
EN	Native English speakers
JP	Native Japanese speakers Japanese learners of English
Adv/Beg JP	Japanese advanced/beginner learners of English

2. Acquisitions of English Rhythm by Native Japanese Speakers

2.1. Introduction

This dissertation investigates the foot rhythm of *JE*, focusing especially on the manifestation of its isochrony. English and Japanese have different rhythmic systems. English is a stress-accent language (Beckman 1986, Abercrombie 1967), in which lexical stress (i.e. accent which “uses to a greater extent material other than pitch”; Beckman 1986, p 1) is said to repeat with roughly equal intervals. The locus of stress is foot, which consists of a stressed syllable and some unstressed syllables (See Section 2.2.1 for details). Since lexical stress repeats with roughly equal intervals, durations of feet, each of which could consist of one or more syllables (i.e. one or more vowels), are also regulated to be identical. That is, the number of the vowels, by definition, does not influence the duration of each foot. In contrast, Japanese is a mora-timed language (Port 1987), in which morae are regulated to have equal intervals. Since mora is a unit consisting of one vowel, durations of the utterances are roughly proportional to the number of vowels. Therefore, a certain transfer of Japanese mora rhythm is expected in the *JE* rhythm.

The current chapter will discuss previous studies on the *JE* rhythm. Section 2.2 will summarise previous studies on the *NE* rhythm. Section 2.2.1 will discuss phonological issues regarding definitions of English foot. Section 2.2.2 will summarise previously conducted acoustic-phonetic studies on isochrony of the *NE* foot rhythm. Section 2.2.3 will examine lexical stress, which manifests the rhythmic beats of the English foot. In Section 2.2.4, the phonetic realisations of the *NE* lexical stress will be discussed, focusing on the results of the previous studies. Section 2.3 will discuss the Japanese rhythm, which is expected to influence the *JE* rhythm. Section 2.3.1 will summarise previously conducted phonological studies on the basic rhythmic units of Japanese, especially focusing on mora. Section 2.3.2 will discuss the isochrony in the Japanese rhythm. Following these, Section 2.4 will discuss previous studies on the *JE* rhythm focusing on the estimated phonetic correlates of *JE* foot rhythm. Section 2.4.1 will summarise the results of previous studies on phonetic realisations of lexical stress in *JE*. Section 2.4.2 will focus on vowel epenthesis, which could be an obstacle in manifesting isochronous foot rhythm in *JE* because they would increase the duration of feet with epenthetic vowels but not those without them. Lastly, Section 2.4.3 will discuss *JE* foot rhythm in relation to previous studies.

2.2. English rhythm

2.2.1. Foot rhythm in English

As in many other languages, foot is the fundamental unit of the speech rhythm in English (Abercrombie 1967, Pike 1945). A common metaphor to describe the foot-based accent is that it is the *rhythmic beats* that repeat periodically; each foot bears the beat. Although it is not fully clear from the previous studies whether each beat is realised with lexical stress (“a phonologically delimitable type of accent in which the pitch shape of the accentual pattern cannot be specified in the lexicon”; Beckman 1986, p 1) or pitch-accent (“a system of syntagmatic contrasts used to construct prosodic patterns”; ibid.), the majority of them (e.g. Abercrombie 1967, Cruttenden 2014, Féry 2018) support the position that foot is realised with lexical stress rather than pitch accent. Cruttenden (2014) argues that the assumption that pitch-accented syllables bear “rhythmical stress” is “exceedingly counter-intuitive” (p 272). According to Féry (2018), all prosodic domains above foot are “prominence based” (p 47), which implies pitch accent is not a property of foot rhythm.

In metrical phonology, foot (φ) is a minimal constituent of phonological words (ω). A foot constitutes multiple syllables (σ) and every foot has one (and only one) stressed syllable. Since a stressed syllable is minimally bimoraic in English (Féry 2018), a foot consists of at least two morae. A phonological word (ω) may consist of multiple feet; there, a phonological word may bear two or more stressed syllables, one of which bears primary stress. Figure 1 is the schematic representation of the prosodic structure of the word *intonation*. σ denotes a syllable, φ a foot, and ω a phonological word. The stress on the second foot *na-tion* is primary stress while one on the first foot *in-to* is not. Alternatively, monosyllabic feet, as well as monosyllabic phonological words, are also allowed in English. In fact, there are many monosyllabic lexical words in English (e.g. *cool, keep, pen, tea*). Function words cannot form a prosodic word by themselves. Instead, they attach to an adjacent stressed syllable to form a prosodic word.

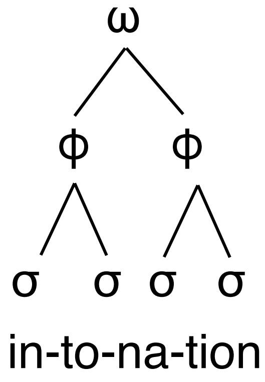


Figure 1 English foot in its prosodic hierarchy

Previous studies suggest trochaic foot (with the strong-weak rhythm) is dominant in English over iambic one (with the weak-strong rhythm). Most disyllabic words have a trochaic stress pattern (Féry 2018). The exceptions are verbs and adjectives, which usually bear stress on their final syllables, as well as loan words (e.g. *bidet*). Cutler and Carter (1987) investigated the rhythmic patterns of an English dictionary with more than 33,000 entries and found that 73 % of the words had either primary or secondary stress on their initial syllables. The result was replicated by Cutler (1989), who examined a corpus of 13,000 most common words of British English and found that the initial syllables of over 70 % of lexical words bore either primary or secondary stress. Furthermore, many of the four-syllable words and compounds had two trochaic feet (e.g. *dictionary, kindergarten*, *television*). Accordingly, English is referred to as "trochaic language" (Féry 2018, p 47), which prefers the trochaic rhythm over the iambic one.

It has been argued that the preference on trochee in English foot rhythm helps identification of word and syntactic boundaries. Taft (1984) conducted a perceptual experiment in which subjects were asked to disambiguate pairs of one-word vs two-word sequences, e.g. "lettuce" vs "let us." The result showed native speakers' preference of one-word interpretations with the trochaic stress pattern (e.g. "lettuce" was preferred over "let us") and two-word interpretations with the iambic stress patterns (e.g. "in vest" was preferred over “invest”). According to the rhythmic segmentation hypothesis by Cutler and Carter (1987), the first phase of English speech segmentation is conducted on the assumption that each strong syllable signals the beginning of a lexical word. Tajima (1998) expresses this as a “statistical preference for word-initial stress in the English lexical system” (p16). The hypothesis was empirically supported by Cutler and Butterfield (1991).

Strictly speaking, a trochaic foot is defined to be disyllabic, consisting of a stressed syllable (its head) and a following unstressed syllable (its tail). The strict definition of English foot is also disyllabic (Gussenhoven 2004) and some of the previous studies did not accept English foot consisting of only one syllable (e.g. Burzio 1994) or more than two syllables (e.g. Hayes 1995). A common way to treat non-disyllabic rhythm based on the binary foot structure is the introduction of the notion of *degenerate feet*, which have no stressed syllable and adjoin the preceding or the following foot at the level of phonological word (Féry 2018). Figure 2 shows examples of a degenerate foot in a word with the dactyl (i.e. strong-weak-weak) rhythm. σ denotes a syllable, φ a foot, and ω a phonological word. The third syllable (σ) of the word *interest* has no other syllable with which to form a binary foot (φ) because the first two syllables form a trochaic foot. Therefore, it forms a degenerate foot by itself and attaches to the preceding foot at the level of phonological word (ω). Another example of degenerate feet is the first syllables of trisyllabic words with no word-initial stress. In the word *succession* (Figure 2), the final two syllables form a trochaic foot while the first syllable forms a degenerate foot. According to Féry (2018), all these varying examples can be assumed to have trochaic structures (i.e. a trochaic foot preceded or followed by degenerate feet).

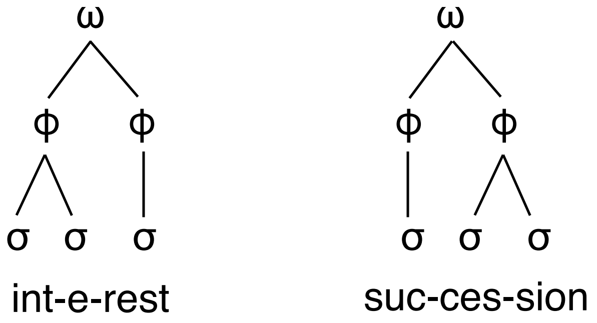


Figure 2 Examples of degenerate feet in English

In other studies (e.g. Erickson et al. 2012, Mott 2011), English foot has been defined to contain more than one unstressed syllable following a stressed syllable. Actually, the dactyl rhythm is very common in English at the level of lexical word (e.g. *beautiful, interest, pendulum*). The current study adopts a broader definition of foot, in which a foot can contain more than two syllables because, with an assumption of degenerate feet, it would be difficult to explain the mental isochrony in English foot rhythm (i.e. each foot manifests equal duration; See 2.2.2). However, sequences of unstressed syllables without preceding or following stressed syllables within the prosodic boundary will be treated as degenerate feet. With this rule, a phrase *American dream* will be parsed as the sequence of a degenerate foot /ə/, followed by a trisyllabic foot /mer.ɪ.kən/ and a monosyllabic foot /driːm/. Figure 3 is a schematic representation of its prosodic structure where σ denotes a syllable, φ a foot and ω a phonological word. The first foot is a degenerate foot since it has no preceding stressed syllable to adjoin one while the last one is not a degenerate foot since it can attach to the preceding foot *mer-i-can* to form a four-syllable foot.

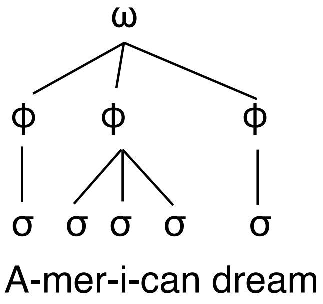


Figure 3 The structure of English foot adopted in the current study

With all the above assumptions, English feet are also referred to as *interstress intervals* (e.g. Erickson et al. 2012, Lehiste 1975, 1977, Mochizuki-Sudo & Kiritani 1991, Mori et al. 2014). The term can be defined as the units of speech delimited by lexical stress (i.e. between the onset of a stressed syllable and the onset of the following stressed syllable (Anderson 1993). Because lexical stress, rather than pitch accent, manifests a foot, the number of feet can also be measured with the number of stressed syllables. The dominance of the trochaic rhythm in English suggests lexical stress marks the beginning of a new foot, or an interstress interval. Furthermore, adopting a broader definition of foot, according to which a foot can contain more than two syllables, most unstressed syllables, which would otherwise be treated as degenerate feet, would be part of a foot containing the preceding stressed syllable. Hence, a foot would be mostly identical to the interval demarcated by two adjacent stressed syllables, i.e. *interstress* intervals.

2.2.2. Isochrony in English foot rhythm

Although English is a stress accent language (See Section 2.1), the isochrony of the English foot rhythm has not been empirically supported. No previous study was successful in demonstrating the isochrony of two interstress intervals (Cruttenden 2014). According to Tajima (1998), even in laboratory settings, isochrony in speech is hindered by many effects including syllable structures and so-called final lengthening.

While the results of the previous studies suggest the absence of physical isochrony in English feet, the results of other studies suggest native speakers' attempts to manifest isochronous foot rhythm. One of the most typical examples is so-called *compensatory shortening* observed within each foot, i.e. the more syllables there are in a foot, the more each syllable will be shortened. The shortening of unstressed vowels was also observed in (Ikoma 1993, 1998), which is also a stress-timed language (Abercrombie 1967).

In English, the negative correlation between the duration and the number of foot-internal syllables has been observed on stressed syllables rather than unstressed ones (Fowler 1977, 1981, Huggins 1973, Lehiste 1972, Rakerd et al. 1987). Lehiste (1972) compared the rhythms of monosyllabic base words and their inflected counterparts, which were disyllabic or trisyllabic, uttered by two native speakers. The result showed that the duration of stressed syllables of the base part of the inflected words was compressed compared to when the base part was uttered as an independent word. Huggins (1973) prepared a sentence with three stressed syllables *Cheese bound out* and added unstressed syllables between the stressed syllables (e.g. *Cheese(s) (a)bound(ed) (ab)out*). The recorded speech of two native speakers showed shortening of stressed syllables were observed when followed by an unstressed syllable across a word boundary. However, the shortening was not observed when the unstressed syllable was in the same word, perhaps because of the effect of final lengthening which increased the duration of the unstressed syllable. Rakerd et al. (1987) compared the shortening in the two conditions, i.e. when the following unstressed syllable was within the same word or across a syntactic boundary. As a result, the shortening was observed regardless of the syntactic boundary. Hence, Rakerd et al. (1987) explicitly state that "'[f]oot-level shortening' refers to shortening of stressed syllables by unstressed syllables in the same metrical foot" (p 147).

The attempt to aim for isochronous foot rhythm was also observed in various phonological phenomena of English. An example is the so-called *stress shift*. When there are two adjacent syllables with primary stress (e.g. thirteen /ˌθɜːˈtiːn/ people /ˈpiːpəl/), the stress on either of the two syllables will shift onto the syllable with canonical secondary stress (i.e. /ˈθɜːˌtiːn ˈpiːpəl/) within the same prosodic phrase. This is expected to regulate durations of interstress intervals and is also observed in many other languages including German, Dutch and Italian (Féry 2018). However, it does not happen if an iambic word has no other stressed syllable (e.g. *taboo subject* /təˈbuː ˈsʌbdʒɪkt/ but not */ˈtəbuː ˈsʌbdʒɪkt/). A similar phenomenon that is observed especially in rapid speech is one of the two successive syllables that are both canonically stressed will lose its stress (Mott 2011). Another example would be so-called *schwa deletion* (The word *probably* /ˈprɒbəbli / could be pronounced /ˈprɒbli/), which changes the syllable structures but keeps the number of feet. The fact that more or less equal timing is maintained with different numbers of syllables further supports the stability of foot as the basic rhythmic unit. According to Lehiste (1977), isochronous rhythm is integrated into English grammar at the syntactic level, and a deviation from isochronous rhythm, i.e. an interstress interval large enough to be perceived, could signal a syntactic boundary.

Results of previous studies also imply native listeners' psychological preference for isochronous foot rhythm. Lehiste (1975) demonstrated that the actual differences in the physical durations of interstress intervals could be below the perceptual levels of listeners. In the experiment, she presented four non-speech stimuli, three of which had identical durations of 300, 400 or 500 milliseconds (ms) and the other had either shorter or longer duration with nine 10 ms steps. Durational difference smaller than 30 ms was not identified reliably. According to Lehiste (1977), this was about 10% of the mean foot duration and was the threshold to perceive isochrony. This psychological preference to hear isochrony might be only relevant to speech sounds. Lehiste (1977) generated some combinations of four sounds differing only in duration and asked the subjects to judge the shortest and longest ones. The result that subjects' performance was better with non-speech sounds than speech sounds implies that psychological preference to 'hear isochrony' is specific to speech sounds.

2.2.3. Lexical stress as rhythmic beat in English foot rhythm

Among multiple functions of English lexical stress such as identification of syntactic categories (most nouns have the trochaic stress pattern while not many verbs do), one of the most important functions is its effects on speech rhythm. According to Cruttenden (2014), English rhythm is governed by equal timing of interstress intervals. As discussed in the previous sections, the intervals are manifested by the so-called rhythmic beats (See Section 2.2.1). Its locus is foot and its physical realisation is lexical stress. Lexical stress is on the rightmost or leftmost syllable of a foot (Gussenhoven 2004). The majority of them are on the leftmost syllable because of the dominance of the trochaic foot rhythm in English.

Some stressed syllables, usually the last one in the intonational phrase, also bear a pitch accent. Vowels with a pitch accent have notably higher or lower fundamental frequency (F0), or a considerable change in the F0, than vowels without a pitch accent. In the Tone and Break Indices (ToBI) notation, pitch events including pitch accent, phrasal tones and boundary tones, are categorised as intonational events and distinguished from rhythmic ones. According to Beckman (1986), lexical stress is syntagmatic while tone events are paradigmatic. Pitch accent is paradigmatic since there are several pitch shapes to imply different pragmatic meanings. In contrast, lexical stress, the locus of pitch accent, is related to syntax, as well as the foot rhythm realised with lexical stress.

In English, syllable weight, i.e. the number of morae in the rime, correlates with stress placement. A syllable that bears lexical stress must be minimally bimoraic, i.e. it has a long vowel or a coda. If the syllable is monomoraic, it would become bimoraic by lengthening (e.g. the first syllable in *decrease* is monomoraic /dɪ/ with the iambic accent but bimoraic /diː/ with the trochaic accent) or ambisyllabicity (e.g. /p/ in *taping* /teɪpɪŋ/ is not ambisyllabic because the first syllable /teɪ/ is bimoraic but /p/ in *tapping* /tæpɪŋ/ is ambisyllabic because the first syllable /tæp/ would be monomoraic without it) (Féry 2018). The two rules are respectively called "Stress-to-Weight" and "Weight-to-Stress." Both of them are observed in many other languages including Italian and Bantu (Gussenhoven 2004, p 16).

One of the main characteristics of English lexical stress is that the quality of unstressed vowels tends to be reduced. They are often realised as a schwa /ə/. Because of the centralisation of the vowel quality in lexically unstressed positions, there is a smaller variety of vowels than in the stressed positions (Gussenhoven 2004). In addition to these phonological reductions, some reductions are spontaneous, which are typically on short function words and pronouns (Shockey 2003). A schwa is sometimes deleted (i.e. *schwa deletion*) or produced as a consonant of a similar place of articulation. According to Gussenhoven (2004), an unstressed high vowel preceding a stressed one is optionally produced as a glide; an example is the first vowel of the word *piano*, which could surface either as [i] or [j]. Mott (2011) argues English unstressed vowels undergo a much more drastic reduction compared to the weakening of vowels in syllable-timed languages such as French and Catalan.

Among the various factors associated with vowel reduction in English, two important correlates are speech rate and register. That is, there are more frequent reductions in fast and/or sloppy rather than slow and/or careful speeches. However, Shockey (2003) argues against the effect of formality. She proposes instead that spontaneous speech elicits more reductions; less frequent reductions are observed in speeches in formal settings because most of them are scripted. Reductions are also assumed to be associated with word frequency, an empirical data of which is the corpus-based analysis by Greenberg and Fosler-Lussier (2000). Other relevant factors include phonetic environments (e.g. preceding and following segments), timing and discourse structure (e.g. words with old information tend to have more frequent reduction than those with new information) (Shockey 2003).

Some previous studies suggest that vowel quality (i.e. whether it is a full vowel or not) is a more promising correlate of English rhythm than stress accent. Cruttenden (2014) claims that identifying lexical stress is often difficult on syllables with no pitch accent. The Borrowing Rule (Bolinger 1981) states syllables with reduced vowels *borrow* time from the preceding full-vowelled syllable. The implication is that syllables with a full vowel become shorter preceding those with reduced vowels. Therefore, durational regulation is not within each foot but between syllables with full vowels and those with reduced ones.

Cruttenden (2014) supports the vowel-based rhythm with an example sentence *Sparrows aren't common* (p 272). Table 2 shows his analyses of the sentence rhythm based either on vowel qualities (vowel-based) or lexical stress (stress-based). In the stress-based analyses, *S* denotes a stressed syllable and (*U)* denotes an unstressed syllable. In the vowel-based analysis, *F* denotes a full-vowelled syllable and *(R)* denotes a reduced-vowelled syllable. The main point of Cruttenden's (2014) argument is that the stress-based analyses are counter-intuitive here. When the copula *aren't* is in the weak form /ənt/[2] (i.e. *Stress-based analysis 1*), two syllables bear stress, i.e. /spa/ of *sparrows* and /kɒm/ of *common*. He argues the first three syllables /spæ.rəʊz ənt/ (a full vowel /æ/

[2] Since the word is in the weak form, the quality of its vowel is reduced to /ə/.

and two reduced vowels /ə/) are intuitively much longer than the last two /kɒm.ən/ (a full vowel /ɒ/ and a reduced vowel /ə/). When the copula is in the strong form /ɑːnt/ (i.e. *Stress-based analysis 2*), three syllables bear stress, i.e. /spa/, /ɑːnt/ and /kɒm/. However, according to Cruttenden, the first foot /spæ.rəʊz/ is still longer than the other two, /ɑːnt/ and /kɒm.ən/. In contrast, with the *Vowel-based analysis*, each full-vowelled syllable (plus a following reduced-vowelled syllable) has equal duration so that there are four rhythmic beats with equal durations, i.e. /spæ/, /rəʊz/, /ɑːnt/ and /kɒm.ən/. At least with this example, the vowel-based analysis seems more intuitive.

Table 2 Example analyses of English rhythm by Cruttenden (2014, p 272)

S: stressed syllable (U): unstressed syllable

F: full-vowelled syllable (R): reduced-vowelled syllable

Stress-based analysis 1	spa.rəʊz ənt kɒm.ən S (U) (U) S (U)
Stress-based analysis 2	spa.rəʊz ɑːnt kɒm.ən S (U) S S (U)
Vowel-based analysis	spa.rəʊz ɑːnt kɒm.ən F F F F (R)

However, there are also counterexamples to the effectiveness of vowel-based analysis. An example is the phrase *English speaking people* (Table 3). With the vowel-based analysis, the duration of the word *English* (with two rhythmic units, i.e. full-vowelled syllables, /ɪŋ/ and /ɡlɪʃ/) and *speaking* (with two rhythmic units /spiːk/ and /ɪŋ/) are intuitively much longer than *people* (with one rhythmic unit, a full-vowelled syllable /piː/, followed by a reduced-vowelled one /pəl/). In contrast, the stress-based analysis treats the duration of the three words identically. Hence, each word has one rhythmic unit. Here, the stress-based analysis seems to be better at describing native speaker intuition.

Table 3 A counterexample to the vowel-based analysis by Cruttenden (2014)

S: stressed syllable (U): unstressed syllable

F: full-vowelled syllable (R): reduced-vowelled syllable

Vowel-based analysis	ɪŋ.ɡlɪʃ spiːk.ɪŋ piː.pəl F F F F F (R)
Stress-based analysis	ɪŋ.ɡlɪʃ spiːk.ɪŋ piː.pəl S (U) S (U) S (U)

The current study follows the stress-based contrast. The primary motivation is that previous studies (Huggins 1973, Lehiste 1972, Rakerd et al. 1987) investigated the compensatory shortening with the assumption of stress-based, rather than vowel-based, rhythm. While some studies (Bolinger 1981, Cruttenden 2014) support the effectiveness of the vowel-based rhythm, there are also counterexamples.

2.2.4. Phonetic correlates of English lexical stress

In examining the results of the previous studies on the phonetic correlates of English lexical stress, this dissertation follows the latest theory of metrical phonology (e.g. Beckman & Edwards 1994, de Jong et al. 1993, de Jong 2004, Gussenhoven, 2004) that assumes the binary feature of lexical stress (either stressed or not). Based on the theory, the difference between primary stress and secondary stress is whether a pitch accent is associated with the stressed syllable in its phonetic realisation. As shown in Table 4, syllables with primary stress have both lexical stress and pitch accent while those with secondary stress have lexical stress but no pitch accent. With this, the so-called *stress shift* (cf. Section 2.2.2) can be interpreted to reallocate pitch accents (i.e. a post-lexical prosodic feature) but keeps the locations of lexical stress. According to Beckman (1986), the shape of pitch accent in stress-accent languages is specified post-lexically. The relevance of the three-way distinction in analysing prosody was also empirically supported for English by Plag et al. (2011) and for Dutch by Sluijter and van Heuven (1996b). The results of the analyses by Plag et al. (2011) suggest lexically defined primary and secondary stressed syllables are not phonetically different unless the word is pitch-accented.

Table 4 Three-way distinction of English stress in the latest theory of metrical phonology (Beckman & Edwards, 1994; de Jong, 2004; Gussenhoven, 2004)

Lexical stress	**Pitch accent**	
+	+	Primary stress
+	-	Secondary stress
-	N.A.	No stress

There are some language universal differences between stressed and unstressed syllables. These include a) longer durations of stressed vowels compared to unstressed ones, b) greater spectral balance, i.e. even distribution of energy across frequency, of stressed vowels compared to unstressed ones and c) a more centralised vowel quality of unstressed vowels (which is closer to schwa) compared to stressed ones (Gussenhoven 2004). These are due to greater articulatory effort in producing stressed vowels and it would be highly unlikely to find a language that realises stress "in phonetically conflicting ways" (Gussenhoven 2004, p 15). According to Beckman (1986), it is perhaps the durational pattern that ensures the identification of accent in stress-accent languages.

Despite the abundance of previously conducted acoustic-phonetic studies on English lexical stress, there have been no unanimous results as to its actual phonetic correlates. The inconsistency is perhaps because most studies did not consider the effects of pitch accent (or prominence). According to Volín and Zimmermann (2011), many acoustic-phonetic studies on English lexical stress, including Fry (1955, 1958), had target words in pitch accented positions. As claimed by Gussenhoven (2004), those studies were ignoring the effect of other prosodic phenomena, most notably change of F0 due to pitch accent. Particularly, when a word is uttered in isolation or embedded in a carrier sentence, its stressed syllables typically bear a pitch accent. Sluijter and van Heuven (1996b) also points out this “covariation of accent and stress” (p 2473). According to Plag et al. (2011), those studies “have suffered from... the fact that they only looked at accented words” (p 363).

The extracted values in such experiments, especially those of F0, are supposed to have been the effect of overwritten pitch accent. Reviewing past studies (e.g. Vanderslice and Ledefoged, 1972; Huss, 1978; Pierrehumbert, 1980; Beckman and Edwards, 1994), Sluijter and van Heuven (1996b) claims “[p]itch movement is the correlate of accent, rather than of lexical stress” (p. 2471). Gussenhoven (2004) more clearly discounts the relevance of F0 to lexical stress, stating it is a “misconception” to assume lexical stress is realised with F0 (p 13). Ladd (2008) also supports this position, arguing that word-internal tones, as well as word-edge tones, have post-lexical functions in English.

Table 5 is the summary of previously conducted acoustic-phonetic studies on English lexical stress in pitch-accented position. The first column shows the study and the second column shows whether it is a production or perception study (or both). The first row shows the parameters investigated in each study. A tick “✓” shows that the factor is distinguished between primary stressed (i.e. stressed and pitch-accented) and unstressed vowels. Blank cells denote the factors that were not investigated in the respective study. Different studies adopted different measurements of values. Although most studies measured peak or mean intensity of vowels, some measured spectral tilt (or spectral balance), i.e. the difference of the energy distributions between the high-frequency and low-frequency spectrum (Sluijter and van Heuven 1996a). Rather than mean or peak F0, pitch slope (the highest F0 minus lowest F0 within each vowel; a relevant factor to identify pitch accents) was adopted in some studies. In addition, some studies measure vowels while others measure syllables. In order to see the big picture, Table 5, Table 6 and Table 7 did not distinguish between such differences.

Table 5 Previous studies on the comparison of English vowels with primary stress VS no stress

	Production or perception	Duration	Intensity / spectral tilt	F0 / pitch slope	Quality
Fry (1955)	Production	✓	✓		
Fear et al. (1995)	Production	✓	✓	Mean F0 only	✓
Sluijter & van Heuven (1996a)	Production	✓	✓	✓	✓
De Jong (2004)	Production	✓			✓
Okobi (2006)	Both	✓	✓		Only for some test words

The results of the previous studies indicate primary stressed vowels can be distinguished from unstressed vowels with intensity, duration, F0 and vowel quality (the first and second formants; hereafter F1 and F2). Duration seems to be the most stable parameter to detect stress in pitch accented positions. Its effect is statistically significant in all studies listed in Table 5. In the study conducted by Sluijter and van Heuven (1996b) on Dutch, duration appeared to be the strongest correlate of the distinction. Although there are intrinsic differences of syllable duration in relation to its position in a word, the production experiment by Okobi (2006) demonstrated stressed second syllables are consistently longer than the first syllable.

Most of the previous studies investigated intensity or spectral tilt and none of them reported the absence of its effect. The effect of F0 was observed in Fear et al. (1995) and Sluijter and van Heuven (1996a). However, the latter reported the absence of the effect of pitch slope. The effect of vowel quality was observed by Fear et al. (1995), Sluijter and van Heuven (1996a) and de Jong (2004) but not observed in all test words of Okobi (2006). Sluijter and van Heuven (1996a) also reported vowel quality is a weaker correlate in pitch-accented positions compared to intensity and F0 peak. Because of the confounding influence of lexical stress and pitch accent, it is not clear from Table 5 which factor is relevant to lexical stress, and which to pitch accent.

Table 6 summarises the results of the previous studies which investigated the factors to manifest the difference between primary and secondary stresses. Based on the theory of the latest metrical phonology (e.g. Beckman & Edwards 1994, de Jong et al. 1993, de Jong 2004, Gussenhoven, 2004), they can be interpreted as stressed syllables with and without accompanying pitch accent. As for Table 5, the first column shows the study and the second column shows whether it is a production or perception study (or both). The first row shows the parameters investigated in each study. A tick "✓" shows that the factor is distinguished between primary stressed (i.e. stressed and pitch-accented) and secondary stressed (i.e. stressed but not pitch-accented) vowels. Blank cells denote the factors that were not investigated in the respective study.

Table 6 Previous studies on the comparison of English vowels with primary VS secondary stress

	Production or perception	Duration	Intensity / spectral tilt	F0 / pitch slope	Quality
Fear et al. (1995)	Perception	✓	No	No	only in fast speech rate
Sluijter & van Heuven (1996a)	Production	✓	✓	✓	
Mattys (2000)	Perception	✓	No	✓	
De Jong (2004)	Production	✓			✓
Plag et al. (2011)	Production	No	Only in left prominent words	Only in left prominent words	

Again, duration seems a promising correlate of the distinction except in Plag et al. (2011). The other three factors show contradictory results of the previous studies. Regarding intensity, more studies are against its effect. Also, Plag et al. (2011), although concluding that spectral balance is the most reliable acoustic parameter to estimate accent, admit it is heavily dependent on the position of the syllable in the word. In the study by Sluijter and van Heuven (1996b) on Dutch, the overall intensity was not so much correlated to stress but to accent. The effect of F0 was observed in most studies but, surprisingly, not observed by Fear et al. (1995) and only partially observed by Plag et al. (2011). This also supports the view that pitch is not a constituent of lexical stress but post-lexical accents. This could be explained by the report by Kochanski et al. (2005) that the use of F0 in accentuation has some variation in British accents. The effect of vowel quality was supported by de Jong (2004), according to whom, vowels in pitch-accented positions are hyperarticulated. However, the effect was observed only in a fast speech in Fear et al. (1995).

Table 7 is the summary of previous studies that investigated the effect of factors on stress contrast in non-pitch-accented positions. Again, the first column shows the study and the second column shows whether it is a production or perception study (or both). The first row shows the parameters investigated in each study. As for the contrast made between primary stress and no stress, the effect of vowel duration was supported in all studies, suggesting it is the most robust correlate of English lexical stress. Furthermore, the linear discriminant analysis by Sluijter and van Heuven (1996a) implied vowel duration to be the strongest cue, followed by spectral tilt and vowel quality. The result is bolstered by both production and perception experiments by Okobi (2006), which indicated syllable duration was the strongest correlate and cue of lexical stress. Ladefoged (2011) also argues vowel duration is the primary perceptual cue to detect stressed syllables.

Table 7 Previous studies on the comparison of English vowels with secondary stress VS no stress (or between stressed VS unstressed vowels in non-pitch-accented positions)

	Production or perception	Duration	Intensity / spectral tilt	F0 / pitch slope	Quality
Fear et al. (1995)	Perception	✓	✓	Only mean	✓
Sluijter & van Heuven (1996a)	Production	✓	No	No	Weakest correlate
De Jong (2004)	Production	✓			✓
Okobi (2006)	Both	✓	No	No	

The use of intensity and F0 was supported by Fear et al. (1995) but not by Sluijter and van Heuven (1996a) and Okobi (2006). The study by Plag et al. (2011) investigated differences in phonetic realisations between primary stress and secondary stress as defined by dictionaries. They found that F0 affected some stress contrast when the words were pitch-accented but not when the words were not. That is, primary and secondary stresses as defined by dictionaries are not phonetically different unless the word bears a pitch accent. This further enhances the validity of the above interpretation of secondary stress that it is a combination of lexical stress and pitch accent. Okobi (2006) argues peak F0 and intensity are correlates of pitch accent rather than lexical stress. Although vowel quality was a reliable correlate in Fear et al. (1995) and de Jong (2004), the linear discriminant analysis by Sluijter and van Heuven (1996a) suggested it was the weakest correlate of lexical stress.

2.3. Japanese rhythm

2.3.1. Units of Japanese rhythm

Japanese is a mora-timed language (Port et al. 1987) and mora is the basic unit of its speech rhythm. This contrasts with English, which has a stress-timed rhythm (Beckman 1987, Abercrombie 1967). Japanese vowels have quantitative contrasts phonologically so that all five vowels (/a/, /e/, /i/, /o/, /ɯ/) have their longer counterparts (/aː/, /eː/, /iː/, /oː/, /ɯː/), which count as two morae. Although syllable is relevant in limited instances of morphology and phonology (e.g. the unit to bear lexical accent is syllable rather than mora; Kubozono 1998), it is not a rhythmic unit in Japanese. Syllables are open, rather than closed, except for those with a moraic nasal /N/ or the first part of a geminate consonant in the coda position. Table 8 lists examples of closed syllables in Japanese where ‘C,’ ‘V,’ ‘N,’ and ‘QC’ denotes ‘consonant,’ ‘vowel,’ ‘moraic nasal’ and ‘geminate consonant’ respectively. Syllables usually consist of one or two morae. Those with more than two morae are phonologically marked and avoided except “when the morphology allows no other choice” (Poser 1990, p 79).

Table 8 Examples of closed syllables in Japanese

Syllable structure	Example	Gloss
VN	/eN/	bond
CVN	/hoN/	book
VQ.(CV)	/iQpo/	one step
CVQ.(CV)	/seQta/	sandal

It has been argued that Japanese has the rhythm regulated by bimoraic foot (Poser 1990). The preference for the bimoraic rhythm has been supported by several morphological and prosodic phenomena. Poser (1990) argues a suffix *-tyan* (a degenerate form of the honorifics *-san*; Kubozono 1999) adjoins bimoraic stem. Table 9 is an excerpt of the examples introduced by Poser (1990), which lists possible and impossible truncations of stems of Japanese names attaching to the suffix *-tyan*. For example, when the name *Taro* /taroo/ (/aa/ is a long vowel which has the phonological length of two morae) attaches to the suffix *-tyan*, it will be truncated to a bimoraic unit /taro/. These truncated stems are always bimoraic (i.e. consisting of even-numbered morae) so that /taro/ and /hana/ are possible but */ta/ and */ha/ are not. Rather irregular truncations such as /hiroko/ into /hii/ and /akiko/ into /ako/ also conform the rule of bimoraicity. In the former example, vowel lengthening is observed to make the truncated stem bimoraic rather than monomoraic (*/hi/). When the original name is longer (e.g. /wasaburoo/), it will be truncated either to /wasaburo/ (consisting of four morae) or /wasa/ (consisting of two morae). Another common example to support the bimoraicity of

Japanese foot comes from the claim by Bekku (1977) that Japanese prose *haiku* which was said to have an odd-numbered rhythm (5-7-5) actually has a bimoraic rhythm counting the pauses in each line. The theory is empirically supported by the results of the perceptual study on the native speakers of Japanese by Cole and Miyashita (2008).

Table 9 Possible and impossible truncations of stems attaching to the suffix -tyan in Japanese (Poser 1990)

Original stem	**Possible truncations**	**Impossible truncations**
/taroo/	/taro/	
/hanako/	/hana/	
/hiroko/	/hii/	*/hi/
/akiko/	/ako/	
/wasaburoo/	/wasaburo/ /wasa/	*/wasabu/ */wa/

2.3.2. Isochrony in Japanese mora rhythm

Japanese isochrony has been investigated in relation to its mora rather than foot. It is well known that word durations are proportional to the number of morae (Bradlow et al. 1995, Port et al. 1987). The temporal rhythm of mora is not affected by its lexical pitch accent (Beckman 1992) unlike English, in which stressed vowels are markedly longer than unstressed vowels. Japanese unaccented vowels are not reduced either. The presence of phonological durational contrast of vowels (i.e. short and long vowels) seems to be adding further stability to its isochronous mora rhythm (Warner & Arai 2001b).

Although the mora rhythm is fairly isochronous (at least more than English foot), it is still susceptible to several phonetic effects. These include intrinsic differences of segmental durations (e.g. low vowels are longer than high vowels) and the mora structure (i.e. morae without onset consonants tend to be shorter) (Warner & Arai 2001a). Due to final lengthening, morae become longer in the final positions of intonation phrases. In addition, so-called special morae (i.e. moraic nasals, the first parts of geminates and the second part of long vowels) are often shorter compared to other *CV* (or *V*) morae (Arai 1999, Beckman 1982). According to the study on a spoken corpus by Campbell and Sagisaka (1991), long vowels were only 1.5 times as long as their short counterparts. Devoiced vowels also become shorter than their voiced counterparts (Komatsu & Aoyagi 2005, Shaw & Kawahara 2018), adding further variations to the mora durations of Japanese.

To manifest the isochrony, compensatory shortening is observed as for the English foot. While shortening of foot-internal syllables is said to have been observed in English (Section 2.2.2), several studies have shown that mora durations of Japanese speech are adjusted between segments. Previous studies (Miyagawa-Kawai 1999, Otake 1988, 1989, Port et al. 1980) investigated the effects of consonant durations on the adjacent vowel durations and found negative correlations. That is, vowel durations adjust themselves to maintain isochronous mora durations. In addition, other studies have shown a correlation between the number of morae and speech durations in larger units of utterances (Bradlow et al. 1995, Port et al. 1987, Sato 1995). For example, in Port et al.'s study, this linear increase of duration was observed in one- to seven-mora utterances. This linear relation suggests durational compensations not only between adjacent consonants and vowels but among segments in the utterance[3]. However, it is not clear whether the durational compensations observed for Japanese mora would influence those for syllable durations in the *JE* rhythm.

[3] The unit in which the compensation takes place (e.g. phonological word) was not specified in any of the studies.

2.4. Japanese English rhythm

2.4.1. Lexical stress in Japanese English

English and Japanese use different parameters to manifest their phonological contrasts. For its phonological contrasts, English uses vowel duration (e.g. stressed vowels usually have considerably greater durations than unstressed vowels), vowel intensity (e.g. stressed vowels have significantly greater intensity than unstressed vowels), vowel F0 (e.g. vowels with a pitch accent have notably higher or lower F0, or a much more considerable vowel-internal change in the F0, than vowels without pitch accent) and vowel quality (e.g. there is vowel reduction associated with the absence of lexical stress so that the F1 and F2 values of unstressed vowels are more centralised, i.e. approach those of schwa, compared to stressed vowels) (e.g. Knight 2012, Roach 2009; See Section 2.1). On the other hand, Japanese relies on duration to manifest its segmental contrast (i.e. phonologically long vowels have significantly greater durations than short vowels) and F0 for its accent (i.e. vowels with pitch accent have significantly higher F0) as well as intonation (e.g., phrasings are manifested with downsteps, i.e. stepwise lowering of the F0 peak within the same intonational phrase). It has no phonological contrast with intensity and no prosodic contrast associated with the quality of vowels. Table 10 summarises the parameters used for phonological contrasts in English and Japanese. A tick "✓" denotes the parameter is used for the phonological contrasts in the language.

Table 10 Parameters used for phonological contrasts in NE and JE

	Vowel intensity	**Vowel F0**	**Vowel duration**	**Vowel reduction**
English	✓	✓	✓	✓
Japanese		✓	✓	

Given the difference, it is likely that parameters used in the *JE* lexical stress are different from those in the *NE* lexical stress. As discussed in Section 2.2.4, vowel duration and intensity are two primary phonetic correlates of the *NE* lexical stress (De Jong 2004, Fear et al. 1995, Okobi 2006, Sluijter & van Heuven 1996a). Since Japanese uses vowel duration but not intensity for its phonological contrasts, it could be hypothesised that *JE* only uses vowel duration for its lexical stress contrast (Table 11).

Table 11 Hypothetical correlates of lexical stress in NE and JE

	Intensity	**Duration**
NE	✓	✓
JE		✓

However, the results of the previous studies suggest that, as well as vowel duration, vowel intensity is also a correlate of *JE* stress contrasts. Lee et al. (2006) investigated the groups of learners whom they call *early bilinguals* and *late bilinguals* of Japanese and English. They were all native speakers of Japanese who had resided in the United States for about 10 to 20 years. The result of the phonetic analyses suggests nativelike manifestations of lexical stress by both groups, i.e. there was no significant difference in the vowel duration and vowel intensity between the native speaker control group and either of the two bilingual groups. Although the *JP* groups in Konishi and Kondo (2015) had much lower proficiency (e.g., not many of them had lived in any English-speaking country), their manifestation of lexical stress contrasts with intensity were not significantly different from that of the *EN* group either. The results of these studies suggest intensity contrasts of lexical stress are easy for *JP Beg*, not only *JP Adv*.

On the other hand, those previous studies showed contrastive results regarding vowel durations in relation to manifestations of lexical stress. While Lee et al. (2006) found no significant difference in the use of vowel durations between the *EN* control group and the Japanese-English bilingual groups, in Konishi and Kondo (2015), a significant difference of durational contrasts of vowels was found between the *EN* group and *JP* group of lower proficiencies. The difference should have been due to the different proficiencies of the *JP* groups in the two studies. While the *JP* groups in Lee et al. (2006) were proficient enough to be called *bilinguals*, the majority of the *JP Beg* group in Konishi and Kondo (2015) had no experience living in English speaking countries. Since the *JP* subjects in the current study are much more similar to those of Konishi and Kondo (2015) than those of Lee et al. (2006), similar results to the former study are expected for the analyses of the vowel intensity.

As well as Lee et al. (2006) and Konishi and Kondo (2015), many other previous studies on L1 and L2 English stress analysed target words embedded in a carrier sentence (e.g. de Jong 2004, Plag et al. 2011, Visceglia et al. 2010, Zhang et al. 2008). Furthermore, except for de Jong (2004) and Plag et al. (2011), such studies did not take into account the effect of post-lexical prosody, e.g. pitch accent and focus prosody. Hence the stressed vowels in most of these previous studies would have been affected by post-lexical prosody, which may have contaminated their data. The largest effect would have been observed for the F0 of the stressed vowels. Target words embedded in carrier sentences usually bear an H* pitch accent, which is realised with a notably higher F0 than the mean F0 of the sentence. In addition, it has been reported that vowels of the words with a sentence focus are realised with a greater duration (Kochanski et al. 2005, Sluijter & van Heuven 1996a). Since it is evident with the discussion of the *NE* phonology that F0 is not a correlate of lexical stress (See Section 2.2.3), it will not be examined in the current study.

2.4.2. Vowel epenthesis in Japanese English

English and Japanese have different phonotactics. English has complex syllable structures with complex consonant clusters both in the syllable onset and coda positions. In contrast, Japanese has much simpler syllable structures with no consonant clusters and coda consonants. The exceptions are moraic nasal (coronal nasal /N/) and the first elements of geminates (See Section 2.3.1). Because of the phonotactic constraint, *JE* speakers are known to epenthesise vowels in consonant clusters (i.e. between consonants) and after syllable-final consonants (i.e. between a consonant and a pause), except for coronal nasals (e.g. Dupoux et al. 1999, Masuda & Arai 2010, Mazuka et al. 2011, Tajima et al. 2000). For example, *street* /striːt/ is often produced as /s**u**t**o**riːt**o**/ (epenthetic vowels in the onset consonant cluster and after the coda consonant) especially in *Beg JE.* Yazawa et al. (2015) reported a negative correlation between the frequency of vowel epenthesis and the perceived goodness of *JE* speech (i.e. *JE* with less frequent vowel epenthesis tended to be perceived more proficient).

It has been known that the kind of epenthetic vowel depends on the type of preceding consonant, which is similar to the patterns observed in Japanese loanword phonology. The rules are summarised in Table 12. The first column shows the types of epenthetic vowels and the second column shows the environment, i.e. preceding consonants. The most commonly epenthesised vowel is /ɯ/ because of its shortest duration and lowest sonority in the five vowels of Japanese (Kubozono 1999). Following post-alveolar affricates /tʃ/ and /dʒ/, /i/ is epenthesised because of its similar place of articulation to that of the consonants. Following alveolar plosives /t/ and /d/, /o/ is epenthesised because the consonants would be produced as [tʃ] and [dʒ] before /i/. Because each vowel phoneme has different intrinsic acoustic characteristics (Fairbanks

& House, 1950; Fairbanks, 1953; Whalen & Levitt,1995), the degree of influence of an epenthetic vowel on foot durations could vary depending on the preceding consonant. Based on the result of the study by Yazawa et al. (2015), the speaker proficiency does not affect the quality of the epenthetic vowels in *JE.*

Table 12 The environments and types of epenthetic vowels in JE

Epenthetic vowel	Preceding consonants
i	tʃ, dʒ
o	t, d
u	all other consonants

Vowel epenthesis has been attributed to the perception of illusory vowels, which is caused by the transfer of L1 phonotactics. A famous experiment by Dupoux et al. (1999) demonstrated that *JP* tend to hear /ebzo/ as /eb**u**zo/, with an epenthetic vowel between the consonants, while French speakers do not. This was supported by acoustic and articulatory studies (e.g. Funatsu et al. 2008, Shibuya & Erickson 2010, Yazawa et al. 2015), which indicated that vowel epenthesis in *JE* is influenced by the phonotactic constraints in Japanese.

Other studies explain vowel epenthesis by articulatory difficulty, i.e. effects of Japanese phonotactics in articulating English. For example, it has been claimed that the production of epenthetic vowels is affected by the phonetic details of the consonant sequences. That is, when a non-native consonant cluster is produced with insufficiently overlapping configuration or mistiming due to the articulatory difficulty, a transitional vowel-like structure appears within the cluster (Davidson 2011, Davidson et al. 2015, Hall 2003). Previous studies reported that the frequency of vowel epenthesis by *JE* was subject to adjacent phonetic environments such as voicing of the following consonants, which makes it difficult to articulate the sound sequence without an epenthetic vowel in between (e.g. Masuda & Arai 2010, Tajima et al. 2000).

Most previous studies on vowel epenthesis focused on the conditions and environments of epenthesis (Davidson 2011, Davidson et al. 2015, Dupoux et al. 1999, Hall 2003, Masuda & Arai 2010, Tajima et al. 2000) or qualities of the epenthetic vowels (Funatsu et al. 2008, Shibuya & Erickson 2010, Yazawa et al. 2015); not many studies have been conducted on vowel epenthesis in relation to speech rhythm. However, vowel epenthesis is assumed to have considerable effects on the *JE* rhythm. Because epenthetic vowels are *extra* vowels added to some feet but not to others, *JE* speech with more frequent vowel epenthesis would be less isochronous, i.e. have a greater variance of foot durations. This will be examined in the current study.

2.4.3. Foot rhythm in Japanese English

Although the rhythmic unit of Japanese is different from that of English (mora in Japanese and foot in English), its rhythm has also been discussed in relation to isochrony. It is well known that the duration of utterances in Japanese speech tends to be proportional to the number of morae (Bradlow et al. 1995, Port et al. 1987). Despite some phonetic obstacles in the realisations of isochronous mora rhythms (e.g. special morae are shorter than other morae; Arai 1999, Beckman 1982, Campbell and Sagisaka 1991), Japanese mora rhythm is known to be much more isochronous than English foot rhythm. In contrast to English foot rhythm, which is manifested with lexical stress, Japanese rhythm is not affected by the presence or absence of lexical pitch accent either.

The differences in the rhythmic units in English and Japanese, as well as their different syllable structures, would result in the transfer of the Japanese rhythm to the *JE* rhythm. The L1 transfer would make the *JE* rhythm less isochronous since Japanese speech is more proportional to the number of morae than English speech is to the number of syllables. Despite the potential negative transfer, previous studies on *JE* prosody did not focus on the manifestations of isochronous rhythm. Most of them investigated other prosodic factors which influence English foot rhythm, e.g. lexical stress (Lee et al. 2006, Konishi & Kondo 2015) and consonant and vowel intervals (Grenon & White 2008, Kawase et al. 2016, Ozaki et al. 2017).

One study which investigated *NE* and *JE* foot rhythm is Mochizuki-Sudo and Kiritani (1991), who conducted both production and perception experiments of the *JE* foot (which they called *interstress intervals*). They examined the relationship between the foot duration and the number of syllables in the read speech. As a result of their production experiment, they observed smaller proportional increases of foot duration with additions of foot-internal unstressed syllables in the speech of more proficient speaker groups. That is, the increase was smaller for the *EN* group than the *JP Adv* group and smaller for the *JP Adv* group than the *JP Beg* group. However, the difference was not tested statistically. In addition, they investigated compensatory shortening of durations of stressed vowels by comparing feet containing one syllable (i.e. a stressed syllable only) to four syllables (i.e. a stressed syllable and a following three unstressed syllables). The result showed greater shortening in *NE* and *Adv JE* than *Beg JE* when one-syllable feet and two-syllable feet were compared*.* Also, the shortening was consistently observed from one- to four-syllable feet in the *Adv JE* and *Beg JE*. However, the shortening was not consistently observed for three-syllable and four-syllable feet in the *NE*.

Mori et al. (2014) statistically compared durations of interstress intervals of the read speech between *NE* and *JE.* They had the subjects read the same sentences several times and calculated the mean and variance of the duration of each interstress interval. The result indicated significant differences between the *NE* and *JE*. However, when the ratios of two adjacent interstress intervals were compared, the difference between the *NE* and *JE* was not statistically significant. Although the results of their study imply some difference in the manifestations of interstress intervals between *NE* and *JE*, they cannot so much be interpreted in relation to isochrony, which can be measured with the variance of durations *among* different feet (or interstress intervals) rather than those *within* each foot.

2.5. Summary

The current thesis aims to investigate the acquisition of English rhythm by native speakers of Japanese. The chapter first summarised previous studies on English and Japanese rhythm, which are the basis of the discussions of *JE* speech rhythm. Foot is the basic unit of speech rhythm in English, which is a stress-accent language (Beckman 1986, Abercrombie 1967). Especially, trochaic, rather than iambic, feet are dominant in English, which is said to help identify syntactic boundaries (Cutler and Butterfield 1991, Cutler and Carter 1987, Taft 1984). Previous studies are rather inconsistent as to the maximum number of syllables within each foot. Some studies seem to have allowed only disyllabic foot (Burzio 1994, Gussenhoven 2004, Hayes 1995) while others (Erickson et al. 2012, Mott 2011) defined foot to contain more than one unstressed syllable following a stressed syllable. The current study adopts the latter definition.

Since English feet consist of stressed and unstressed syllables, lexical stress is supposed to have a considerable influence on its speech rhythm. One important characteristic of English associated with its lexical stress is reductions of unstressed vowels, which are often realised as a schwa. A schwa is sometimes deleted or approximates its quality to a consonant of a similar place of articulation (Gussenhoven 2004). Although some previous studies (e.g. Bolinger 1981, Cruttenden 2014) support vowel-based rather than stress-based rhythm in English, the current study adopts the stress-based contrast following Lehiste (1972), Huggins (1973) and Rakerd et al. (1987).

Like many other languages, no complete isochrony has been observed in English foot rhythm. However, studies have shown *NE* speakers' efforts towards isochronous rhythm and *NE* listeners preference of it. Fowler (1977, 1981), Huggins (1973), Lehiste (1972) and Rakerd et al. 1987 observed compensatory shortening of foot-internal stressed syllables with the increase in the number of foot-internal syllables. No study has reported the shortening of foot-internal unstressed syllables. *EN*'s preference on isochronous foot rhythm has also been empirically supported by Lehiste (1975, 1977).

In contrast to the stress-based rhythm in English, Japanese has the mora-based rhythm (Port et al. 1987). Although many phonological phenomena have been explained with its bimoraic feet (Poser 1990), the isochronous rhythm has been investigated in relation to morae rather than bimoraic feet (Bradlow et al. 1995, Port et al. 1987). In addition, native Japanese speakers' preference for the bimoraic rhythm has been supported by several morphological and prosodic phenomena (Bekku 1977, Cole & Miyashita 2008, Poser 1990). However, the isochronous rhythm is still susceptible to various phonetic phenomena including intrinsic durations of segments (Warner & Arai 2001a), those of special morae (Arai 1999, Beckman 1982, Campbell and Sagisaka

1991) and those of devoiced vowels (Komatsu & Aoyagi 2005, Shaw & Kawahara 2018). As in English foot, compensatory shortening has been observed in Japanese speech. Miyagawa-Kawai (1999), Otake (1988), 1989 and Port et al. (1980) have shown that adjacent segments adjust durations to maintain isochronous mora timing. The durational adjustment was also observed in larger units of speech (Bradlow et al. 1995, Port et al. 1987, Sato 1995). Overall, isochrony of the Japanese mora rhythm seems to be more robust than that of the English foot rhythm.

Based on the different rhythmic structures in English and Japanese, *JE* rhythm is assumed to have influences of L1 Japanese rhythm. However, regarding manifestations of lexical stress contrasts, the results of the previous studies (Konishi & Kondo 2015, Lee et al. 2006) indicate the use of the same phonetic parameters in the *JE* as *NE*, namely vowel duration and intensity. Still, a statistical difference exists between the degree of the contrasts made. *EN* and *JP Adv* are known to manifest greater stress contrasts with vowel durations than *Beg JP*.

Many other previous studies have investigated the effects of F0 on the manifestation of *JE* lexical stress. However, based on the latest theory of metrical phonology (See Section 2.2.4), F0 is considered as a correlate to pitch accent, which in most of such previous studies was overwritten onto the target stressed syllables to be analysed; the stressed syllables of target words embedded in carrier sentences usually bear an H* pitch accent. Especially since the current study investigates rhythm, rather than intonation, F0 is not going to be analysed.

Another pervasive problem in *JE* is vowel epenthesis because of the transfer of L1 Japanese phonotactics, in which syllable structures are much simpler compared to that of English. Especially, vowel epenthesis is frequently observed between consonant clusters and after syllable-final consonants. Despite extensive studies conducted on vowel epenthesis, the phenomenon has not been investigated in relation to speech rhythm. Since epenthetic vowels would affect the variance of foot durations, making the speech sound less isochronous, the current study aims to investigate their actual effects.

Lastly, the chapter summarised previous studies on the *JE* foot rhythm, the investigation of which is the primary goal of the current study. English and Japanese have different prosodic units which are relevant to the isochrony in their speech. English is said to manifest isochrony with its foot, which could consist of more than one vowel. In contrast, isochrony in Japanese is discussed in relation to its mora, which consists of only one vowel. In other words, Japanese speech is roughly proportional to the number of vowels while English speech is not. Hence L1 transfer of Japanese mora rhythm is expected in the isochrony of *JE* foot rhythm. One previous study (Mochizuki-Sudo and Kiritani 1991) supports the existence of the transfer, no statistical analysis was conducted to support it. Another study (Mori et al. 2014) statistically compared durations of *JE* feet. However, it focused on the mean and variance of each individual foot in relation to utterance tokens rather than comparing durations of different feet.

The aspects of *JE* rhythm discussed in the current chapter will be statistically analysed in the later chapters. Manifestations of *JE* lexical stress (discussed in Section 2.4.1) will be investigated in Chapter 5. Epenthetic vowels (discussed in Section 2.4.2) will be analysed in Chapter 6. Foot durations (discussed in Section 2.4.3) will be examined in Chapter 7.

3. The Corpus

3.1. Introduction

Studies in the field of second language acquisition used to be conducted on a small number of subjects, usually due to limitations of time and resources. However, it has been said a wide range of individual differences exist in second language acquisition (e.g. Dörnyei 2014, Roberts & Meyer 2021, Skehan 1991), which makes it difficult to investigate the general tendency of learners with a small amount of data. This is true of the acquisition of L2 speech as well (Saito 2019, Slevc & Miyake 1996). Therefore, a growing number of studies began to be conducted using large-scale data, i.e. linguistic corpus. Vienna-Oxford International Corpus of English (VOICE)[4] and Written ELF in Academic Settings (WrELFA) Corpus (ELFA)[5] are among the biggest corpora of the non-native varieties of English. However, these large-scaled corpora are usually written, rather than spoken, ones perhaps because constructions of speech corpora take much more time than written corpora mainly due to time for annotating the recorded speech.

Despite the difficulty of construction, the Asian English Speech cOrpus Project (AESOP; cf. Meng et al. 2009, Visceglia et al. 2009, Kondo, 2012) was launched in 2008 to construct large-scaled corpora of L2 English spoken by native speakers of different languages in Asian regions. The AESOP team aims to construct corpora in different Asian countries including Japan, Taiwan, Hong Kong, China, Indonesia, Thailand, and Vietnam. The corpora are being constructed on the same platform, e.g. the same recording devices and the scripts in case of the read speech, which makes it easier to compare L2 English by speakers of different L1s.

[4] https://www.univie.ac.at/voice/page/corpus_description

[5] https://www2.helsinki.fi/en/researchgroups/english-as-a-lingua-franca-in-academic-settings/research/wrelfa-corpus

In order to examine the effects of Japanese on L2 English rhythm, the current study analysed the data extracted from the *J-AESOP* corpus, which is the *AESOP* corpus consisting of *JE* speech. With reference to Kondo (2012), the following sections will introduce the design of the corpus.

3.2. Subjects

J-AESOP consists of speech data of 183 *JE* speakers (115 females and 68 males) as well as 25 *NE* speakers (16 females and 9 males). At the time of recording, all of them were undergraduate or postgraduate students at a Japanese university located in Tokyo or surrounding areas. Their ages ranged from 18 to 38 (mean = 20.3). Out of the 183 *JE* subjects, 63 had experienced living in one or more English-speaking countries; the length of residence (LOR) ranged from one month to 11 years (mean = 29.3 months). However, as shown in Chapter 4, their proficiency was determined not according to their LOR but by ratings of experienced phoneticians.

3.3. Methods

The recording was conducted at the home university of each subject. Depending on the recording venue, subjects recorded their speech in a sound-proof or quiet room. For the recordings of the read speech, scripts were provided on a computer screen, as well as the instructions. In the picture-description task, the pictures were presented on the screen. Only for the recording of *the North Wind and the Sun*, the data of which is investigated in this dissertation, subjects were asked to practise reading in advance. The speech was recorded with Sennheiser PC 166 Headset (sampling frequency 16 kHz, 16-bit quantization).

There are eight sets of recording tasks: 1) target words in career sentences, 2) target words at prosodic boundaries, 3) target words in narrow focus, 4) reduced and unreduced function words, 5) prosodic disambiguation, 6) the North Wind and the Sun, 7) Computer-prompted dialogue and the 8) picture description task. Scripts for the read speech is shown in the Appendix.

3.4. Data for the current study

Among the sets of recordings in the *J-AESOP* corpus, the read speech of *the North Wind and the Sun* was adopted in order to examine the speech rhythm of a long discourse rather than utterances of individual sentences. The whole script of the passage is presented below. As the subjects were instructed to practise reading the text prior to the recording session, not much dysfluency was expected. In addition, the subjects were instructed to go back to the beginning of the sentence they stuttered while reading, which ensured for each subject a complete set of fluent utterances of all the sentences. Since it was a read speech with a script, relatively few spontaneous vowel reductions are expected (Shockey 2003).

> *The North Wind and the Sun were disputing which was the stronger when a traveler came along wrapped in a warm cloak. They agreed that the one who first succeeded in making the traveler take his cloak off should be considered stronger than the other. Then the North Wind blew as hard as he could. But the more he blew, the more closely did the traveler fold his cloak around him. And at last, the North Wind gave up the attempt. Then the Sun shone out warmly. And immediately the traveler took off his cloak. And so the North Wind was obliged to confess that the Sun was the stronger of the two.*

3.5. Annotation

The audio files were first annotated with forced alignment based on the acoustic information of each segment as well as its transitions across segments. The automatic alignment was done using Hidden Markov Model Toolkit (HTK)[6] based on the TIMIT speech corpus[7]. Since the TIMIT speech corpus is a collection of speech data of native American English speakers reading sentences, it was insufficient to annotate *JE* speech. Therefore, the word dictionary was modified to accommodate variations of pronunciations by *JE* speakers.

Table 13 lists the variants of the pronunciations of the word *closely* added to the pronunciation dictionary for the automatic alignment (Kondo 2012). The first column lists the variants of the pronunciations. The *Vowel* and *Consonant* columns respectively show whether the respective pronunciation variant contains different consonants/vowels from the target ones. The *Epenthesis* column shows whether the variant contains epenthetic vowels. The variant in the top row, [klousli], is the *NE* pronunciation of the word. The actual examples of the pronunciations with vowel variants are [kloːsli], [kloːsɹi], [kɹoːsli], [kɹoːsɹi], [kʊɹoːsli], [kʊɹoːsɹi]. Those with consonant variants are [klousɹi], [kloːsɹi], [kɹousli], [kɹoːsli], [kɹousɹi], [kɹoːsɹi], [kʊɹousli], [kʊɹoːsli], [kʊɹousɹi], [kʊɹoːsɹi], [kʊɹousʊɾi]. Those with epenthetic vowels include [kʊlousli], [kʊɹousli], [kʊɹoːsli], [kʊɹousɹi], [kʊɹoːsɹi], [kʊɹousʊɾi].

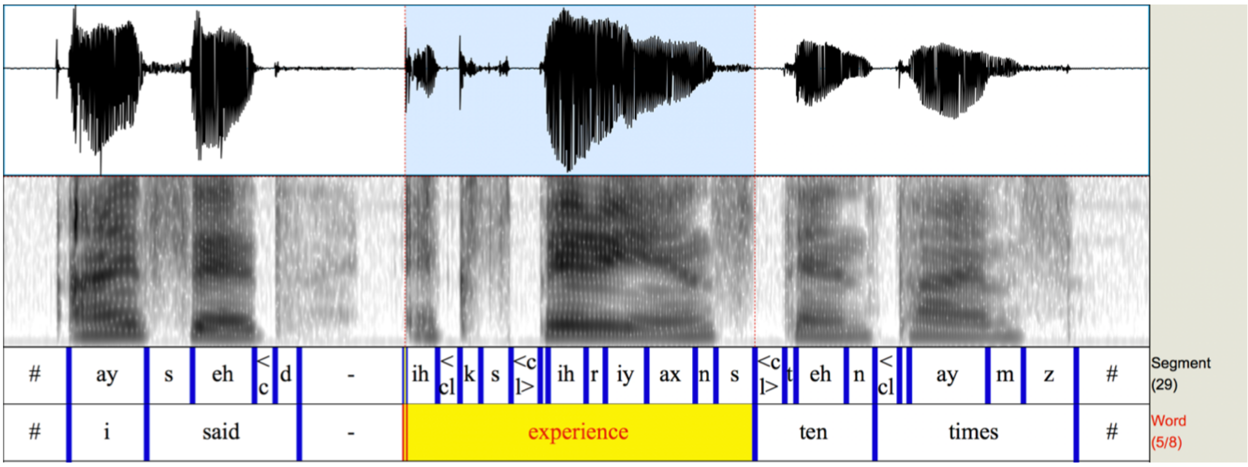
[6] http://htk.eng.cam.ac.uk/

[7] https://catalog.ldc.upenn.edu/ldc93s1

Table 13 The variants of the pronunciations of the word ‘closely’ in the pronunciation dictionary for the automatic alignment of the J-AESOP corpus (Kondo 2012)

	Vowel	**Consonant**	**Epenthesis**
[klousli]			
[kloːsli]	✓		
[klousɹi]		✓	
[kloːsɹi]	✓	✓	
[kɹousli]		✓	
[kɹoːsli]	✓	✓	
[kɹousɹi]		✓	
[kɹoːsɹi]	✓	✓	
[kʊlousli]			✓
[kʊɹousli]		✓	✓
[kʊɹoːsli]	✓	✓	✓
[kʊɹousɹi]		✓	✓
[kʊɹoːsɹi]	✓	✓	✓
[kʊɹousʊɾi]		✓	✓

The annotation of segments and words has been provided in the *TextGrid* format of Praat[8] (Boersma & Weenik 2020). Figure 4 is an example of annotation in the *J-AESOP* corpus. Below the oscillogram and spectrogram, there are two tiers, the *segment* tier and the *word* tier. The segments are transcribed with ARPABET, which is an ASCII-character-based transcription system developed by Advanced Research Projects Agency.



# ay s eh <c d - ih <cl k s <cl> ih r iy ax n s <cl> t eh n <cl ay m z #
Segment (29)
# i said - experience ten times #
Word (5/8)

Figure 4 Example of the annotation in the *J-AESOP* corpus

Segment boundaries of the output data were then modified by phonetically trained annotators. This ensured accurate representations of segment insertions and deletions. However, no segments were inserted or deleted unless acoustic evidence was clear on the spectrogram. A segment was labelled as a vowel if it had a clear formant structure. In case a *JE* vowel was devoiced because of the influence of L1 phonology (high vowels /i/ and /ɯ/ are devoiced between voiceless consonants or when preceded

[8] https://www.fon.hum.uva.nl/praat/

by a voiceless consonant and followed by a pause), the relevant segment was deleted. In order not to discriminate dialectal variations, diphthongs and rhotic vowels were treated as single segments. If a boundary was not clear between two segments, it was defined proportional so that the intervals of the vowel and the consonant would be 2 to 1. Figure 5 is an example of such segment boundaries. Since the boundary between /n/ in the word *in* and /m/ in the word *making* (highlighted red on the waveform) is not clear on the spectrogram, it was defined proportionally. As /n/ and /m/ are consonants, the intervals were defined to have equal durations.

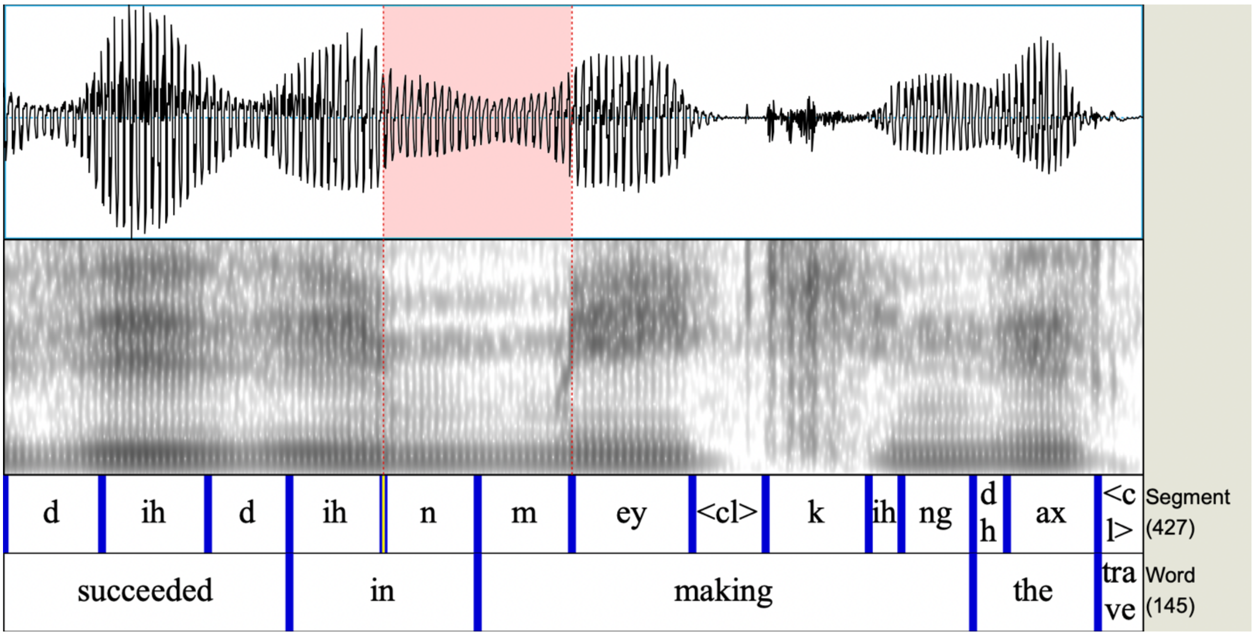


Figure 5 An example of segment boundaries defined proportionally

The segments were annotated phonemically, not phonetically. For example, even when a voiced consonant is produced as voiceless, the segment is still labelled as voiced. This also applies to L2 speech labelling notation. For example, various realisations of *JE* /l/ and /r/ were all labelled with the target phoneme. However, when it was clear on the spectrogram that a segment had not been realised (such as unreleased phrase-final plosives), the segment interval was removed.

In addition, some supplementary labels mark various events such as pauses and dysfluencies. In Figure 6, the *(WRG)* label denotes that the original word *blew* is misread as *blow*. The *(RPT)* shows the word *as* in the script is repeated. This ensures the exclusions of the same words repeated where necessary. The *(DSF)* label (at the end of the highlighted section) marks observed dysfluency that could not be transcribed. Epenthetic vowels and consonants were marked so that their durations could be extracted. Also, when the realisation of a word was considerably deviant from its canonical pronunciation, the segment was marked as *(DVN)* so that it would be excluded from the data for the analyses where necessary.

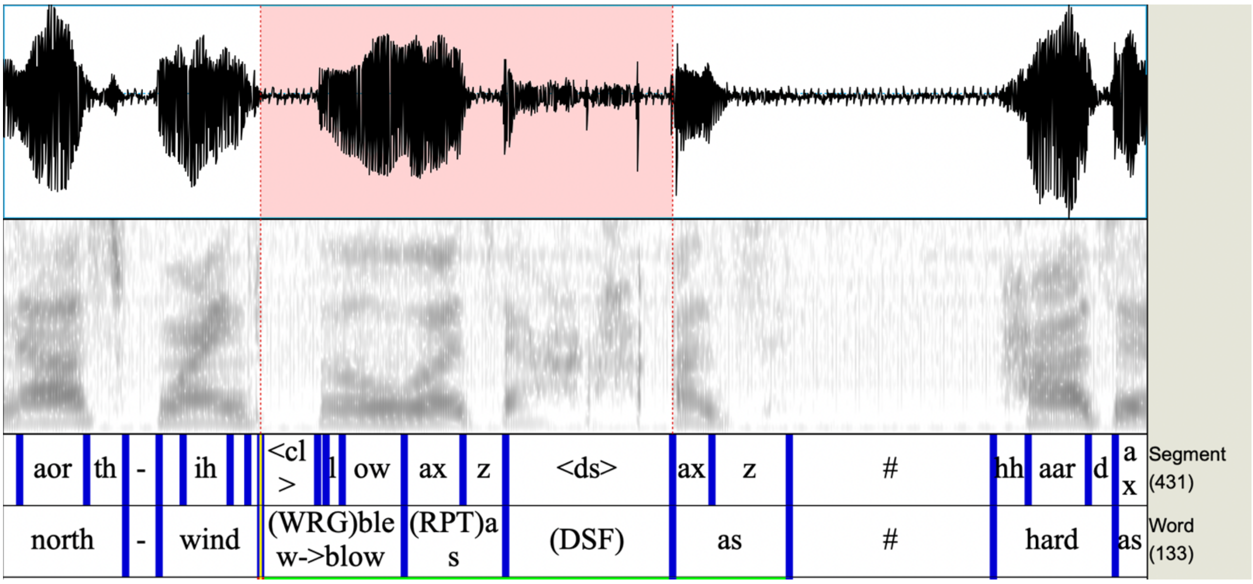


Figure 6 Examples of mispronunciation, repetition and dysfluency

The original script is '(Then, the) north wind blew as hard as (he could)'

The (WRG) label denotes that the word 'blew' is mispronounced as 'blow.'

The (RPT) label denotes that the first 'as' in the phrase 'as hard as' is repeated.

The (DSF) label marks dysfluency observed after the first 'as.'

Because of the difficulty of identifying epenthetic vowels and consonants with the HTK, the annotation was conducted manually. Four criteria in Tajima et al. (2000) were adopted to identify epenthetic vowels, namely 1) clear vowel-like formant structure (especially F1) on the spectrogram; 2) periodic waveform; 3) a minimum of two pitch periods; and 4) (if it is adjacent to a consonant with formant structure,) the evidence as a vowel. If all the four conditions were met, the relevant portion was marked as an epenthetic vowel. Table 14 lists examples of words and word sequences in the *J-AESOP* corpus in which vowel epenthesis was observed. The ‘Pronunciation’ column shows the transcription of the word with the International Phonetic Alphabet. <V> represents a potential epenthetic vowel in the word. The ‘Environment’ column shows whether the epenthesis occurs within the consonant cluster or in the syllable-final position.

Table 14 Examples of words with epenthetic vowels in the J-AESOP corpus

Word	**Pronunciation**	**Environment**
agreed	/əg<V>riːd<V>/	consonant cluster / syllable-final
and	/ənd<V>/	syllable-final
as	/əz<V>/	syllable-final
at	/ət<V>/	syllable-final
blew	/b<V>lu/	consonant cluster
could	/kʊd<V>/	syllable-final
closely	/k<V>loʊsli/	consonant cluster

gave	/geɪv<V>/	syllable-final
hard	/hɑːrd<V>/	syllable-final
immediately	/ɪmiːdiət<V>li/	consonant cluster
making	/meɪkɪŋ<V>/[9]	syllable-final
north	/nɔːθ<V>/	syllable-final
of	/əv<V>/	consonant cluster
obliged	/əb<V>laɪdʒ<V>d<V>/	consonant cluster / syllable-final
should	/ʃəd<V>/	syllable-final
stronger	/s<V>t<V>rɔŋgɚ/[10]	consonant cluster
succeeded	/sək<V>siːdəd<V>/[11]	consonant cluster / syllable-final
traveler	/t<V>rævəlɚ/	consonant cluster
warmly	/wɔːm<V>li/	consonant cluster
wind	/wɪnd<V>/	syllable-final

Table 15 shows an actual example of epenthetic vowels in the corpus within the consonant cluster and Table 16 in the coda position. In the example of Table 15, an

[9] A voiced velar plosive /g/ is often inserted between the epenthetic vowel and the preceding velar nasal /ŋ/

[10] Epenthetic vowels between /s/ and /t/ are rather rare since they are in an environment of vowel devoicing in Japanese.

[11] Epenthetic vowels between /k/ and /s/ are rather rare since they are in an environment of vowel devoicing in Japanese.

epenthetic vowel is inserted between the phoneme /t/ and /r/ within the consonant cluster of the word *stronger* /strɒŋgə/. For the word *gave* /geɪv/ on Table 16, an epenthetic vowel is inserted after the syllable-final /v/ of the word. Note the /v/ interval has little frication because the phoneme is produced as /b/ reflecting the Japanese loan-word phonology. After the annotation was done, the duration of each epenthetic vowel was extracted.

Table 15 An example of epenthetic vowels in the J AESOP corpus (within a consonant cluster)

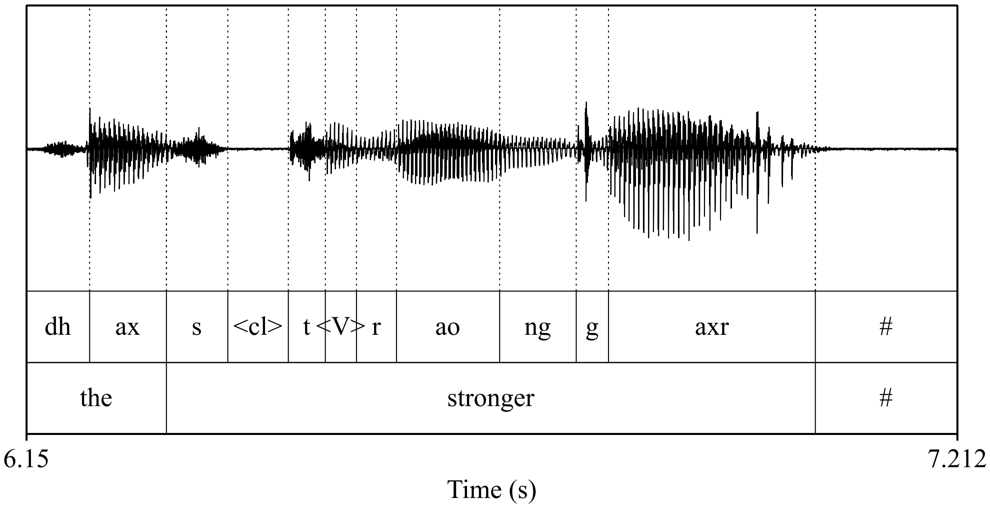
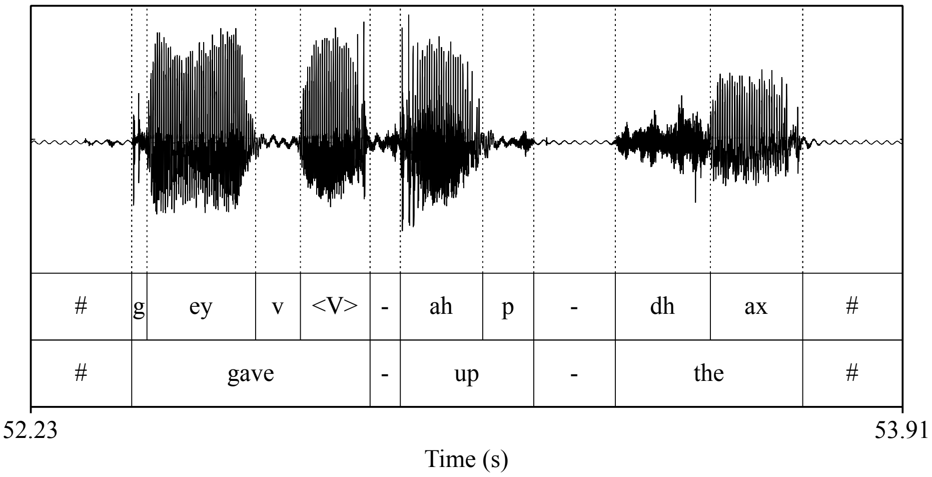


Table 16 An example of epenthetic vowels in the J AESOP corpus (after a coda)



The corpus was not annotated for prosody (e.g. locations of lexical stress and pitch accents). This was due to a lack of objective criteria to identify realisations of lexical stresses and pitch accents. Even though some *JE* speakers seemed to have read some words with non-canonical stress placements, there were no appropriate parameters to judge whether the stress had actually been placed on a non-canonical syllable; the investigation of such parameters itself is one of the primary aims of the current study.

3.6. Summary

This chapter detailed the data for the analyses in the current study, which was extracted from the *J-AESOP* corpus. The corpus consists of the speech data of 183 *JE* speakers as well as 25 *NE* speakers recorded at Japanese universities. Among the eight sets of recordings in the corpus, the read speech of *the North Wind and the Sun* is going to be analysed in the following chapters in investigating realisations of lexical stress, vowel epenthesis and the foot rhythm in *JE*. In order to accommodate variations of pronunciations by *JE* speakers, the audio data was first annotated with forced alignment using HTK based on the TIMIT speech corpus with a modified word dictionary. Segment boundaries of the output data were then modified by phonetically trained annotators. The segments were annotated phonemically, not phonetically so that variants of pronunciations in *JE* were all labelled with their target phoneme. Some supplementary labels mark various events such as dysfluencies, mispronunciations and word repetitions. The annotation of the epenthetic vowels was conducted manually based on the criteria proposed by Tajima et al. (2000). The annotation is especially relevant in Chapter 6, which investigates the phonetic factors of epenthetic vowels in *JE.* The corpus was not annotated for prosody due to a lack of objective criteria to identify realisations of prosodic events in *JE* speech. Therefore, the analyses in Chapter 5 and Chapter 7 will be conducted based on the assumption that all lexical stresses are placed on the canonical syllables in *JE* as well as *NE.*

4. Perceived Proficiency of Japanese English

4.1. Introduction

In order for the dissertation to investigate the developmental change of the *JE* rhythm, the current chapter aims to provide an appropriate measure with which to divide *JP* into proficiency groups. The analyses of *JE* lexical stress (Chapter 5), epenthetic vowels (Chapter 6) and foot (Chapter 7) will all be based on the measure, which will be obtained in this chapter.

Currently, it is common in the field of second language phonetic acquisition to compare groups of learners of different proficiency levels, grouping the subjects according to their scores of some proficiency test (e.g. TOEIC), duration of learning the language or, in case of bilingual studies, ages of arrival. However, such measures do not necessarily correlate with the proficiency to be analysed. Especially, no English tests practically available solely evaluate pronunciations; speaking sections of major tests such as TOEFL and IELTS assess grammar and vocabulary as well as pronunciations. Excluding cases of early bilinguals, acquisitions of pronunciation often show large individual variance due to learners' aptitudes so that they are not necessarily dependent on the duration of learning. Although human ratings would be ideal, not many studies, especially ones with large data, can afford them.

Most of the few studies with human ratings rely on native speakers in evaluating proficiency. For example, a large-scale study on *JE* conducted by Saito et al. (2017) used ratings by 10 native speakers. However, the majority of English speakers in the 21st century are *NNE* (Bolton 2004, Crystal 2003) and it is not ideal to assume native listeners only.

In those human evaluations, *intelligible* or *comprehensible* speech (hereafter collectively referred to as *intelligible* speech)[12] is clearly distinguished from *nativelike* speech. With an increasing number of *NNE* speakers, recent studies of L2 acquisition and teaching focus on intelligible speech. Derwing (2016) argues "even heavily foreign-accented pronunciation can be perfectly intelligible and well-suited to effective communication" (p 28). Since the data for the current study were the read speech, intelligibility could not be measured. Instead, it investigated three phonetic factors estimated to influence intelligibility: *segmental accuracy*, *prosody* and *fluency*.

In order to examine the effect of raters' L1 in evaluating *JE* speech, the current study investigated consistency among raters with different L1s, including *EN*. Section 4.2 will introduce methods and rationale for recruiting raters for the current study. Section 4.3 will explain the detailed procedures of rating. In Section 4.4, inter-rater reliability will be examined for different L1 groups as well as individuals. Section 4.5 will investigate the effects of correlates of intelligibility on perceived nativelikeness. Section 4.6 will explain the scores used for the remaining analyses in this dissertation.

[12] Different scholars have defined the terms *intelligibility* and *comprehensibility* differently. For example, according to Smith and Nelson (2019), *intelligibility* refers to whether individual words are recognised while *comprehensibility* is related to understanding of meanings (so-called locutionary force in pragmatics). As done in many other studies, they will be collectively referred to as *intelligibility* in this thesis.

4.2. Raters

The raters recruited for the project were four *EN*, four *JP* and eight native speakers of other languages (hereafter *OT*). All raters are listed in Table 17 below. The ID numbers of *OT* raters reflect the order of recruitment. Although not intended when recruiting, all *EN* raters happened to be native speakers of American English. All *JP* raters spoke Japanese with the Tokyo accent (the standard accent). One of them was a native speaker of Hokkaido accent, whose main difference from the Tokyo accent regarding pronunciation is mainly on the placement of lexical pitch accent, rather than post-lexical prosody and rhythm (Dallyn 2008).

Table 17 Raters of the J-AESOP corpus

ID	**First language**	**ID**	**First language**
EN01	American English	**OT01**	Spanish
EN02	American English	**OT02**	German
EN03	American English	**OT03**	Mandarin Chinese
EN04	American English	**OT04**	Cantonese
JP01	Japanese	**OT05**	Polish
JP02	Japanese	**OT06**	French
JP03	Japanese	**OT07**	Korean
JP04	Japanese	**OT08**	Punjabi

The *OT* raters were chosen to have diverse linguistic backgrounds both regionally and typologically. Among the languages, one had the stress-based rhythm like English

(German) some had the syllable-based rhythm (e.g. Spanish, Mandarin Chinese, French). In addition, some of the languages had complex syllable structures with frequent consonant clusters like English (e.g. German, Polish, Punjabi) while others had relatively simple syllable structures like Japanese (e.g. Spanish, Mandarin Chinese). Furthermore, the group of languages includes a stress-timed language like English (German), a lexical pitch-accent language like Japanese (Punjabi), tone languages (Mandarin Chinese, Cantonese) and those unspecified for word-level prosody (e.g. French, Korean). See Jun (2005) for a detailed prosodic typology of languages although not all languages in the *OT* group for the current study are discussed.

All raters had completed a postgraduate course in phonetics, second language acquisition or related fields (e.g. speech engineering, language teaching). In Saito et al. (2017), there were some differences between “experienced” raters (those with experiences in linguistics or teaching) and “inexperienced” raters in the rating of “more complex linguistic categories, such as suprasegmentals (i.e. word stress, intonation, rhythm)” although high interrater agreement (Cronbach Alpha > .9) was obtained in “more intuitive and conceptually simpler categories, such as global speech judgement (comprehensibility, accentedness)” (p 449). Recruiting such experienced raters would ensure no inconsistency other than the one potentially arising from L1 differences since the rating criteria in the current study assumed some knowledge of linguistics.

4.3. Procedures

Human-rated proficiency scores of the read speech of *the North Wind and the Sun* were assigned to each *JE* speaker for four categories: a) *segment*, b) *prosody*, c) *fluency* and d) *nativelikeness*. Although not necessarily a common term, the expression *nativelikeness* was adopted rather than a more common *accentedness* in order to make the scale consistent with the other three (i.e. higher score denotes better proficiency). Table 18 shows the actual instruction given to each rater. In order to distinguish between rhythm and intonation, factors related to speech rhythm were all included in the *prosody* category while those related to intonation were included in the *fluency* category.

Table 18 Instruction given to each rater of the J-AESOP corpus

Segment	- Sound of individual vowels and consonants (e.g. place and manner of articulation, voicing contrast)
Prosody	- Clear lexical stress - Speech rhythm - Speech free of wrong insertions/elisions of segments
Fluency	- Intonation - Speech rate - Pausing - Stumbling
Nativelikeness	- Speech free of foreign accent

Prior to the rating, the recorded speech was divided into three sections (Table 19) in order to improve the accuracy of the rating by having raters evaluate the same speaker three times and avoiding local dysfluencies to affect the impression of the whole utterance. Recordings of 25 *NE* speakers were included with the hope of helping to eliminate raters' bias towards *NNE* speakers. The randomised 624 tokens (3 sections x (25 *EN* + 183 *JP*)) were evaluated by each rater. Each rating was on a 10-point scale from 1 (poor) to 10 (good).

Table 19 Three sections of the North Wind and the Sun for the rating of the J-AESOP corpus

Section 1	The North Wind and the Sun were disputing which was the stronger when a traveler came along wrapped in a warm cloak. They agreed that the one who first succeeded in making the traveler take his cloak off should be considered stronger than the other.
Section 2	Then the North Wind blew as hard as he could. But the more he blew, the more closely did the traveler fold his cloak around him. And at last, the North Wind gave up the attempt.
Section 3	Then the Sun shone out warmly. And immediately the traveler took off his cloak. And so the North Wind was obliged to confess that the Sun was the stronger of the two.

4.4. Inter-rater reliability

To investigate the consistency of ratings, Cronbach's Coefficient Alpha (Cronbach 1970; See DeVellis 2005 and Saito et al. 2015 for its application to test inter-rater reliability) was calculated for each category (i.e. *segment, prosody, fluency* and *nativelikeness*). The calculations were done with *alpha()* of the *psych* package on *R* (Ver 4.1.1.). The alpha values were first compared between the rater groups: the *EN* group, the *JP* group and the all-rater group, which include all *EN, JP* and *OT* raters. In addition, the alpha value for the all-rater group was compared with those excluding each rater in the group. This was to ensure whether each of the raters was consistent with other raters. Excluding the 75 tokens of the *NE* utterance (3 sections x 25 speakers), 549 tokens of the *JE* utterance (3 sections x 25 speakers) were analysed.

4.4.1. Segment

The alpha value of the *segment* scores was .93 for the *EN* rater group, .89 for the *JP* rater group and .97 for the all-rater group, which included all *EN, JP* and *OT* groups. As alpha values are known to increase with the increase of the number of items to be compared, the higher value of the all-rater group (.97) compared to those of the *EN* and *JP* groups (.93 and .89) was not solely due to greater inter-rater consistency. Nonetheless, the alpha values indicated high inter-rater reliability in all three groups.

In addition, excluding no individual rater from the all-rater group resulted in greater consistency. Table 20 lists the mean and standard deviation (SD) values of the scores of each rater, and the alpha values if the rater in the row was dropped from the group. Despite variances in the means and SDs of the scores, dropping any one individual rater resulted in the same alpha value. This implies high inter-rater consistency was obtained even in the all-rater group, which included raters with various L1s.

Table 20 Mean and standard deviation of the segment scores of each rater, and the Cronbach's Coefficient Alpha value for the all-rater group when the rater in the row is excluded

	mean	**SD**	**alpha w/o the rater**
EN01	4.0	2.4	.97
EN02	5.8	2.3	.97
EN03	5.2	2.0	.97
EN04	6.7	2.0	.97
JP01	5.8	1.9	.97
JP02	5.6	2.5	.97

JP03	4.4	2.4	.97
JP04	5.7	1.6	.97
OT01	4.4	2.0	.97
OT02	4.8	2.6	.97
OT03	6.3	2.1	.97
OT04	6.9	2.0	.97
OT05	3.8	1.6	.97
OT06	5.1	2.3	.97
OT07	4.7	2.9	.97
OT08	4.6	2.2	.97

4.4.2. Prosody

The alpha value of the *prosody* scores was .92 for the *EN* rater group, .89 for the *JP* rater group and .97 for the all-rater group. The values were almost the same as those of the *segment* scores. Again, the relatively higher value for the all-rater group would perhaps have been due to the greater number of raters. Table 21 summarises the means and SDs of the scores and alpha values for the all-rater group if the relevant subject is dropped. Dropping some of the raters would lower the alpha value but dropping none would raise it, again implying a high consistency in the all-rater group despite the inconsistency in the raters' L1s.

Table 21 Mean and standard deviation of the prosody scores of each rater, and the Cronbach's Coefficient Alpha value for the all-rater group when the rater in the row is excluded

	mean	**SD**	**alpha w/o the rater**
EN01	4.9	2.4	.97
EN02	6.7	2.3	.97
EN03	5.5	2.0	.96
EN04	6.6	2.2	.97
JP01	6.4	2.1	.97
JP02	6.1	2.3	.97
JP03	5.0	2.4	.97
JP04	5.3	1.6	.97
OT01	3.8	2.2	.97

OT02	6.4	2.4	.97
OT03	6.7	1.7	.96
OT04	7.1	2.9	.97
OT05	4.7	1.5	.97
OT06	5.3	2.3	.97
OT07	5.9	2.8	.97
OT08	4.8	2.1	.97

4.4.3. Fluency

The alpha value of the *fluency* scores was .92 for the *EN* rater group, .89 for the *JP* rater group and .97 for the all-rater group. The values were exactly the same as those of the prosody scores and again implied high consistency. Table 22 shows that dropping any subject in the all-rater group would result in the same alpha value.

Table 22 Mean and standard deviation of the fluency scores of each rater, and the Cronbach's Coefficient Alpha value for the all-rater group when the rater in the row is excluded

	mean	**SD**	**alpha w/o the rater**
EN01	4.9	2.4	.97
EN02	6.7	2.3	.97
EN03	5.5	2.0	.97
EN04	6.6	2.2	.97
JP01	6.4	2.1	.97
JP02	6.1	2.3	.97
JP03	5.0	2.4	.97
JP04	5.3	1.6	.97
OT01	3.8	2.2	.97
OT02	6.4	2.4	.97
OT03	6.7	1.7	.97
OT04	7.1	1.9	.97
OT05	4.7	1.5	.97

OT06	5.3	2.3	.97
OT07	5.9	2.8	.97
OT08	5.2	2.2	.97

4.4.4. Nativelikeness

The alpha value of the *nativelikeness* scores was .95 for the *EN* rater group, .91 for the *JP* rater group and .98 for the all-rater group. The values were higher than those of the other three categories. Again, dropping no rater would increase the alpha value for the all-rater group (Table 23).

Table 23 Mean and standard deviation of the nativelikeness scores of each rater, and the Cronbach's Coefficient Alpha value for the all-rater group when the rater in the row is excluded

	mean	**SD**	**alpha w/o the rater**
EN01	4.3	2.4	.98
EN02	5.4	2.5	.98
EN03	5.2	1.9	.98
EN04	6.8	2.0	.98
JP01	5.8	2.1	.98
JP02	5.9	2.4	.98
JP03	4.6	2.4	.98
JP04	5.3	1.6	.98
OT01	4.2	1.9	.98
OT02	4.6	2.8	.98
OT03	6.1	2.1	.98
OT04	6.8	2.0	.98
OT05	3.7	1.5	.98

OT06	5.3	2.2	.98
OT07	3.9	2.9	.98
OT08	4.2	2.4	.98

4.5. Constituents of nativelikeness

Compared to the other three categories, (*segment, prosody* and *fluency*), *nativelikeness* (more commonly referred to as *accentedness*) is a holistic and intuitive notion (Table 18). This section aims to investigate how the raters' perceptions of the former three categories affect those of *nativelikeness*. The relative influences of the former three on *nativelikeness* were investigated with multiple regression analyses using *lm()* on *R* (Ver 4.1.1.). The dependent variable was the *nativelike* score and the predictor variables were those of *segment, prosody* and *fluency*. To see potential differences in relation to rater languages, separate analyses were conducted for the data of the *EN* raters and *JP* raters. Hence the score data of one category (e.g. *segment*) was a vector containing all scores of the *EN* or *JP* rater group. Assuming multicollinearity, the interactions of the predictor were not integrated into the model. Although multicollinearity was also assumed between the predictor variables, the Variance Inflation Factor was below 10 for all three predictor variables in the models for both rater groups.

The result of the regression analysis for the *EN* rater group yielded significant main effects of all three predictor variables (ps < .001). The adjusted R^2 was .925, indicating that a quite high proportion of the whole variance of the *nativelike* score was explained by the other three scores. The estimated coefficient was 0.516 for the *segment* score, 0.325 for the *prosody* score and 0.160 for the *fluency* score (Table 24). Thus, the evaluation of segmental accuracy had the strongest effect on the judgement of nativelikeness, i.e. the more segmentally accurate a speech was, the more nativelike it was judged.

Table 24 Standardised coefficient of the main effect of segment, prosody and fluency score on the nativelikeness score (the EN rater group)

Factor	*Segment*	*Prosody*	*Fluency*
Statistical significance	$p < .001$	$p < .001$	$p < .001$
Standardised coefficient	0.516	0.325	0.160

The result of the regression analysis for the *JP* rater group also yielded significant main effects of all three predictor variables (*p*s < .001). The adjusted R^2 was .944, which was as high as that of the model for the *EN* rater group. The coefficient was 0.450 for the *segment* score, 0.289 for the *prosody* score and 0.278 for the *fluency* score. As was observed for the *EN* rater group, segmental accuracy contributed the most to the *JP rater* group's judgement of nativelikeness. However, the effect was weaker than that on the *EN* rater group. In contrast, fluency affected the *JP* rater group's judgement of nativelikeness more than did that of the *EN* rater group.

Table 25 Standardised coefficient of the main effect of segment, prosody and fluency score on the nativelikeness score (the JP rater group)

Factor	*Segment*	*Prosody*	*Fluency*
Statistical significance	*p* < .001	*p* < .001	*p* < .001
Standardised coefficient	0.450	0.289	0.278

4.6. Proficiency score used in the remaining analyses

Among the four kinds of scores, the most relevant to the current study is the *prosody* score. The score was based on evaluations of a) lexical stress, b) rhythm and c) wrong insertions and elisions of segments. These three are respectively relevant to the upcoming analyses of a) lexical stress (Chapter 5), foot duration (Chapter 7) and epenthetic vowels (Chapter 6). Although elisions of vowels and consonants (especially vowels) would also hinder manifestations of accurate rhythm, the phenomenon is not as common in *JE* and therefore was not analysed. Since intonation was in the criteria of the *fluency* score, the *prosody* score is expected to be rather independent of intonation, which is not the focus of the current study. By definition, it is unlikely to correlate with the *segment* score. In addition, the results of the analyses in Section 4.5 suggests the score is also independent of nativelikeness.

Since the inter-rater consistency of the scores was quite high (Section 4.4), the mean *prosody* scores of the 48 tokens per speaker (3 sections x 16 raters) were averaged to elicit the grand mean score for each speaker. Figure 7 shows the distribution of scores both for the *EN* group (on the left) and the *JP* groups (on the right). The scores in the *JP* group (n = 183) ranged from 1.771 to 9.479 (mean = 5.628, SD = 1.775) while those in the *EN* group (n = 25) from 8.792 to 9.854 (mean = 9.345, SD = 0.287).

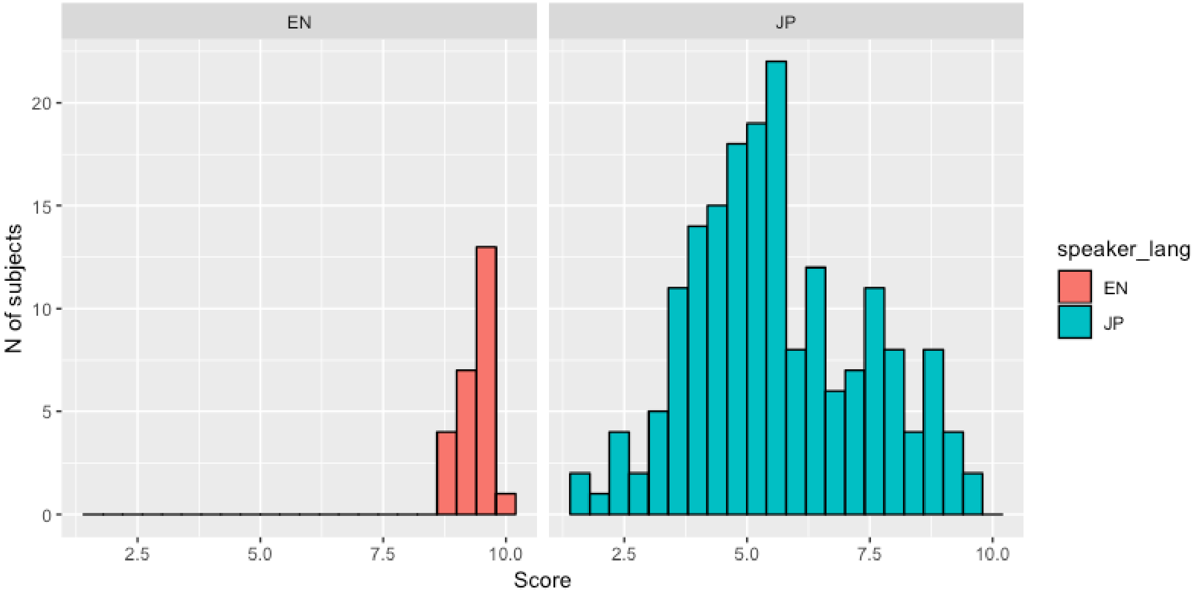


Figure 7 Distribution of the grand mean scores

In order to investigate developmental changes of the phonetic parameters, the *JP* group was divided into two proficiency groups with the median score of 5.438: the more proficient *JP Adv* group (with higher scores than the median, n = 90, mean = 7.033, SD = 1.228) and the less proficient *JP Beg* group (with scores lower than or as low as the mean, n = 93, mean = 4.234, SD = 0.897). Figure 8 shows the distribution of prosody scores in relation to the proficiency groups. The red bars denote the distribution of the *JP Adv* group and the blue bars that of the *JP Beg* group. In the analyses of the remaining chapters (Chapters 5, 6 & 7), the data of the two groups will be compared in relation to the acquisition of the relevant phonetic factor.

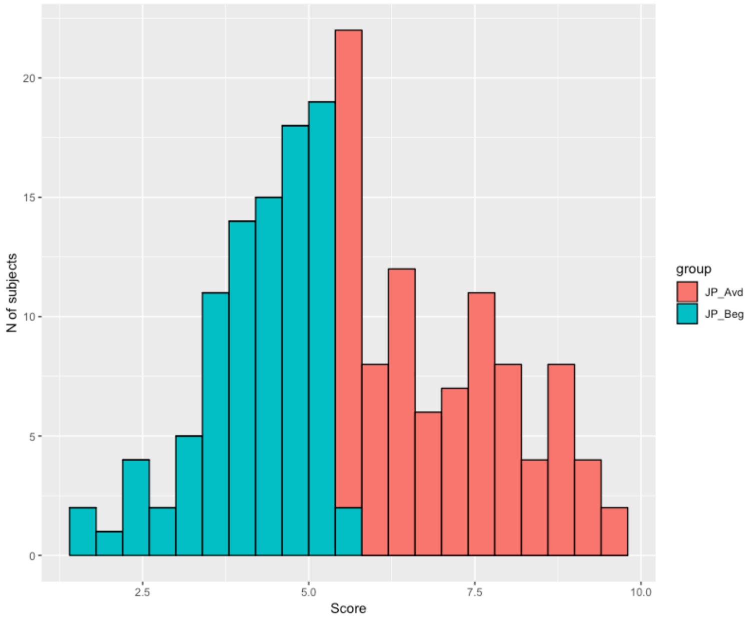


Figure 8 Distributions of prosody scores of the JP Adv and JP Beg groups

4.7. Summary

The aim of this dissertation is to compare phonetic factors of *JE* rhythm in relation to learners' proficiency. In order to investigate the effect of proficiency, this chapter examined proficiency scores which were estimated to correlate to the perception of the *JE* rhythm. In contrast to previous studies, most of which relied on *EN* in evaluating *NNE* speech, the current studies also included *NNE* raters in evaluating *JE* speech. The scores of all four rating categories, i.e. *segment, prosody, fluency* and *nativelikeness* yielded high consistency among different L1 rater groups as well as among individual raters. In the analyses in the remaining chapters, the *JP* subjects will be divided into two proficiency groups, the *JP Adv* and *JP Beg* groups, based on the *prosody* score, which is estimated to be reflecting speech rhythm the most and was not correlated with any of the other three scores.

5. Lexical Stress in Japanese English

5.1. Introduction

The goal of this dissertation is to investigate the phonetic correlates of *JE* foot rhythm. This chapter examines lexical stress, which manifests repetitive beats in English foot rhythm (Chapter 2). Lexical stress is especially relevant in the current study which assumed trochaic, rather than iambic, feet (See Section 2.1) in speech segmentation. With the assumption of the trochaic bias (Cutler, 1989), each stress, or a rhythmic beat, marks the beginning of a new foot. Therefore, manifestations of lexical stresses would affect the listeners' perception of foot rhythms.

It is known that lexically stressed vowels in English usually have greater duration and intensity compared to unstressed vowels (Section 2.2.4). Since lexical stress associates with vowel quality in English, stressed vowels have intrinsically different acoustic characteristics from those of unstressed vowels. In fact, unstressed vowels are intrinsically shorter than stressed vowels because most of the former are realised as a schwa while most of the latter are full vowels, some of which are diphthongs. Therefore, the comparisons of the two groups of vowels will be much more meaningful when investigated in relation to those of the *NNE* speakers, who are assumed to be poorer at manifesting the canonical *NE* rhythm. Schwa has a substantial dialectal variation in American English (Ladefoged 1999). In addition, previous studies reported that distributions of formants (i.e. F1 and F2) of *JE* schwa were notably different from those of *NE* schwa (e.g. Lee et al. 2006). Cruttenden (2014) argues that manifestations of weak forms of English words are difficult even for advanced learners. However, since vowel quality is a segmental feature, it is not investigated in the current study. Also, F0 was not included in the current analysis since it is a relevant parameter in English prosody and not directly much with the foot rhythm, which is manifested with lexical stress (See Section 2.2.4).

Although Japanese uses different phonetic correlates to manifest its phonological contrasts from those in English (See Table 10 in Section 2.4.1), *JP* are known to use the same phonetic parameters as *EN* to manifest the lexical contrasts in English. According to the previous studies (De Jong 2004, Fear et al. 1995, Okobi 2006, Sluijter & van Heuven 1996a), vowel duration and intensity are two primary phonetic correlates of the *NE* lexical stress. These two parameters were shown to be used in *JE,* as well as *NE*, lexical contrasts (Konishi & Kondo 2015, Lee et al. 2006). In addition, the extent of the stress contrasts manifested by intensity was not significantly different between *EN* and *JP.* However, the results of Konishi and Kondo (2015) also implied significantly greater use of the vowel duration to manifest lexical contrasts by *EN* than *JP* of lower proficiency. The difference was not observed between *EN* and *JP* of higher proficiency in Konishi and Kondo (2015) as well as highly proficient *JP*, called "early" and "late bilinguals" in Lee et al. (2006). Table 26 summarises the differences in the use of the two phonetic parameters between *Adv/Beg JE and NE.* The *n.s.* in the cell denotes no statistically significant difference observed in the previous studies. *'Significantly smaller'* means the relevant *JE* relies on the parameter to a smaller extent in manifesting stress contrasts.

Table 26 Degree of stress contrasts made with the phonetic parameter compared to NE

	Intensity	**Duration**
Adv JE vs NE	*n.s.*	*n.s.*
Beg JE vs NE	*n.s.*	Significantly smaller

The effects of other prosodic phenomena on the duration and intensity of vowels might be more dominant than those of lexical stress, especially in *NNE*. For example, overwritten pitch accent and/or sentence focus could hide the potential difference in the extent of durational contrast between *NE* and *JE* vowels*.* That is, the vowel duration could vary depending on whether they carry a pitch accent or sentence focus, which could obscure differences in the duration because of lexical stress. Due to a lack of objective criteria, the *J-AESOP* corpus is not annotated for prosody. It is difficult for annotators to determine whether lexical stress is placed on a given word, and if it is, whether it is placed on the canonical syllable. It is also difficult to judge whether a pitch accent is placed on a given syllable. Hence the current study will not exclude pitch accented and sentence focused words. However, it aims to minimise those effects by analysing a long discourse containing both accented and unaccented words. Likewise, although effects of phrase-final intonations and final-lengthening were not accounted for, such effects are expected to be mitigated by excluding outliers of the data with mean ± 3 standard deviation thresholds.

Based on the above discussion, three hypotheses were formulated prior to the analyses. In accordance with the results of Konishi and Kondo (2015), greater durational contrast is expected between stressed and unstressed vowels of *NE* than between those of *JE* (Hypothesis I). In addition, a developmental change is expected between *Adv JE* and *Beg JE* so that the former manifests greater durational contrast of vowels (Hypothesis II). Regarding intensity contrast of stressed and unstressed vowels, in accordance with the results of Lee et al. (2006) and Konishi and Kondo (2015), no significant difference is expected in relation to proficiency (Hypothesis III).

Hypothesis I: Stress is realised with greater contrasts in vowel duration by *NE* than *JE*.

Hypothesis II: Stress is realised with greater contrasts in vowel duration by *Adv JE* than *Beg JE*.

Hypothesis III: There is no difference between *NE* and *JE* in the extent of stress contrast by vowel intensity.

The next section (Section 5.2) will introduce the methods for extractions of the vowel duration and intensity as well as how those extracted data were normalised for speech rate. In Section 5.3, the above three hypotheses will be tested with the normalised data. Lastly, the results will be interpreted in relation to the results of the previous studies (Section 5.4).

5.2. Data coding and normalisations

The correlates of stress, i.e. vowel duration and intensity, were examined both for *NE* and *JE*. The parameters were extracted from the read speech of the *North Wind and the Sun* in the *J-AESOP* corpus, based on the annotated vowel boundaries. The extracted values were normalised where necessary.

Since the annotation of the *J-AESOP* corpus does not contain prosodic information, lexical stress was defined according to whether a given syllable is canonically stressed or not. The stress placement followed Ladefoged (1999) and the stressed syllables are underlined in the script below. Syllabifications followed the Longman Pronunciation Dictionary (3rd ed.).

*The **North** **Wind** and the **Sun** were dis**put**ing **which** was the **strong**er **when** a **trav**eler **came** a**long** **wrapped** in a **warm** **cloak**. They a**greed** that the **one** who **first** suc**ceed**ed in **mak**ing the **trav**eler **take** his **cloak** off should be con**sid**ered **strong**er than the **oth**er. Then the **North** **Wind** **blew** as **hard** as he **could**. But the **more** he **blew**, the **more** **close**ly did the **trav**eler **fold** his **cloak** a**round** him. And at **last**, the **North** **Wind** **gave** **up** the at**tempt**. **Then** the **Sun** **shone** out **warm**ly. And im**med**iately the **trav**eler **took** **off** his **cloak**. And **so** the **North** **Wind** was o**bliged** to con**fess** that the **Sun** was the **strong**er of the **two**.*

5.2.1. Duration

The vowel duration was calculated by subtracting the annotated left boundary (the starting point) of each vowel segment from the right boundary (the end point). Epenthetic vowels (n = 2,142) were excluded from the data. To alleviate fluctuating speech rates within the same speaker, the vowel duration was normalised with the word duration. Since some words in *the North Wind and the Sun* were monosyllabic, the vowel duration was normalised with the total duration of the five words (i.e. the word containing the vowel and two words before and after). Equation 1 shows the formula to calculate the normalised vowel duration (*ND*). The numerator is the raw vowel duration (*VD*) and the denominator was the sum of the duration of each of the five words (*WD*) divided by the number of vowels in the words (*nV*). The word duration (*WD*) was defined to be the sum of the duration of all vowels within the word, including the epenthetic vowels. If there was a pause within the five words, ones before the preceding pause and ones after the following pause were excluded from the normalisation[13]. Out of the 29,175 tokens of the normalised vowel duration, outliers were excluded with mean ± 3 standard deviation thresholds, resulting in 305 vowels excluded from the data. Altogether, 11,886 stressed vowels and 16,984 unstressed vowels were analysed.

Equation 1 Normalised vowel duration for the analysis

$$ND = \frac{VD}{WD/nV}$$

[13] For example, if there was a pause just before the word with the target vowel but no other pause within the five words, three words were used for the normalisation.

5.2.2. Intensity

Vowel intensity was defined to be the maximum intensity within the vowel interval. On *Praat,* intensity objects of the sound files were first created, and each vowel intensity was calculated using the function *"Get maximum."* The interpolation method was "*parabolic*." See the formula on the official homepage of *Praat*[14].

In order to alleviate the effect of prosody on the vowel intensity (e.g. vowels with a pitch accent are expected to have greater intensity), they were normalised with the word intensity so that the normalised intensity was calculated by dividing the vowel intensity by the mean intensity of the word containing the vowel. Equation 2 shows the formula to calculate the normalised vowel intensity (*NI*). The numerator is the raw vowel intensity (*VI*) and the denominator is the mean word intensity (*WI*) as calculated by the *Praat* function *"Get mean."* The averaging method was *energy*. See the formula on the official homepage of *Praat*. As for the vowel duration, epenthetic vowels (n = 2,142) were excluded from the data. Out of the 29,175 data of the normalised vowel intensity, outliers were excluded with mean ± 3 standard deviation thresholds, resulting in 420 vowels being excluded. Altogether, 11,963 stressed vowels and 16,792 unstressed vowels were analysed.

Equation 2 Normalised intensity for the analyses

$$NI = \frac{VI}{WI}$$

[14] https://www.fon.hum.uva.nl/praat/

5.3. Analyses

5.3.1. Stress and vowel duration

To investigate inter-group differences in the vowel duration in relation to stress contrast, a mixed analysis of variance (hereafter ‘mixed ANOVA’) was conducted with *anova_test()* in the *rstatix* package (Ver 0.7.0) on *R* (Ver 4.1.1). The predictor variables were a fixed effect of lexical stress (i.e. whether the vowel was canonically stressed (1) or not (0); a within-subject factor), that of the speaker’s language group (*EN*, *JP Adv* or *JP Beg*; a between-subject factor) as well as their interaction. The dependent variable was the normalised vowel duration (See Section 5.2.1).

The result showed significant main effects both of stress ($F(1, 205) = 1175.17$, $p < .\text{-}001$) and the speaker’s language group ($F(2, 205) = 26.12$, $p < .001$). The interaction was also significant ($F(2, 205) = 79.72$, $p < .\text{-}001$). Figure 9 shows the distributions of the normalised vowel duration in relation to the presence or absence of lexical stress. The left pane shows the distributions of the unstressed vowels (n = 16,984) and the right pane those of the stressed vowels (n = 11,886). In each pane, the red, green and blue boxes respectively show the distributions of the *NE*, *Adv JE* and *Beg JE* feet. The post-hoc pairwise comparisons with Bonferroni adjustments (hereafter calculated with *pairwise.t.test()* on *R*) showed that unstressed vowels of the *EN* group had significantly smaller duration than those of the *JP Adv* group, which also had significantly smaller duration than those of the *JP Beg* group (both *p*s <.001). On the other hand, there was no significant difference among the stressed vowels of the three groups (*p*s >.9) The result indicates *NE* vowels had the greatest durational contrast for stress, followed by those of the *Adv JE*. The *JP Beg* group manifested the smallest contrast.

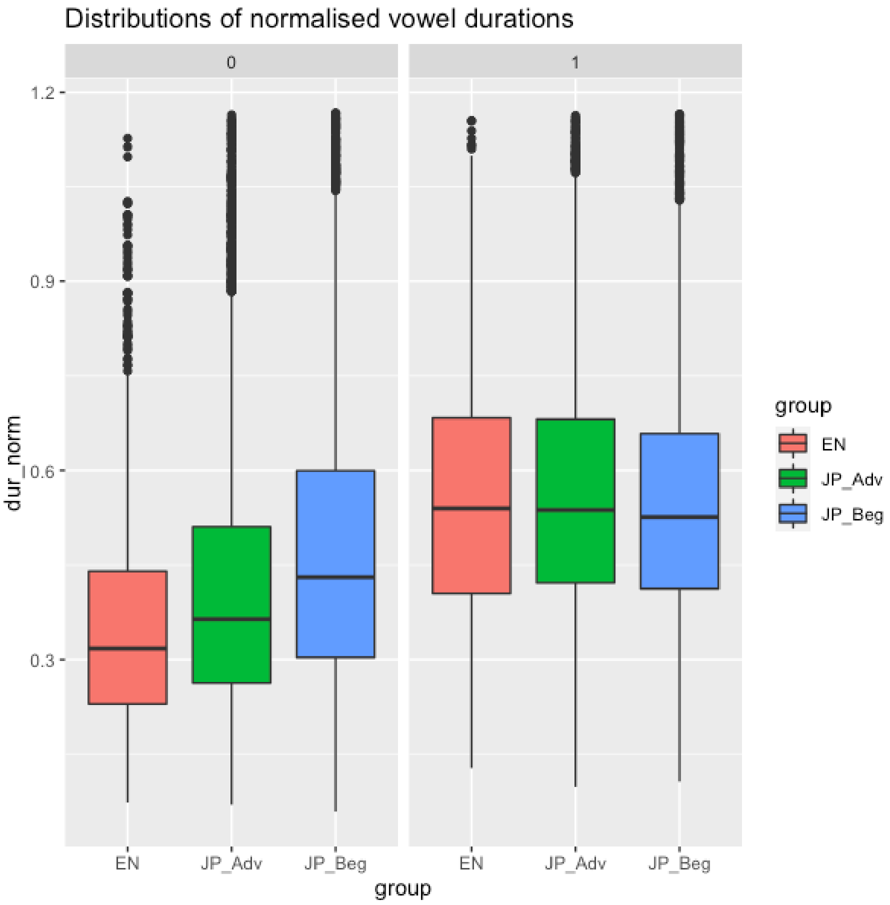


Figure 9 Normalised duration of stressed and unstressed vowels

5.3.2. Stress and vowel intensity

As for duration, a mixed ANOVA was conducted. Again, the predictor variables were a fixed effect of stress (i.e. whether the vowel was canonically stressed (1) or not (0); a within-subject factor), that of the speaker's language group (*EN*, *JP Adv* or *JP Beg*; a between-subject factor) as well as their interaction.

The result showed a significant main effect of stress ($F(1, 205) = 885.14$, $p < .001$) and its significant interaction with speaker's language group ($F(2, 205) = 41.58$, $p < .001$). However, the main effect of speaker's language group was not significant ($F(2, 205) = 0.97$, $p > .3$). Figure 10 shows the distributions of the normalised vowel intensity in relation to the presence or absence of lexical stress. The left pane shows the distributions of the unstressed vowels (n = 16,792) and the right pane those of the stressed vowels (n = 11,963). In each pane, the red, green and blue boxes respectively show the distributions of the *NE*, *Adv JE* and *Beg JE* feet. The post-hoc pairwise comparisons showed, as for the vowel duration, the unstressed vowels of the *EN* group had significantly smaller intensity than those of the *JP Adv* group, which had significantly smaller intensity than those of the *JP Beg* group (both *p*s < .001). On the other hand, stressed vowels of the *EN* group had significantly greater intensity than those of the *JP Adv* group and the *JP Beg* group (both *p*s < .001). No significant difference existed between the stressed vowels of the latter two groups ($p > .05$). The result shows stress is realised with the greatest intensity contrast by the *EN* group, followed by the *JP Adv* group.

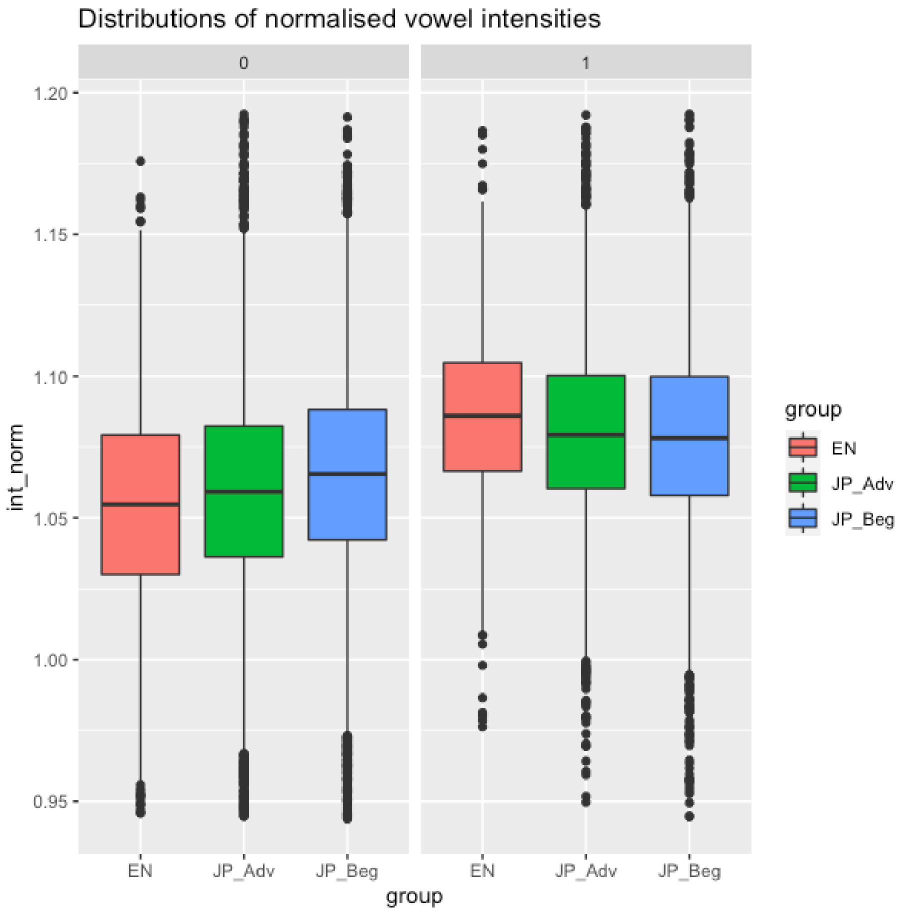


Figure 10 Normalised vowel intensity of stressed and unstressed vowels

5.4. Discussion

The results of the analyses indicate that the greatest stress contrast was manifested by the *EN* group with the difference in the vowel duration, which supports Hypothesis I (Table 27). The difference was observed for unstressed vowels rather than stressed vowels. The *NE* unstressed vowels had significantly smaller duration than those of the *Adv JE* and the *Beg JE* (*p*s <.001), meaning that greater durational reduction was observed for the unstressed vowels of the *NE* speech. On the other hand, the duration of the *NE* stressed vowels were not significantly different from either of the *JE* (*p*s >.9). The result contrasts with that of Lee et al. (2006), in which no significant difference was found between *NE* and *JE* in the manifestation of stress contrast with the vowel duration. However, as explained in Section 5.1, the *JP* groups examined in Lee et al. (2006) were much more proficient than the *JP* subjects in the current study. The subjects in their study were early Japanese-English bilinguals, whose mean length of residence (LOR) in the United States was 20 years, and late Japanese-English bilinguals with a mean LOR of 10 years. In contrast, most of the *JP* subjects in the current study had learned English in Japanese schools, with no experience living in any English-speaking countries. Therefore, the result of the analyses of *JP Adv* in the current study could represent a stage of the developmental change rather than the ultimate attainment. In other words, the result of the current study does not imply necessarily that nativelike manifestations of English stress is impossible for *JP*.

Table 27 The hypotheses and results of the analysis of the NE and JE lexical stress

Hypothesis I	Stress is realised with greater contrast in the vowel duration by *NE* than *JE*.	**Supported**
Hypothesis II	Stress is realised with greater contrasts in the vowel duration by *Adv JE* than *Beg JE*.	**Supported**
Hypothesis III	There is no difference between *NE* and *JE* in the extent of stress contrast by intensity.	**Not supported**

In addition, the greater stress contrast was manifested with the difference in the vowel duration in the *Adv JE* than the *Beg JE*, supporting Hypothesis II. Again, the difference was observed for unstressed rather than stressed vowels. In contrast to the duration of stressed vowels which had no significant difference between the two proficiency groups (p > .9), unstressed vowels in the *Adv JE* had significantly smaller duration than those in the *Beg JE* (p < .001), suggesting a greater durational reduction in the *Adv JE* speech. This contrasts with the result of Konishi and Kondo (2015), in which the difference was found on the stressed vowels. However, vowel duration in their study was converted to z-scores to alleviate individual differences so that the distributions of stressed and unstressed vowels might have been different from the original ones. Although the vowel duration in the current study was normalised with the word duration to alleviate the effect of fluctuations of speech rate, it was not normalised for individual differences in speech rates.

Contrary to Hypothesis III, as well as the result of Lee et al. (2006), the result of the analysis showed a significant difference in intensity between the *NE* and *JE*. Lee et al. (2006) observed nativelike manifestations of stress contrast with intensity even in the *Beg JE.* In the current study, the *NE* unstressed vowels had significantly smaller intensity than that of the *Adv JE* and *Beg JE* (both ps < .001). On the other hand, the *NE* stressed vowels had significantly greater intensity than those of the *JP* groups (both *p*s < .001). Again, the difference from the result of Lee et al. (2006) is assumed to be due to the much lower proficiencies of the *JP* groups in the current study.

In addition, differences in intensity were also observed in relation to the speakers' proficiency. As it was observed for the vowel duration, the significant differences in the intensity were mainly observed for the unstressed vowels (*p* < .001) rather than between the stressed vowels (*p* > .05). Again, the implication is the greater difficulty for lower-proficiency learners to reduce unstressed vowels. This contrasts with the result of Konishi and Kondo (2015), which found no significant difference in the intensity of unstressed vowels as well as stressed vowels. However, unlike the current analysis, the intensity data for their analysis were converted to z-scores, which could have resulted in the contrastive results.

Altogether, the results indicate the developmental change in the manifestation of *JE* lexical stress is mainly observed in the reduction of unstressed vowels. In addition to the results of the previous studies that suggests the reduction of unstressed vowels was difficult for *JP* in terms of their qualities, i.e. the first and second formants (e.g. Kondo 2000), the current study also observed the difficulty in relation to the speech rhythm, i.e. the duration and intensity of the vowels. Although the *Adv JP* group managed to manifest more nativelike stress contrasts with duration and intensity, the realisations were still significantly different from those of the *EN* group*.*

5.5. Summary

As one of the phonetic correlates of *JE* rhythm, this section investigated lexical stress, i.e. acoustic realisation of the repetitive beats in English foot rhythm. Two acoustic correlates of lexical stress, the vowel duration and intensity, were compared between *EN, JP Adv* and *JP Beg*. Results of previous studies suggest stressed vowels are realised with greater duration and intensity than unstressed vowels. Although some previous studies also investigated vowel F0, it was not investigated in the current study, following the convention of metrical phonology that assumes F0 associates with pitch accent but not lexical stress (Section 2.2.4).

Since Japanese uses different phonetic parameters to manifest its phonological and prosodic contrasts than those in English, L1 interference was expected in *JE*. In contrast to the results of previous studies which found no significant difference in the use of the vowel duration and intensity in *Adv JE* compared to those in *NE,* the results of the current study indicated significantly greater duration of unstressed vowels in the *Adv JE* and *Beg JE* than in the *NE*, i.e. smaller reductions were manifested in the unstressed vowels of the lower-proficiency speech. The significant difference between the *NE* and *JE* in the use of intensity was also primarily observed for the unstressed vowels, i.e. a lack of reduction, rather than for the stressed vowels. There were also developmental changes so that *JP Adv* manifested a greater extent of vowel reduction than *JP Beg* in terms of both duration and intensity. Since the *JP* subjects of the current study had a great variation in proficiency, the results of the analyses could have been different from those in Lee et al. (2006), all of whose *JP* subjects were Japanese-English bilinguals. The result of the analysis of the vowel duration in the current section will be integrated into the model of the analysis of the foot duration in Chapter 7.

6. Epenthetic Vowels in Japanese English

6.1. Introduction

This dissertation aims to investigate phonetic correlates of *JE* rhythm. Chapter 5 examined lexical stress, the acoustic realisation of rhythmic beats. Based on the results of the analyses, *JE* appears to manifest English stress with smaller contrasts in vowel duration and intensity. The current chapter aims to investigate another estimated correlate of *JE* rhythm, epenthetic vowels, which is estimated to be a phonetic obstacle to the manifestations of isochronous rhythm (Section 2.4.2). The dissertation assumes all feet to be trochaic and a foot may consist of more than one unstressed syllable following a stressed syllable (Section 2.2.1). Because the addition of an epenthetic vowel within a foot unnecessarily increases its duration, *JE* speech with more frequent vowel epenthesis would have greater a variance in the foot duration, making the speech sound less isochronous. As reported by Yazawa et al. (2015), there is a negative correlation between the frequency of vowel epenthesis and speakers' proficiency.

As explained in Section 2.4.2, vowel epenthesis is a pervasive phenomenon in *JE* (e.g. Dupoux et al. 1999, Masuda & Arai 2010, Mazuka et al. 2011, Tajima et al. 2000). This is due to much more complex consonant clusters in English than Japanese, as well as coda consonants, which are frequent in English but not allowed in Japanese except for moraic nasals. Previous studies were conducted on many aspects of vowel epenthesis in Japanese: the environments and conditions for epenthesis (Davidson 2011, Davidson et al. 2015, Dupoux et al. 1999, Hall 2003, Masuda & Arai 2010, Tajima et al. 2000), frequency of epenthesis (Yazawa et al. 2015), qualities of the epenthetic vowels (Funatsu et al. 2008, Shibuya & Erickson 2010, Yazawa et al. 2015). However, not much has been studied about the phonetic realisations of epenthetic vowels. Although the frequency of vowel epenthesis is expected to negatively correlate with the nativelike manifestation of English rhythm, there has been no study that actually investigated realisations of epenthetic vowels in relation to *JE* rhythm.

As with vowels with and without lexical accent, the primary correlates of epenthetic vowels to *JE* rhythm are expected to be duration and intensity. As either (or both) of the two factors becomes greater, the epenthetic vowel is more likely to be perceived as a vowel. Hence the two phonetic parameters of epenthetic vowels were investigated. The result of the analysis of lexical stress (Chapter 5) suggests that *JE* has a developmental change in the extent of stress contrasts made with the vowel duration and intensity. The difference was observed primarily for unstressed vowels rather than stressed vowels. If epenthetic vowels produced by *Adv JE* are less conspicuous in terms of their duration and intensity, the developmental change would be observed on the phonetic factors.

Accordingly, three hypotheses and one research question were formulated. Firstly, it can be hypothesised that vowel epenthesis is less frequent in *Adv JE* than *Beg JE* (Hypothesis I), which replicates the result of Yazawa et al. (2015). In addition, since duration and intensity were better controlled by *JP Adv* to manifest unstressed vowels (Section 5.3), it can also be hypothesised that epenthetic vowels of *Adv JE* have smaller duration than *Beg JE* (Hypothesis II) and smaller intensity (Hypothesis III).

Hypothesis I: Vowel epenthesis is less frequent in *Adv JE* than *Beg JE.*

Hypothesis II: Epenthetic vowels have smaller duration in *Adv JE* than *Beg JE*.

Hypothesis III: Epenthetic vowels have smaller intensity in *Adv JE* than *Beg JE*.

Firstly, Section 6.2 will detail the data extractions and normalisations. Section 6.3 will test the above hypotheses. Lastly, Section 6.4 will interpret the results in relation to those of the previous studies, as well as those in Chapter 5.

6.2. Data coding and normalisations

6.2.1. Duration

Since epenthetic vowels were annotated in the *J AESOP* corpus (Section 3.5), the number of epenthetic vowels was counted for each speaker and the duration and intensity of each epenthetic vowel were extracted. Since the difference in speech rates were assumed between *JP Adv* and *JP Beg*, the data were also extracted from other vowels (i.e. lexically stressed and unstressed vowels) for the sake of statistical comparison.

The duration of vowels, including epenthetic vowels, was calculated by subtracting the left boundary (i.e. the starting point) of each epenthetic vowel from the right boundary (i.e. the end point). As in Section 5.2.1, the duration of each vowel, whether epenthetic or not, was normalised to address fluctuating speech rates. Equation 3 shows the formula to calculate the normalised vowel duration (*ND*). The numerator is the raw vowel duration (*VD*) and the denominator was the sum of the duration of all vowels within the five words (*WD*) divided by the number of vowels in the five words (nV).

Equation 3 Normalised duration of the vowels

$$ND = \frac{VD}{WD/nV}$$

6.2.2. Intensity

As in Section 5.2.2, vowel intensity was defined to be the maximum intensity within the vowel interval. Again, the vowel intensity was extracted using the function *"Get maximum."* of *Praat*[15]. To alleviate the effect of prosody on the vowel intensity (e.g. vowels with a pitch accent are expected to have greater intensity), the extracted intensity was normalised with the word intensity. Each vowel intensity was divided by the mean intensity of the word containing the vowel (as calculated by the *Praat* function *Get mean*[16]). Equation 4 shows the formula for the normalisation where *VI* denotes the raw vowel intensity, *WI* the mean word intensity and *NI* the normalised vowel intensity.

Equation 4 Normalised intensity of the epenthetic vowels

$$NI = \frac{VI}{WI}$$

[15] The interpolation method was *parabolic*.

[16] The averaging method was *energy*.

6.3. Analyses

6.3.1. Frequencies of epenthesis

The *EN* group (n = 25) had considerably fewer epenthetic vowels than the *JP Adv* group (n =90) and the *JP Beg* group (n= 93). The mean number of epenthetic vowels was 0.880 per speaker in the *EN* group, with a standard deviation of 0.971 (Figure 11). Out of 25 subjects, 11 had no epenthesis. The mean number of epenthetic vowels per speaker was 6.244 for the *JP Adv* group, with a standard deviation of 5.349. Out of the 90 subjects, 7 had no epenthesis. In contrast, all the 93 subjects in the *JP Beg* group had at least two epenthetic vowels. The mean number of epenthetic vowels was 17.548 and the standard deviation was 10.607.

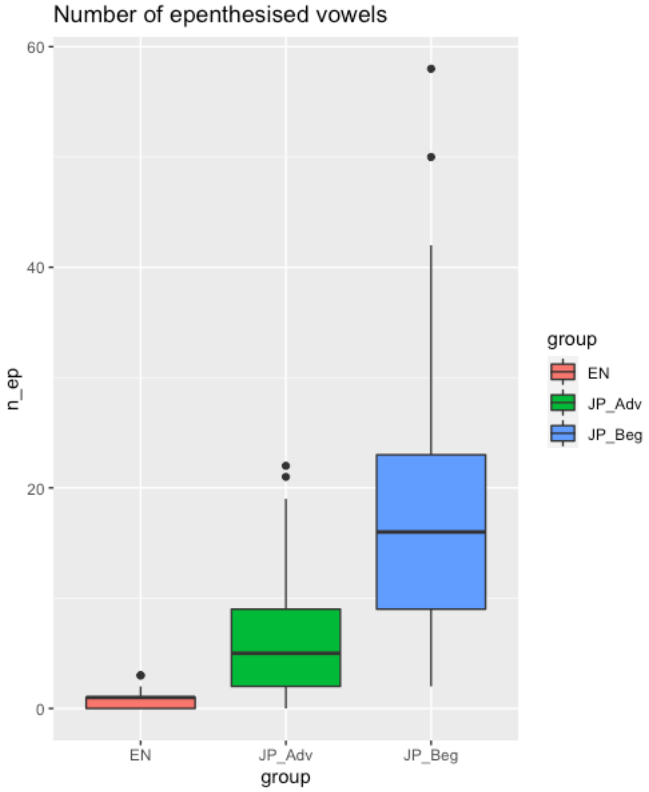


Figure 11 The number of epenthetic vowels for each speaker

Although the differences were seemingly obvious, the mean frequency of vowel epenthesis, i.e. the number of epenthetic vowels, in each speaker's utterance was statistically compared among the three speaker groups. To investigate statistical difference, a one-way analysis of variance was conducted setting the speaker group as the predictor variable and the number of epenthetic vowels as the dependent variable.

The result showed a significant main effect of the speaker group ($F(2, 205) = 67.9$, $p < .001$). Based on the result of the post-hoc comparisons, the numbers of epenthetic vowels were significantly smaller in the *EN* group than the *JP Adv* group ($p < .01$) and in the *JP Adv* group than the *JP Beg* group ($p < .001$). There was also a significant difference between the two *JP* groups ($p < .001$) so that epenthesis was statistically less frequent in *Adv JE.*

6.3.2. Duration of epenthetic vowels

The analysis in this section, as well as Section 6.3.3, will exclude speakers with no epenthesis. Since 11 of the 25 speakers in the *EN* group had no epenthesis and 8 out of the remaining 14 had only one epenthesis, the *EN* group (n = 3,505) was altogether excluded from the analysis. 7 out of the 90 subjects in the *JP Adv* group, who had no epenthesis, were excluded from the data as well. In contrast, all the 93 subjects in the *JP Beg* group epenthesised at least two vowels. Out of the 27,812 vowels, outliers were excluded with the mean ± 3 standard deviation thresholds, resulting in 296 vowels being excluded. Altogether, 10,414 stressed vowels, 14,982 unstressed vowels and 2,120 epenthetic vowels were analysed.

The duration of epenthetic vowels was statistically compared between the *JP Adv* and the *JP Beg* group. Since a difference in speech rate was expected between the two groups, the vowel duration of the *JP Adv* group was expected to be smaller than those of the *JP Beg* group. Therefore, the vowel duration was compared in relation to that of other vowels, i.e. stressed vowels and unstressed vowels. A mixed ANOVA was conducted with *anova_test()* on *R* (Ver 4.1.1). The predictor variables were a fixed effect of vowel type (*stressed, unstressed* or; a within-subject factor), that of the speaker's language group (*JP Adv* or *JP Beg*; a between-subject factor) and their interaction. The dependent variable was the normalised vowel duration (See Section 6.2).

The result showed significant main effects of both vowel type ($F(2, 346) = 1420.56$, $p < .001$) and the speaker's language group ($F(1, 173) = 15.60$, $p < .001$). The interaction was also significant ($F(2, 346) = 22.19$, $p < .001$). The post-hoc pairwise comparisons showed that epenthetic vowels had significantly smaller duration than unstressed, as well as stressed, vowels in both groups (*p*s < .001). There was no significant group difference in the duration of epenthetic vowels ($p > .03$) (Figure 12). See Section 5.3.1 for the statistical difference in unstressed vowels between the two groups and Section 5.4 for its interpretation.

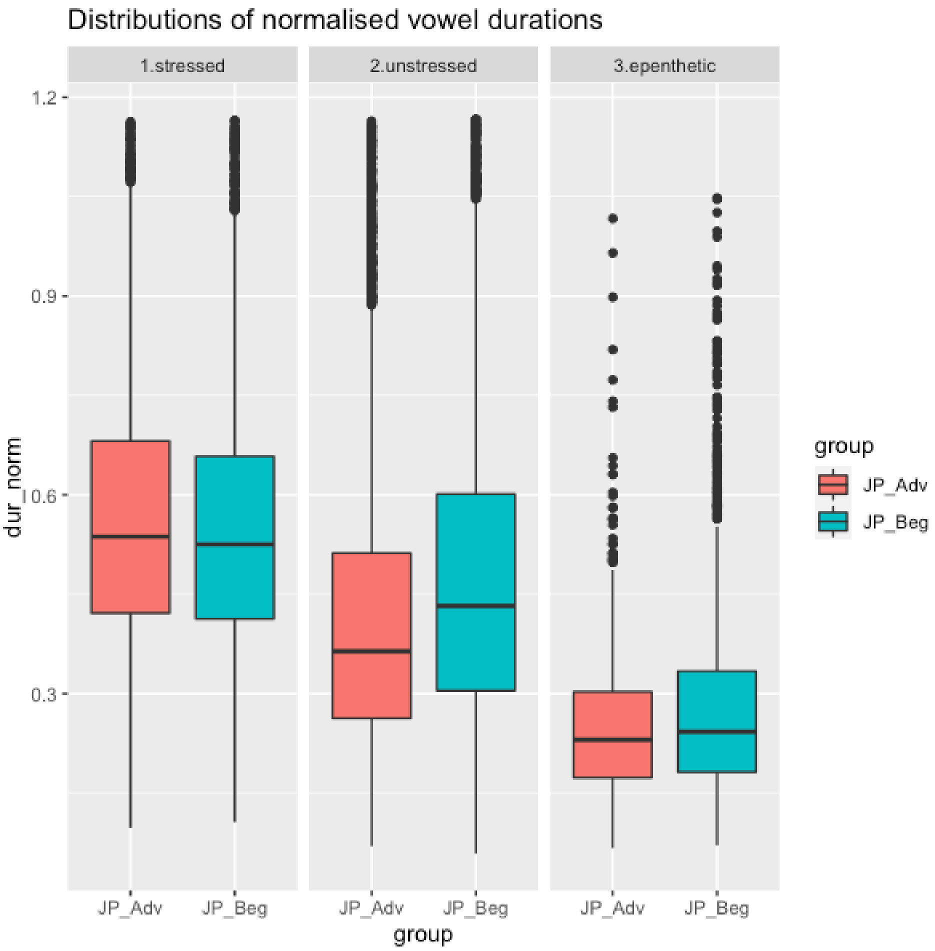


Figure 12 Normalised vowel duration in relation to the vowel type

6.3.3. The intensity of epenthetic vowels

As in Section 6.3.2, the *EN* group (n = 3,505) was excluded from the data for the analysis. Speakers with no epenthesis (7 out of the 90 subjects in the *JP Adv* group and none out of the 93 in the *JP Beg* group) were excluded as well. Out of the 27,812 vowels, outliers were further excluded with the mean ± 3 standard deviation thresholds, excluding 542 vowels. This resulted in 3 subjects in the *JP Adv* group and 3 subjects in the *JP Beg* group being excluded from the data. Altogether, 10,514 stressed vowels, 14,990 unstressed vowels and 1,766 epenthetic vowels were analysed.

As for the duration, the intensity of epenthetic vowels was statistically compared between the *JP Adv* and the *JP Beg* group and in relation to other vowels, i.e. stressed and unstressed vowels. A mixed ANOVA was conducted. The predictor variables were a fixed effect of the vowel type (*stressed, unstressed* or *epenthetic*; a within-subject factor), that of the speaker's language group (*JP Adv* or *JP Beg*; a between-subject factor), and their interaction. The dependent variable was normalised vowel intensity.

The result showed significant main effects of both vowel type ($F(2, 336) = 1379.07$, $p < .001$) and the speaker's language group ($F(1, 168) = 23.45$, $p < .001$). The interaction was also significant ($F(2, 336) = 23.15$, $p < .001$). The post-hoc pairwise comparisons showed, despite the group difference in the intensity of unstressed vowels (See Section 5.3.2, as well as Section 5.4 for the interpretation), the intensity of the epenthetic vowels was significantly smaller than that of the unstressed vowels for both groups (*p*s < .001). In addition, the epenthetic vowels of the *JP Adv* group had significantly smaller intensity than those of the *JP Beg* group ($p < .001$) (Figure 13).

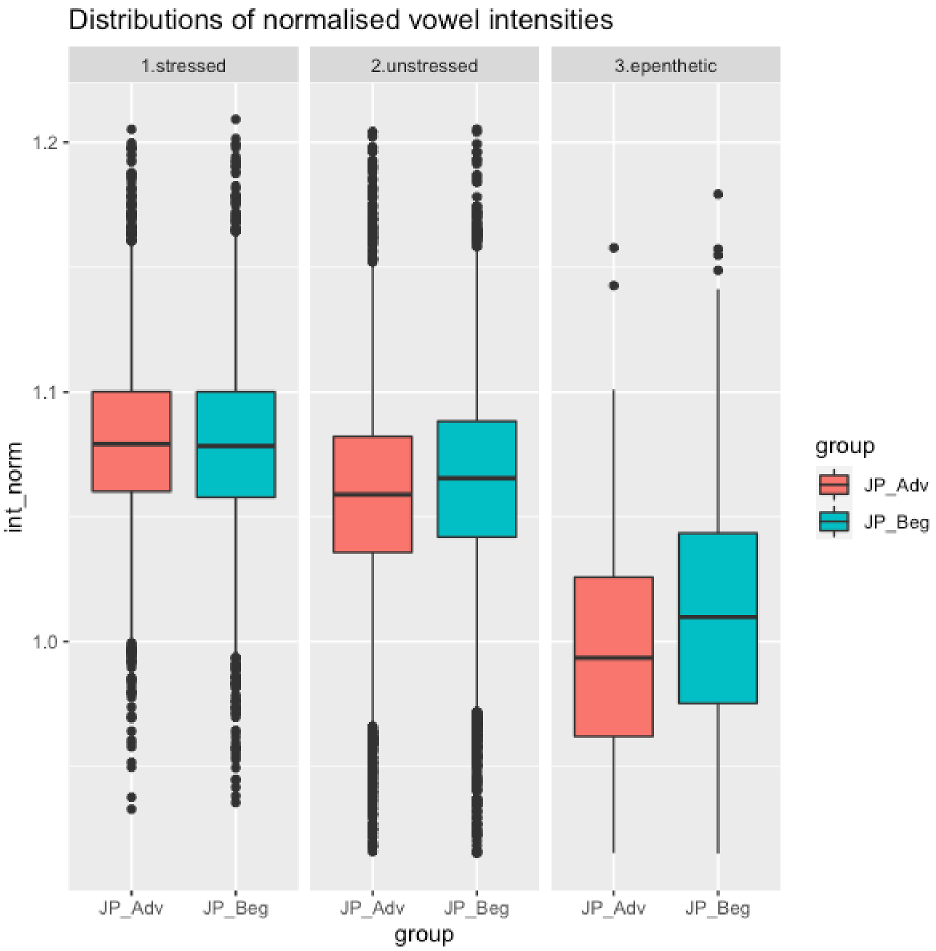


Figure 13 Normalised vowel intensity in relation to the vowel type

6.4. Discussion

The result of the analysis of frequency replicated the result of Yazawa et al. (2015), supporting Hypothesis I (Table 28). That is, a developmental change was observed for the frequency of vowel epenthesis; the *JP Adv* group had statistically less frequent epenthesis (mean = 6.244, SD = 5.349) compared to the *JP Beg* group (mean = 17.548, SD = 10.607). 7 out of the 90 subjects in the *JP Adv* group had no epenthesis while all 93 subjects in the *JP Beg* group had epenthesis more than once. On the other hand, vowel epenthesis by the *JP Adv* group was still significantly more frequent than those by the *EN* group (mean = 0.880, SD = 0.971).

Table 28 The hypotheses and results of the analysis of the JE vowel epenthesis

Hypothesis I	Vowel epenthesis is less frequent in *Adv JE* than *Beg JE.*	**Supported**
Hypothesis II	Epenthetic vowels have smaller duration in *Adv JE* than *Beg JE*.	**Not supported**
Hypothesis III	Epenthetic vowels have smaller intensity in *Adv JE* than *Beg JE*.	**Supported**

The analysis of the duration of epenthetic vowels did not support Hypothesis II. Unlike greater durational reduction which was observed for lexically unstressed vowels by the *JP Adv* group, the duration of epenthetic vowels showed no such difference in relation to the proficiency group. The result is consistent with that of Yazawa et al. (2015), which observed no effect of learner's proficiency in the quality (i.e. first and second formants) of epenthetic vowels. Considering that intrinsic vowel duration depends on the vowel phoneme (Fairbanks & House, 1950; Fairbanks, 1953; Whalen & Levitt,1995), no statistical difference would be expected in the duration of epenthetic vowels unless differences in vowel qualities were observed between *JP Adv* and *JP Beg*. This suggests that the developmental change of *JE* rhythm in relation to vowel epenthesis is on its frequency rather than its actual phonetic realisations.

Contrastively, the analysis of the intensity of epenthetic vowels did show an effect of the speakers' proficiency, supporting Hypothesis III. The intensity of epenthetic vowels was statistically smaller in the *Adv JE* than in the *Beg JE*. Unlike the results of the analysis of the duration, this suggests some effect of proficiency on the realisation of epenthetic vowels.

Furthermore, the results of the analyses of the duration and intensity suggest epenthetic vowels in *JE* are statistically less conspicuous than lexically unstressed vowels, as well as stressed vowels. The produced epenthetic vowels had significantly smaller duration and intensity compared to lexically unstressed vowels. This means it is quite unlikely for an epenthetic vowel to be perceived as a rhythmic beat; hence it does not affect the number of feet. Instead, epenthetic vowels are expected to hinder isochrony in *JE* foot rhythm by adding extra duration to feet containing them. However, the added duration is much smaller than that of an unstressed syllable. Although the statistical difference was found between the intensity of the epenthetic vowels in the *Adv JE* and that of the epenthetic vowels in the *Beg JE*, its effect on the foot duration is expected to be limited compared to that of the duration of epenthetic vowels.

6.5. Summary

This chapter investigated epenthetic vowels, a common phenomenon in *JE*, which is estimated to hinder the manifestations of isochronous foot rhythm. Epenthetic vowels are expected to unnecessarily increase the foot duration, adding greater variance to the foot duration and making the speech sound less isochronous. Based on the results of previous studies and those of Chapter 5 in this dissertation, epenthetic vowels were investigated in relation to their frequency and as well as their actual realisations, i.e. in terms of the vowel duration and vowel intensity.

As a result, the developmental change in *JE* epenthetic vowels was mainly observed on their frequency rather than their realisations. While the *JP Adv* group epenthesised much fewer vowels than the *JP Beg* group, the duration of actually epenthesised vowels showed no effect of proficiency. Although there was a statistical difference in the intensity of epenthetic vowels between the *Adv JE* and the *Beg JE*, their effect on the foot duration is expected to be small. In addition, the epenthetic vowels in *JE* were statistically different from lexically unstressed vowels. They had significantly smaller duration and intensity than lexically unstressed vowels. Therefore, the main effect of epenthetic vowels on the *JE* rhythm does not seem to be on the rhythmic beat but on the greater duration of feet.

7. Foot in Japanese English

7.1. Introduction

As the main goal of this dissertation, the current chapter will investigate *JE* foot rhythm in relation to speakers' proficiency. The results of the analyses in Chapter 5 suggest greater stress contrasts made by more proficient speakers. With the assumption of trochaic bias (Cutler 1989; See Section 2.1), greater stress contrasts are expected to result in easier perceptions of the rhythmic beats, which mark the beginning of each foot. In addition, Chapter 6 examined vowel epenthesis, which is expected to negatively contribute to the isochronous rhythm in *JE* feet by adding extra duration to some feet. The result showed more frequent vowel epenthesis in less proficient *JE*. Based on these results, more proficient speakers are expected to manifest better foot rhythm.

Despite the putative 'mental' isochrony, the physical isochrony of the *NE* feet has not been empirically supported. However, several phonetic and phonological phenomena suggest native speakers' preference for isochronous feet. One of the most well-investigated phonetic phenomena is compensatory shortening (Fowler 1977, 1981, Huggins 1973, Lehiste 1972, Rakerd et al. 1987). In the previous studies, a negative correlation was observed between the number of syllables in the foot and the duration of each syllable. The compensations of the duration were observed between stressed and unstressed syllables rather than among unstressed syllables. That is, foot-internal stressed syllables are shortened more when they are followed by a greater number of unstressed syllables. In addition, perceptual studies by Lehiste (1975, 1977) support native listeners' preference of isochronous feet.

Such regulation of the duration has been observed on Japanese mora. In Japanese, the consonant and vowel duration is said to be adjusted so that the duration of an utterance is proportional to the number of morae (Bradlow et al. 1995, Port et al. 1987; See Section 2.3.2). However, since the speech is proportional to mora, i.e. the number of vowels, the L1 Japanese rhythm is expected to transfer on the manifestations of English foot, which is regulated by foot, a unit that could consist of more than one vowel.

Although some previous studies have been conducted on the *JE* foot rhythm, the methods and analyses are rather unsatisfactory. Mochizuki-Sudo and Kiritani (1991) conducted both production and perception studies on *JE* foot rhythm. The result of their production study implies more isochronous foot rhythm in more proficient speech. That is *NE* had a smaller proportional increase of the duration in relation to the number of foot-internal syllables than *Adv JE,* which also had a smaller proportional increase of the duration than *Beg JE*. However, compensatory shortening, a characteristic phenomenon observed for foot-internal stressed syllables, was not consistently observed in their *NE.* More importantly, the observed differences in the foot duration were not statistically tested and it is not clear whether the differences were statistically significant. Mori et al. (2014) conducted a statistical analysis of foot duration. However, they examined the mean and variance of each foot, i.e. how each speaker is consistent in the produced duration of a given foot interval across tokens of utterances. Since isochrony is the degree of variance of the duration across different feet, rather than the variance within the same foot, the result cannot so much be interpreted in relation to isochrony.

Based on the results of the previous studies on the manifestations of *NE* and *JE* foot rhythm (Fowler 1977, 1981, Huggins 1973, Lehiste 1972, Mochizuki-Sudo & Kiritani 1991, Rakerd et al. 1987), four hypotheses were formulated. Firstly, *NE* should have a smaller variance of the foot duration compared to *JE,* because the Japanese rhythm is not regulated by lexical stress but mora, i.e. the foot duration should be more proportional to the number of syllables (Hypothesis I; See Section 2.3). In addition, since *JP Adv* manifests a better English rhythm in terms of lexical stress (Chapter 5) and vowel epenthesis (Chapter 6), it should have a smaller variation of the foot duration than *Beg JE* (Hypothesis II). Furthermore, as observed in Mochizuki-Sudo and Kiritani (1991), compensatory shortening should be observed more for the *NE* than *JE* of the current data set (Hypothesis III) and more in *Adv JE* than *Beg JE* (Hypothesis IV).

Hypothesis I: The duration of the *NE* feet has a smaller variation than that of the *JE* feet.

Hypothesis II: The duration of the *Adv JE* feet has a smaller variation than that of the *Beg JE* feet.

Hypothesis III: Greater compensatory shortening is observed for the syllables of the *NE* feet than those of the *JE* feet.

Hypothesis IV: Greater compensatory shortening is observed for the syllables of the *Adv JE* feet than those of the *Beg JE* feet.

In addition to the four hypotheses, one research question was formulated. *JE* foot rhythm is assumed to be influenced by lexical stress contrasts with the vowel duration, i.e. phonetic realisations of the rhythmic beats (Chapter 5), and frequency of vowel epenthesis, which would hinder nativelike rhythm in *JE* (Chapter 6). Therefore, the relative contributions of these two factors were investigated (Research Question I).

Research Question I: What are the relative influences of lexical stress contrasts and vowel epenthesis on the *JE* foot rhythm?

The hypotheses will be tested in relevant sections. Section 7.2 will detail the data coding and normalisation. To test Hypotheses I and II, Section 7.3.1 will investigate the variance of the foot duration. To test Hypotheses III and IV, Section 7.3.3 will examine whether compensatory shortening is observed on foot-internal stressed syllables in relation to the number of syllables in the foot. Section 7.3.4 will investigate the foot-internal shortening of unstressed syllables. Finally, Section 7.3.5 will investigate relative contributions of lexical stress and epenthetic vowels on the *JE* foot rhythm.

7.2. Data coding and normalisations

Foot was defined according to the canonical stress placement of the word in the script of *the North Wind and the Sun*. All feet were assumed to be trochaic, consisting of a stressed syllable and following unstressed syllables (See Section 5.1). Therefore, a stressed syllable marks the beginning of a new foot while the end of the foot is marked by a) another stressed syllable, b) a pause and/or c) a prosodic boundary[17]. In the following script of *the North Wind and the Sun*, "//" denotes a foot boundary and "-" a syllable boundary. All stressed syllables are underlined. Shading denotes all trochaic feet which were analysed. Feet consisting only of unstressed syllables were regarded as degenerate feet (Féry 2018; See Section 2.2.1) and excluded from the data for the analyses. Feet consisting of more than three syllables (n = 551) were excluded so that the feet *gave up the at-(tempt)* and *(im)-me-di-ate-ly* were excluded from the data. When the first schwa of the word *travleler (/*ˈtrævələ/) was not realised, the word was treated as disyllabic *(*i.e. /ˈtrævlə/). See Section 3.5 for deletions of segments. Epenthetic vowels were not counted as one vowel but their intervals were included in the foot. Altogether, 1,407 feet in the *NE*, 5,156 feet in the *Adv JE* and 5,182 feet in the *Beg JE* was obtained. Table 29 is the list of all feet for the current analysis. Altogether, 16 types of one-syllable feet, 18 types of two-syllable feet and 12 types of three-syllable feet were analysed.

[17] Prosodic boundaries were defined 1) before complementisers (e.g. They agreed // that; stronger // than), 2) before relative clauses (one // who first), 3) between subjects and preceding adverbial phrases (e.g. at last // the), 4) after a complex subject with more than one phrase (cloak off // should) and 5) before a subject-verb inversion (closely // did the).

*The // **North** // **Wind** and the // **Sun** were dis- // **put**-ing // **which** was the // **strong**-er // **when** a // **trav**-e-ler // **came** a- // **long** // **wrapped** in a // **warm** // **cloak** // They a- // **greed** // that the // **one** // who // **first** suc- // **ceed**-ed in // **mak**-ing the // **trav**-e-ler // **take** his // **cloak** off // should be con- // **sid**-ered // **strong**-er // than the // **oth**-er // **Then** the // **North** // **Wind** // **blew** as // **hard** as he // **could** // But the // **more** he // **blew** // the // **more** // **close**-ly //did the // **trav**-e-ler // **fold** his // **cloak** a- // **round** him // And at // **last** // the // **North** // **Wind** // **gave** up the at- // **tempt** // **Then** the // **Sun** // **shone** out // **warm**-ly // And im- // **me**-di-ate-ly // the // **trav**-e-ler // **took** off his // **cloak** // And // **so** // the // **North** // **Wind** was o- // **bliged** to con- // **fess** // that the // **Sun** was the // **strong**-er // of the // **two***

Table 29 List of the feet for the analysis

Portions of words in parentheses are not included in the respective foot.

A “-” denotes a syllable boundary.

One-syllable	blew cloak could (a)-greed (a)-long last more north one so sun (a)-tempt took two warm wind	Two-syllable	blew as came a-(long) cloak a-(round) cloak off close-ly first suc-(ceed-ed) fold his more he oth-er (dis)-put-ing (a)-round him (con)-sid-ered shone out strong-er take his then the warm-ly when a
Three-syllable	(o)-bliged to con-(fess) (suc)-ceed-ed in hard as he mak-ing the sun was the sun were dis-(put-ing) took off his		

	trav-e-ler which was the wind and the wind was o-(bliged) wrapped in a

Equation 5 shows the formula to calculate the foot duration. The foot duration (*FD*) was calculated by adding the normalised duration of all syllables (*nσ*) in the foot. The syllable duration (*σD*) was calculated by subtracting the left boundary (i.e. the starting point of the first segment of the syllable) from the right boundary (i.e. the end point or the final segment of the syllable). To alleviate the effect of fluctuations of speech rate, the extracted duration was normalised with the duration of the five words (i.e. the word containing the syllable and two words before and after; cf. Section 5.2). The numerator was the sum of the duration of the five words (*WD*) and the denominator is the number of syllables within the five words (*nσW*). If there was a pause, words within the pause and the other five-word boundary (i.e. the target word plus two words) were used for the normalisation. Out of the obtained 11,745 feet (1,407 feet in the *NE*, 5,156 feet in the *Adv JE* and 5,182), outliers were excluded with the mean ± 3 standard deviation (SD) thresholds of the duration of the entire feet, foot-internal stressed syllables and foot-internal unstressed syllables. This resulted in 438 feet being excluded from the data for the analyses. Hence, 1,396 *NE* feet, 5,099 *Adv JE* feet and 5,132 *JP Beg* feet were analysed.

Equation 5 The formula to calculate the foot duration

$$FD = \sum_{i=1}^{n\sigma} \frac{\sigma Di}{WDi/n\sigma Wi}$$

7.3. Analyses

7.3.1. Variances of the foot duration

Foot duration was investigated not considering the number of syllables since at least the mental isochrony of the *NE* foot rhythm is believed to be independent of the number of syllables within the feet. In order to investigate manifestations of isochronous foot rhythm, the variance rather than the mean of the foot duration was compared between the speaker groups. The current analysis was going to investigate standard deviation (SD), the square root value of variance. However, since SD has a positive correlation with the mean, the mean foot duration was first compared.

To compare the mean foot duration, a one-way analysis of variance was conducted with *anova_test()* on *R* (Ver 4.1.1.), setting the speaker group (*EN, JP Adv* or *JP Beg*) as a between-subject factor and the mean foot duration of each speaker as the dependent variable. The result showed a significant main effect of the speaker group (*F* (2, 205) = 13.88, *p* < .001). Figure 14 shows the distributions of the mean foot duration in relation to the speaker group. The red, green and blue boxes respectively show the distributions of the mean foot duration of the *EN* group (n = 25), the *JP Adv* group (n = 90) and the *JP Beg* group (n = 90).

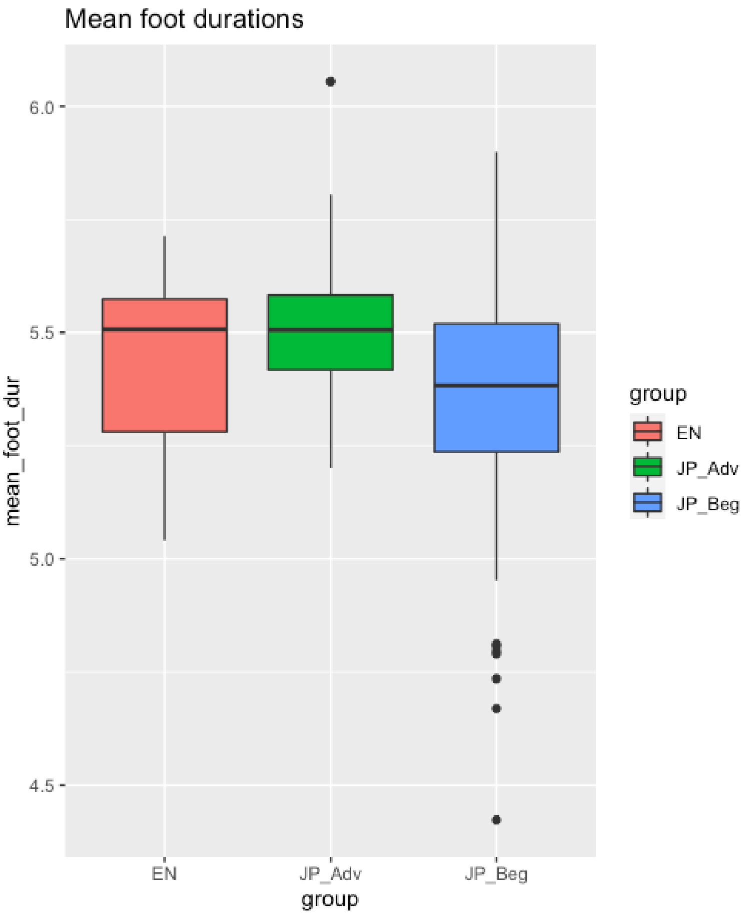


Figure 14 Distributions of the mean foot duration

Thus, instead of the standard deviation, the coefficient of variation (standard deviation divided by mean; hereafter CV) of the foot duration was calculated for each speaker, which was compared between the speaker groups. The same normalisation is adopted for the speech measurements *Variability of coefficient* (*Varco*) proposed by Dellwo and Wagner (2003). Its unnormalised versions (*ΔC, ΔV* and *%V*) proposed by Ramus et al. (1999) were criticised especially in L2 studies since speakers of different proficiency levels often have different speech rates, which affect the values of the SDs.

A one-way analysis of variance was conducted setting the speaker group as a between-subject factor and CV of foot duration as the dependent variable. The result showed a significant main effect of the speaker group (*F* (2, 205) = 57.97, *p* < .001). Post-hoc pairwise comparisons showed the CV of the *NE* foot duration was significantly smaller than that of the *Adv JE* foot duration (p < .01), which was also significantly smaller than that of the *Beg JE* foot duration (p < .001). Figure 15) shows the distributions of the CVs of the foot duration in relation to the speaker group. The red, green and blue boxes respectively show the distributions of the CV values of the *EN* group (n = 25), the *JP Adv* group (n = 90) and the *JP Beg* group (n = 90). The result suggests a smaller variance of the foot duration in speeches of more proficient speakers.

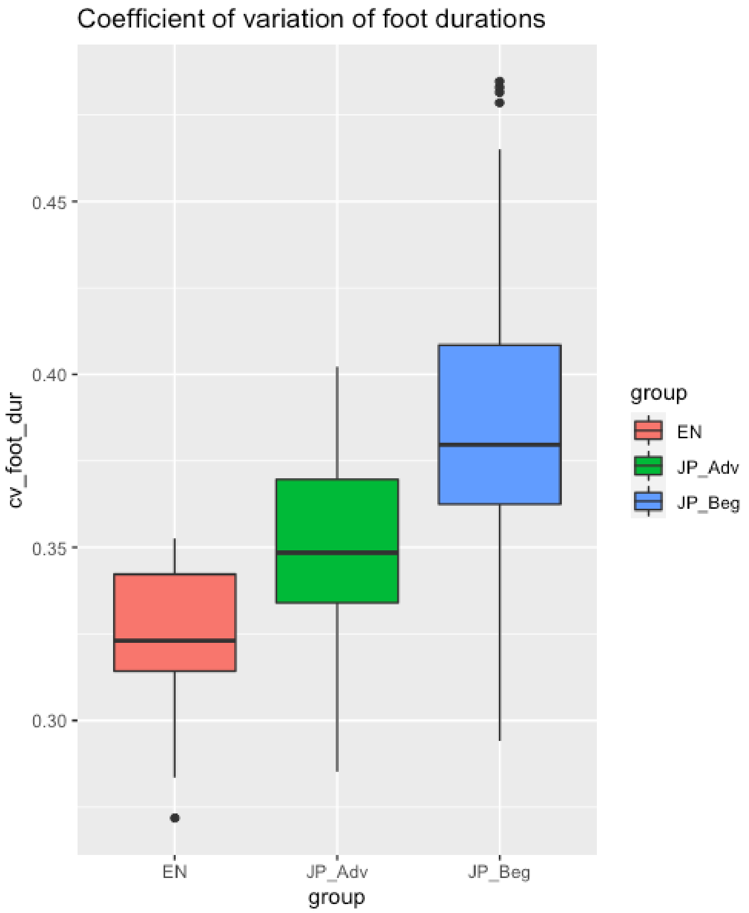


Figure 15 Distributions of standard deviations of the variance of the foot duration

7.3.2. Foot duration and the number of syllables

Foot duration was further investigated in relation to the number of syllables assuming their positive correlation with the number of syllables at least in *JE*. Since Japanese has the mora rhythm, the duration of its speech is proportional to the number of vowels except for moraic nasals and geminates (Bradlow et al. 1995, Port et al. 1987). Therefore, some first language transfer was expected in the speech rhythm of *JE*. In fact, Spearman's rank correlation rho showed even the *NE* foot duration was positively correlated with the number of syllables (rho = 0.704, *p* < .001). That of *Adv JE* and *Beg JE* was also positively correlated with the number of syllables (rho = 0.767 and rho = 0.757, *p*s < .001).

A mixed ANOVA was conducted with *anova_test()* on *R* (Ver 4.1.1.). The predictor variables were the fixed effects and interaction of the number of syllables within one foot (one, two or three; a between-subject factor) and the speaker's language group (*EN*, *JP Adv* and *JP Beg*; a within-subject factor). The dependent variable was foot duration. The result showed a significant main effects of the number of syllables (*F*(2, 410) = 5647.39, *p* < .001), that of the speaker group (*F*(2, 205) = 2.49, *p* < .001) and the interaction (*F*(4, 410) = 24.97, *p* < .001).

Figure 16 shows foot duration in relation to the number of foot-internal syllables and the speaker groups. The left pane shows the distributions of the *EN* group (n = 1,360); the centre pane shows those of the *JP Adv* group (n = 4,978); and the right pane shows those of the *JP Beg* group (n = 4,969). In each pane, the red box on the left shows the distributions of the one-syllable feet; the green box at the centre shows those of the two-syllable feet, and the blue box on the right shows those of the three-syllable feet. Based on the results of the post-hoc pairwise comparisons, in each group, the duration of the one-syllable feet were significantly smaller than that of the two-syllable feet, which was also significantly smaller than that of the three-syllable feet (ps < .001).

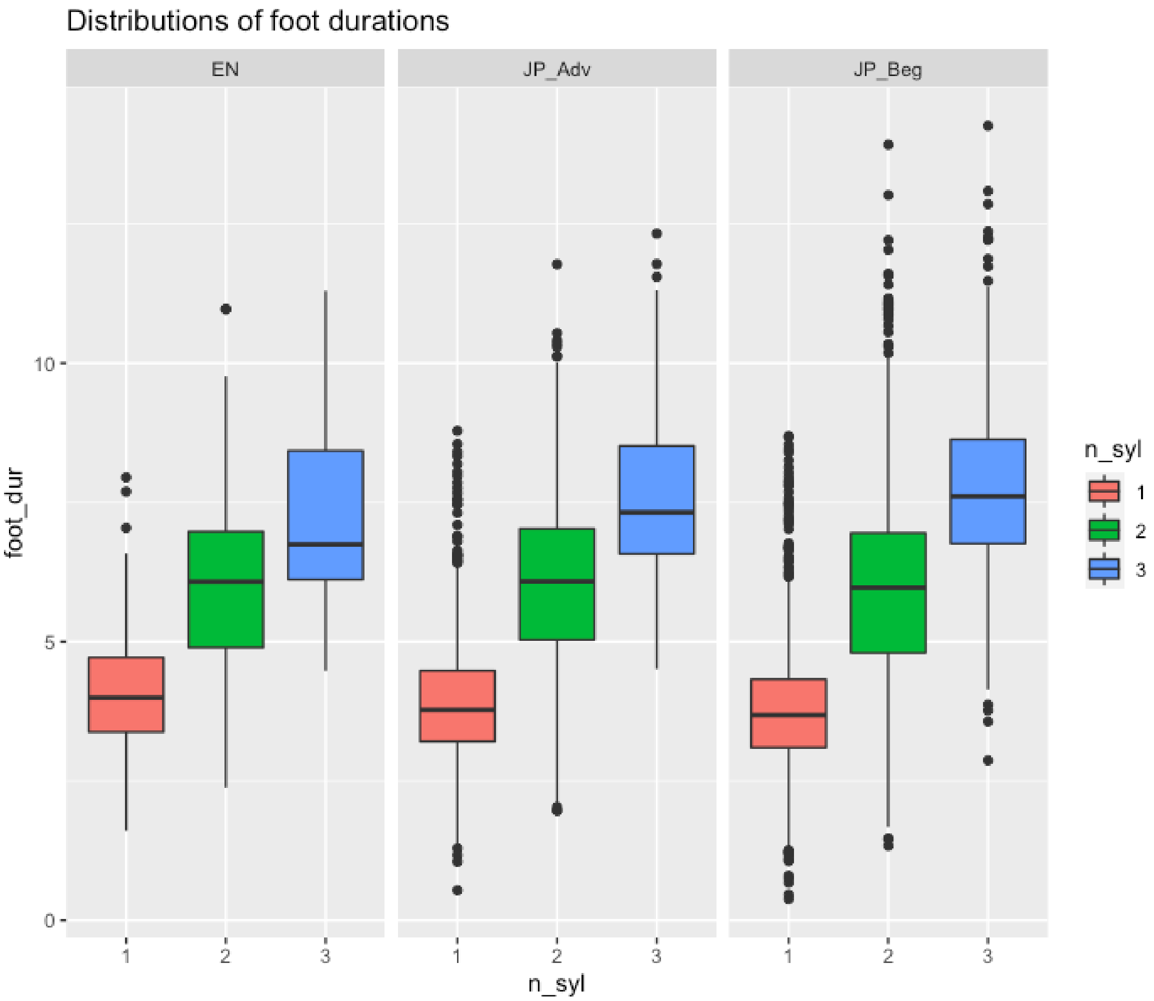


Figure 16 Foot duration in relation to the number of syllables in the foot

In Figure 17, foot duration is compared between the groups and in relation to the number of syllables. The left pane shows the distributions of the one-syllable feet (n = 4,708); the centre pane shows those of the two-syllable feet (n = 4,249); and the right pane shows those of the three-syllable feet (n = 2,350). In each pane, the red box on the left shows the distributions of the *EN* group; the green box at the centre shows those of the *JP Adv* group, and the blue box on the right shows those of the *JP Beg* group. The post-hoc comparisons showed the duration of one-syllable feet of the *EN* group and *JP Adv* group was significantly larger than that of the *JP Beg* group (p < .001 and p < .05 respectively). No significant difference was observed between the former two groups (p > .1). On the other hand, the duration of the three-syllable feet was significantly smaller in the *NE* than the *Adv JE* and the *Beg JE* (*p* < .01 and *p* < .001 respectively). No significant difference was observed between the two *JP* groups (*p* > .1). These results indicate that the variance of the foot duration is smaller in *NE* than in *JE* and that the variance is also smaller in *Adv JE* than in *Beg JE*.

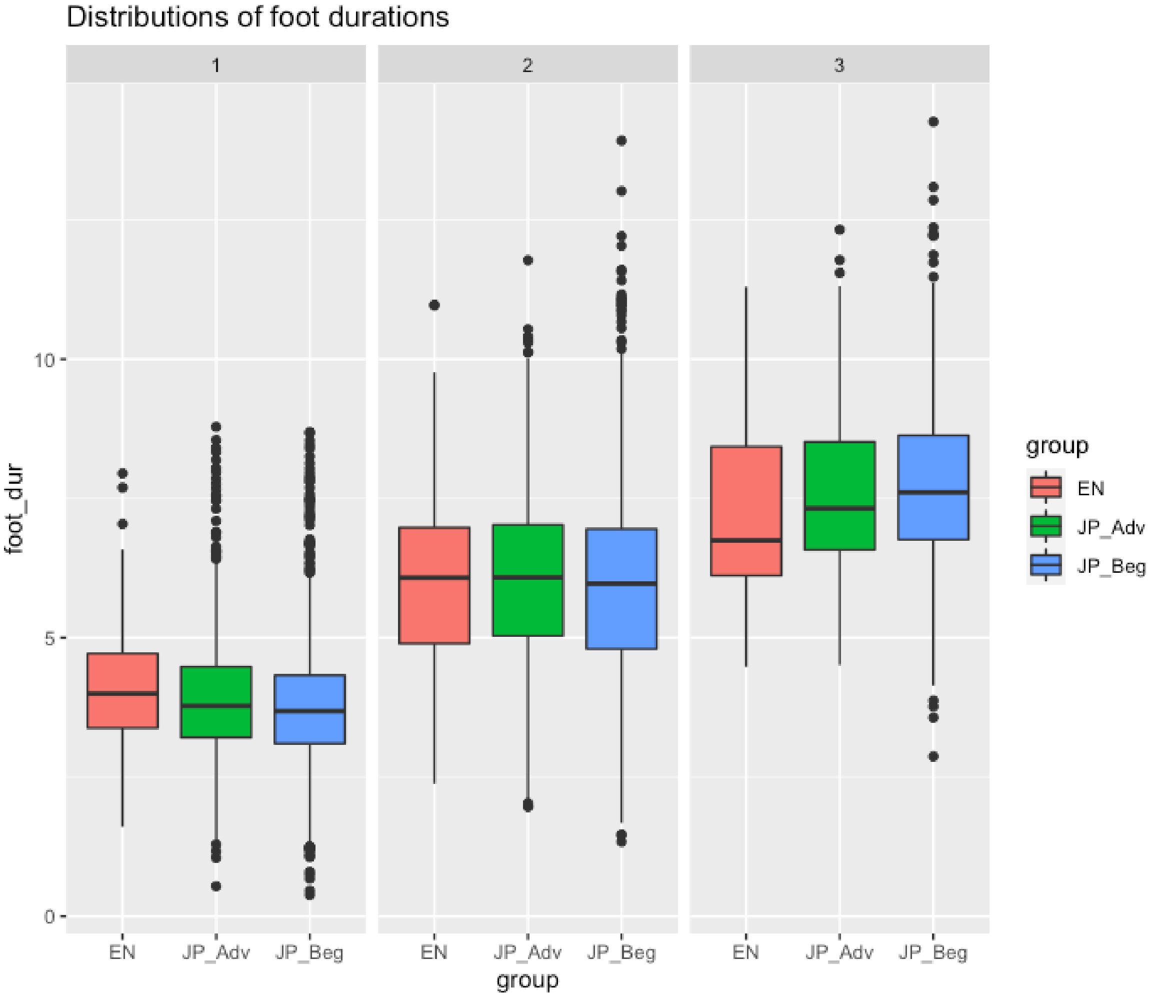


Figure 17 Foot duration in relation to the number of syllables in the foot

Table 30 lists the mean foot duration of the one-syllable to three-syllable feet in each group. The leftmost column shows the number of syllables in the foot and the top row shows the speaker group. For the sake of comparison, all mean values were divided by the mean value of one-syllable feet. Again, they imply the smallest variance of the duration of the *NE* feet (n = 1,396). The mean duration of the three-syllable feet was 1.775 times as large as that of the one-syllable feet. The *JE Adv* feet (n = 5,099) showed the second smallest increase and the duration of its three-syllable feet was 1.949 times as large duration as its one-syllable feet. On the other hand, even the duration of the *JE Beg* feet (n = 5,132) were not completely proportional to the number of syllables, despite the expected transfer of the L1 mora rhythm. The duration of the three-syllable feet was 2.056 times as large as that of the one-syllable feet.

Table 30 Mean normalised duration of foot in relation to the number of syllables

	EN	**JP Adv**	**JP Beg**
1	1.000	1.000	1.000
2	1.483	1.558	1.579
3	1.775	1.949	2.056

7.3.3. Compensatory shortening in stressed syllables

The result of the analysis in the previous section suggests that *NE* feet have a smaller variance in their duration and thus are more isochronous than *JE* feet. As a potential factor of the greater isochrony, this section will investigate compensatory shortening in stressed syllables in the foot. (Compensatory shortening in unstressed syllables will be investigated in Section 7.3.4.) In the previous studies, compensatory shortening was observed on stressed syllables rather than unstressed ones (Fowler 1977, 1981, Huggins 1973, Lehiste 1972, Rakerd et al. 1987). Thus, stressed syllables are expected to adjust their duration based on the number of syllables in the foot.

To investigate compensatory shortening on stressed syllables, a mixed ANOVA was conducted. The predictor variables were the fixed effects and interaction of the number of syllables within one foot (one, two or three; a within-subject factor) and the speaker's language group (*EN*, *JP Adv* and *JP Beg*; a between-subject factor). The dependent variable was the foot duration The showed a significant main effect of group ($F(2, 205) = 72.18$, $p < .001$) and its interaction with the number of syllables ($F(4, 410) = 3.51$, $p < .01$). However, the main effect of the number of syllables was not significant ($F(2, 410) = 1.35$, $p > .2$).

Figure 18 shows the distributions of the duration of the foot-internal stressed syllables in relation to the number of syllables. The left pane shows the distributions of the *EN* group (n = 1,360). The centre pane shows those of the *JP Adv* group (n = 4,978) and the right pane shows those of the *JP Beg* group (n = 4,969). In each pane, the red, green and blue boxes respectively show the distributions of the stressed syllables in the one-syllable, two-syllable and three-syllable feet. Unlike the result of previous studies (Fowler 1977, 1981, Huggins 1973, Lehiste 1972, Rakerd et al. 1987), in which a negative correlation was observed between the duration of the stressed syllables and the number of foot-internal syllables, the post-hoc pairwise comparisons showed no significant difference in relation to the number of syllables for either group. The duration of the stressed syllables in the one-syllable feet of the *EN* group had no significant difference from that of the two-syllable and three-syllable feet (*p*s = 1). The stressed syllables of the *Adv JE* and *Beg JE* feet showed the same tendency (*p*s = 1). Hence it was indicated that the duration of foot-internal stressed syllables is not influenced by the number of syllables in the foot.

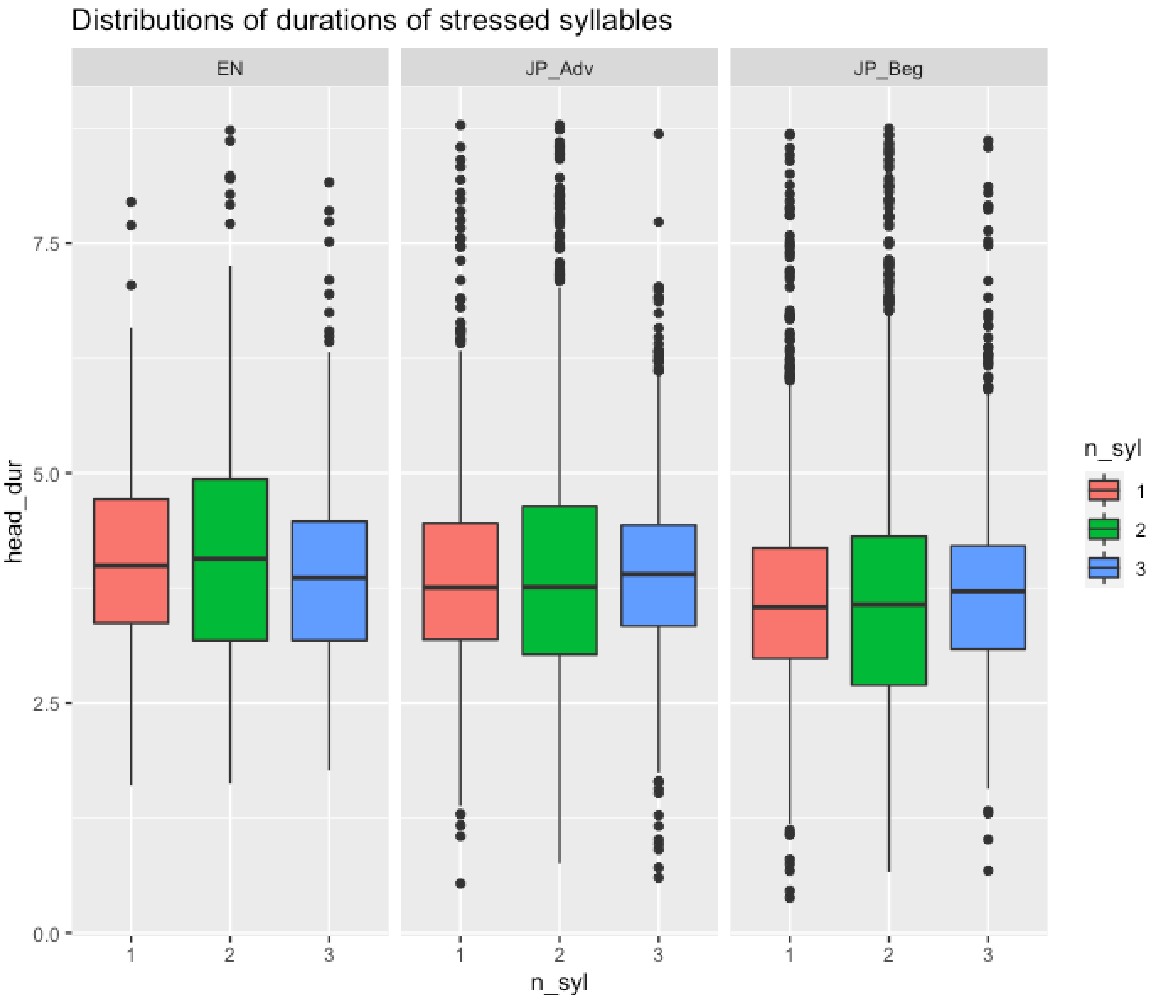


Figure 18 Duration of the foot-internal stressed syllables in relation to the number of syllables in the foot

On the other hand, the post-hoc pairwise comparison showed inter-group differences. Figure 19 shows the distributions of the duration of foot-internal stressed syllables in relation to the speaker groups. The left, centre and right pane respectively show the distributions of the duration of the stressed syllables in the one-syllable feet (n = 4,708), the two-syllable feet (n = 4,249) and the three-syllable feet (n = 2,350). In each pane, the red, green and blue boxes respectively show the distributions of the *NE, Adv JE* and *Beg JE* feet. The duration of the stressed syllables in the *NE* one-syllable feet was significantly greater than that of the *Adv JE* feet (*p* < .05) and the *Beg JE* feet (*p* < .001). The same relations were observed for the two-syllable feet (*p*s < .001). In addition, the duration of the stressed syllables in the three-syllable feet of the *NE* and *Adv JE* was significantly greater than that of the *Beg JE* (*p* < .01). This seems to reflect different extents of durational contrasts manifested by each group; i.e. the greatest contrast is made between stressed and unstressed vowels in *NE,* followed by *Adv JE* (See Section 5.3.1).

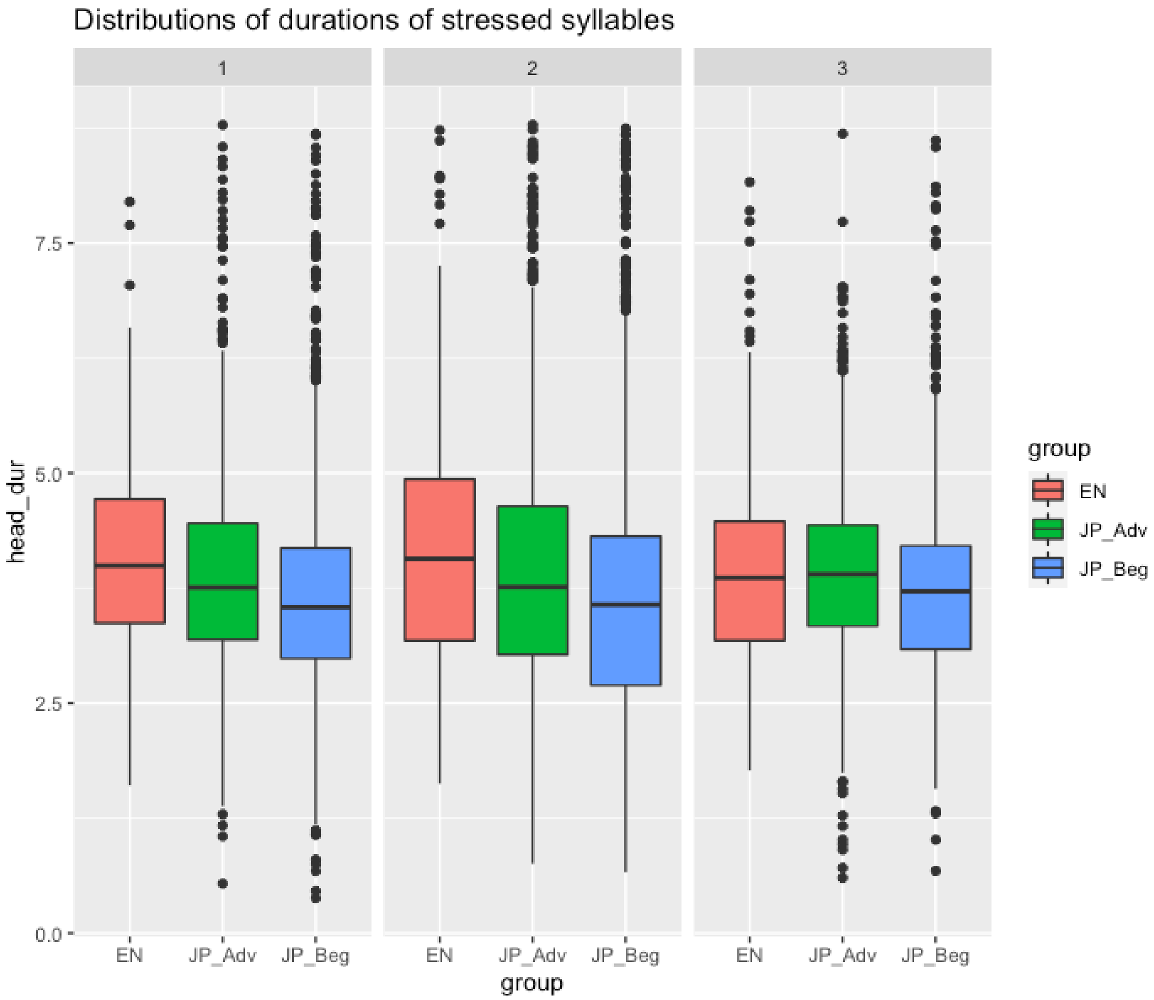


Figure 19 Duration of the foot-internal stressed syllables in relation to the speaker group

7.3.4. Compensatory shortening in unstressed syllables

Compensatory shortening in unstressed syllables was also compared between two-syllable feet and three-syllable feet. Containing no unstressed syllables, one-syllable feet (n = 4,708) were excluded from the data. Hence, 6,599 feet were analysed.

A mixed ANOVA was conducted, setting the number of syllables (two or three; a within-subject factor) and the speaker group (*EN, JP Adv* or *JP Beg*; a between-subject factor) as the predictor variables and assuming their interaction. The dependent variable was the total duration of all unstressed syllables within each foot. As a result, the main effect of the number of syllables ($F(2, 410) = 8536.49$, $p < .001$), that of the group ($F(2, 205) = 77.89$, $p < .001$) and the interaction ($F(4, 410) = 25.37$, $p < .001$) were all significant.

The post-hoc pairwise comparisons showed the duration of the unstressed syllables were greater in three-syllable feet than two syllable-feet for all groups (*p*s < .001). Figure 20 shows the distributions of the duration of the foot-internal unstressed syllables in relation to the number of syllables. The left pane shows the distributions of the *NE* feet (n = 816), the centre pane shows those of the *Adv JE* (n = 2,967) and the right pane those of the *Beg JE* (n = 2,816). In each pane, the red box shows the distribution of the two-syllable feet and the blue box that of the three-syllable feet.

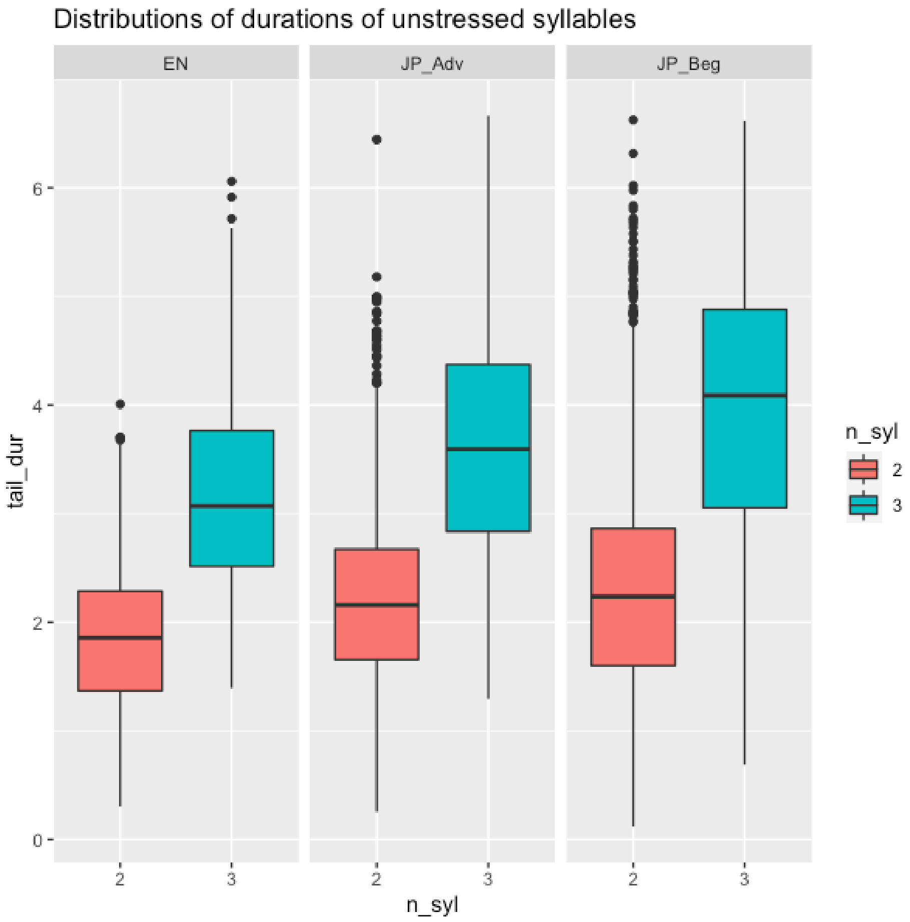


Figure 20 Duration of the foot-internal unstressed syllables in relation to the speaker group

In addition, the duration of the two-syllable feet of the *EN* group was significantly smaller than that of the *JP Adv* group (*p* < .001), which in turn was significantly smaller than that of the *JP Beg* group (*p* < .01). The three-syllable feet showed the same tendency and the duration were smallest in the *NE*, followed by the *Adv JE* (*p*s < .001) Figure 21 shows the distributions of the duration of the foot-internal unstressed syllables in relation to the speaker groups. The left pane shows the distributions of the two-syllable feet (n = 4,249) and the right pane shows those of the three-syllable feet (n = 2,350). In each pane, the red, green and blue boxes respectively show the distributions of the *NE, Adv JE* and *Beg JE* feet.

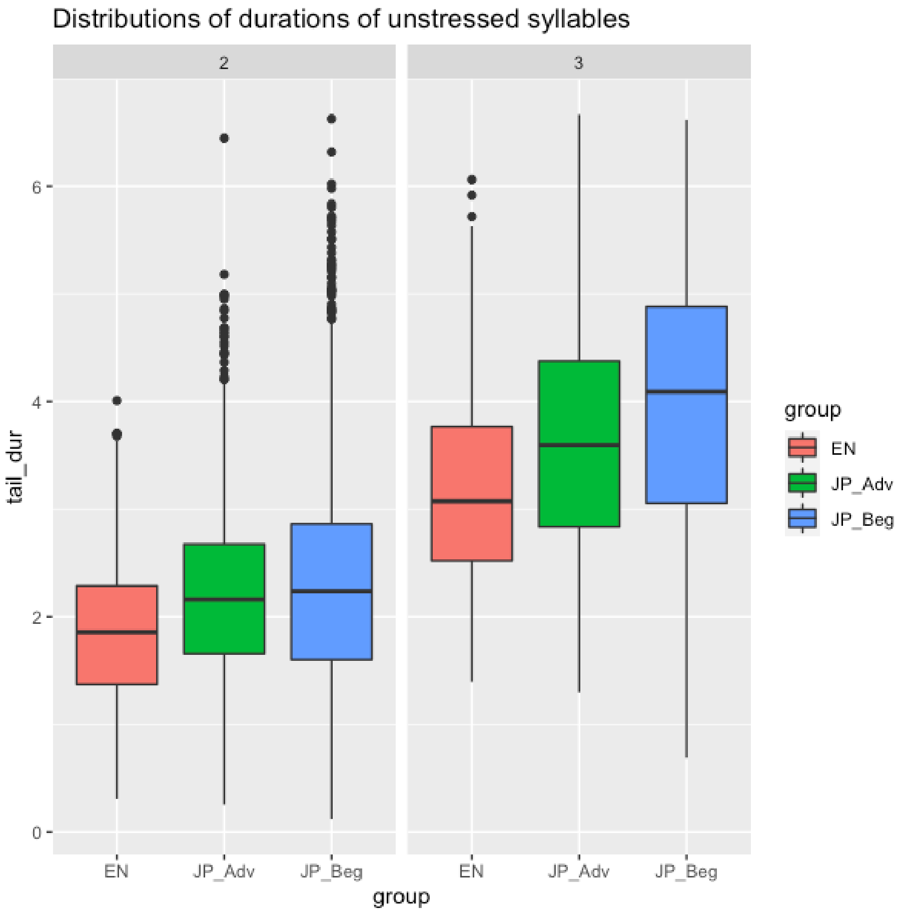


Figure 21 Duration of the foot-internal unstressed syllables in relation to the number of syllables in the foot

Table 31 lists the mean total duration of foot-internal unstressed syllables in each group. The leftmost column shows the number of syllables in the foot and the top row shows the speaker group. For the sake of comparison, all mean values were divided by the mean value of the total duration of unstressed syllables in the two-syllable feet. Despite the smaller duration of the unstressed syllables of *NE* feet compared to that of *JE* feet, the proportional increase of the duration from two-syllable feet to three-syllable feet were similar for the three speaker groups. In addition, compensatory shortening was observed even in the *Beg JE* feet so that the total duration of the unstressed syllables in the three-syllable feet (i.e. the sum of the duration of the two unstressed syllables) was not twice as long as that of the two-syllable feet (i.e. the duration of the one unstressed syllable).

Table 31 Mean normalised duration of unstressed syllables in relation to the number of syllables in the foot

	EN	**JP Adv**	**JP Beg**
2	1.000	1.000	1.000
3	1.714	1.691	1.712

7.3.5. Phonetic correlates of foot rhythm in Japanese English

Lastly, phonetic correlates which are assumed to influence the foot duration of *JE* were investigated. The expected correlates are the degree of durational contrast of lexical stress, which is assumed to positively contribute to the manifestations of the isochronous rhythm by reducing the duration of unstressed vowels, and vowel epenthesis, which is assumed to negatively contribute to the manifestations of the isochronous rhythm by adding extra duration to syllables with epenthetic vowels. They are both important factors of the *JE* rhythm as well as the foot rhythm (See Section 2.4). The effects of the two factors were first discussed below.

The results of the analyses in this Chapter showed the effect of lexical stress contrast on the manifestations of isochronous rhythm in *JE*. In Section 7.3.1 and Section 7.3.2, *NE* had more isochronous feet than *JE* in that the former had a smaller variance of the foot duration. In addition, the feet of *Adv JE* had smaller variance of their duration than those of the *Beg JE*. Unlike the results of previous studies (Fowler 1977, 1981, Huggins 1973, Lehiste 1972, Rakerd et al. 1987), compensatory shortening was observed for foot-internal unstressed syllables (Section 7.3.4) rather than stressed syllables (Section 7.3.3). Although there was no group difference in the extent of compensatory shortening on unstressed syllables (i.e. the proportional increase of the total duration of foot-internal unstressed syllables from two-syllable to three-syllable feet was not different between the three proficiency groups), the actual increases of the foot duration with additional unstressed syllables were smallest in the *EN* group, followed by the *JP Adv* group. This seems to be reflecting greater durational reductions of unstressed vowels by the *EN* group than the *JP Adv* group, and the *JP Adv* group than the *JP Beg* group (Section 5.3.1).

In addition, vowel epenthesis is assumed to negatively contribute to isochronous foot rhythm by increasing the duration of the feet with epenthetic vowels; in the analyses of Chapter 6, epenthetic vowels were not counted as vowels but their duration was added to the foot duration (See Section 7.2). To test this, a mixed-effects multiple regression analysis was conducted on the data of the *JP* group. The predictor variables were the fixed effect of the number of syllables in the foot (one, two or three), that of vowel epenthesis (i.e. whether the foot contains epenthetic vowels), and the interaction between the two. A random intercept of the combinations of syllables was assumed as well. The dependent variable was foot duration. P-values were calculated with backward stepwise regression.

The result yielded a significant interaction between the number of syllables and the speaker group (p < .001). Figure 22 shows the foot duration in relation to the number of syllables and whether the foot contains epenthetic vowels. The left pane shows the distributions of the duration of the one-syllable feet (n = 4,164). The centre pane shows the distributions of the two-syllable feet (n = 3,682). The right pane shows the distributions of the three-syllable feet (n = 2,101). In each pane, the red box shows the distribution of feet without epenthetic vowels while the blue box shows that of feet with epenthetic vowels. The post-hoc pairwise comparisons showed, regardless of the number of syllables in the foot, those containing epenthetic vowels had greater duration (*p*s < .001).

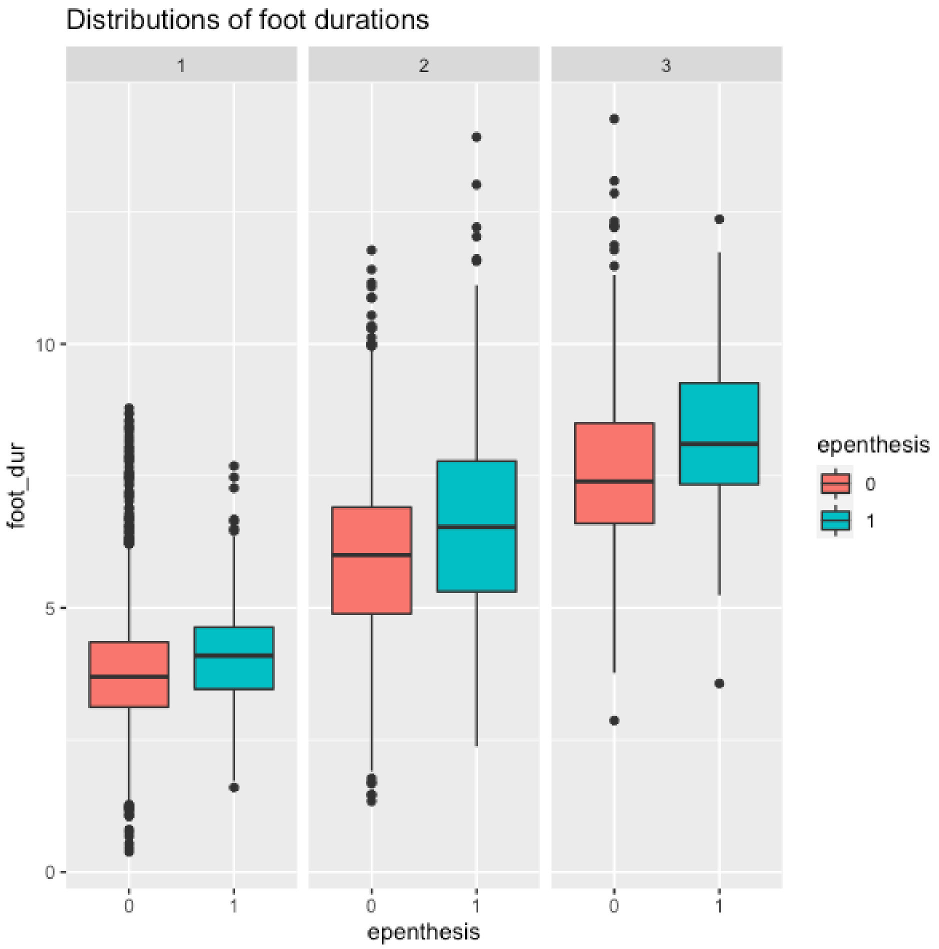


Figure 22 Distribution of the foot duration in relation to the number of syllables and inclusion of epenthetic vowels

Hence both durational contrast of stress and vowel epenthesis were integrated into the model. Since the *EN* group had few epenthetic vowels (See 6.3.1), the group was excluded from the data for the analysis. The frequency of vowel epenthesis was measured by counting the epenthetic vowels in each subject's utterance. Out of the 90 subjects in the *JP Adv* group, 7 had no epenthesis. In contrast, all the 93 subjects in the *JP Beg* group had epenthesis at least once.

Although the number of syllables in the foot positively correlated with the foot duration (Section 7.3.2), the factor was not integrated into the model since the data was extracted from a read speech. In fact, there was a certain influence of proficiency on the mean number of foot-internal syllables perhaps because of frequent pauses and dysfluencies in utterances of lower-proficient speakers. Figure 23 shows the distributions of the mean number of foot-internal syllables (the y-axis) in relation to the proficiency score (the x-axis). Pearson's product-moment correlation showed a significant positive correlation between the proficiency score and the number of foot-internal syllables (r = 0.260, t(181) = 3.624, p < .001). However, the correlation between the coefficient of variation (CV) of the foot duration and the mean number of syllables in the feet was negative rather than positive. Figure 24 shows the distributions of the CV of the foot duration for each speaker (the y-axis) in relation to the mean number of foot internal syllables (the x-axis). Pearson's product-moment correlation showed a significant negative correlation (r = -0.405, t(181) = -5.96, p < .001), which implies smaller foot variance for a greater mean number of syllables in a speaker's feet. The negative correlation was assumed to be due to the confounding effect of proficiency (i.e. more proficient speakers should have greater mean numbers of foot-internal syllables but the smaller variance in the foot duration). Therefore, the factor was not integrated into the model.

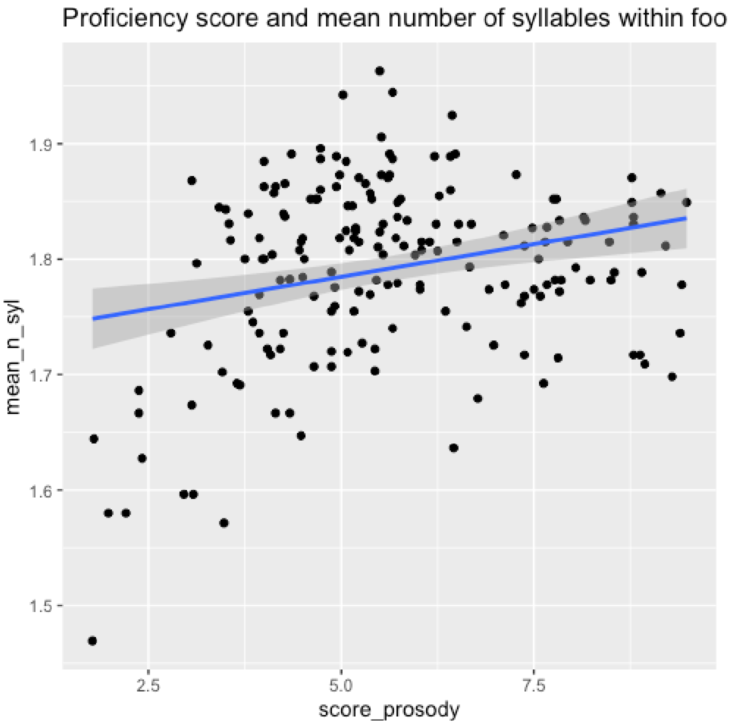


Figure 23 The mean number of foot-internal syllables and the proficiency score

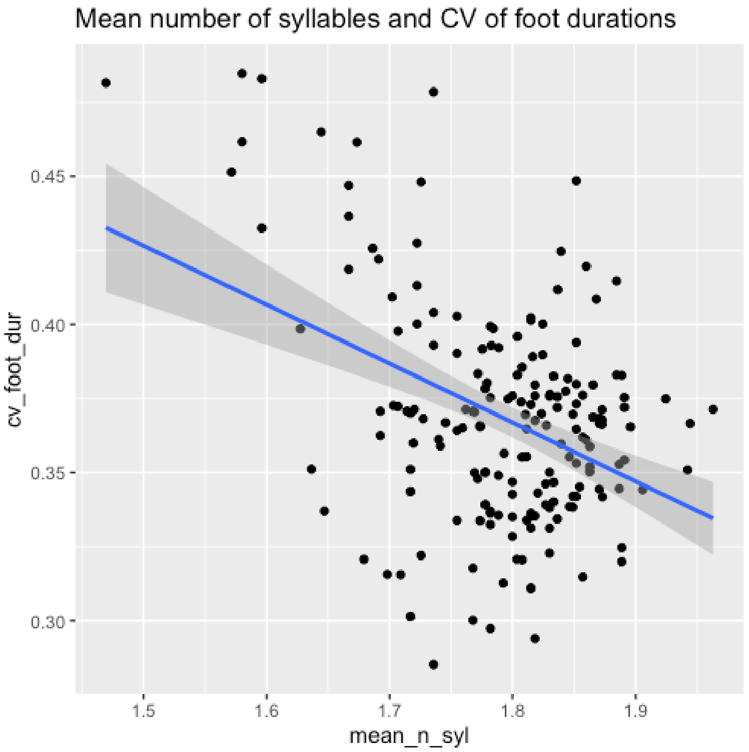


Figure 24 The variance of the foot duration and the mean number of syllables

Data extractions and normalisations of the vowel duration mostly followed the same methods adopted in the previous sections (Section 5.2 and Section 6.2). The duration of stressed and unstressed vowels was measured with annotated boundaries of vowel segments in the *J-AESOP* corpus. Whether a given vowel was stressed or not was determined by its canonical, rather than actual, stress placement as defined by the Longman Pronunciation Dictionary (3rd ed). Out of the 27,812 vowels in the *JE,* epenthetic vowels (n = 2,120) were excluded from the data. To address fluctuating speech rates, the measured duration was normalised with the duration of the five words (i.e. the word containing the vowel and two words before and after). Equation 1 shows the formula for the normalisation where *ND* denotes the normalised duration and *VD* the raw vowel duration. The denominator is the word duration (*WD*), defined as the sum of the duration of the five words, divided by the number of vowels in the words (*nV*). The word duration was defined to be the sum of the duration of all vowels within the word, including the epenthetic vowels. Out of the 25,692 data of the normalised vowel duration, outliers (n = 280) were excluded with mean ± 3 standard deviation thresholds. See Section 5.2.1 for more details. Finally, the data were summarised for each subject by calculating the duration ratio of stressed and unstressed vowels (*DR*) with

Equation 6, where *µVDst* denotes the mean duration of all stressed vowels and *µVDus* the mean duration of all unstressed vowels.

Equation 1 Normalised vowel duration for the analysis

$$ND = \frac{VD}{WD/nV}$$

Equation 6 Duration ratio of stress contrasts calculated for each subject

$$DR = \frac{\mu VDst}{\mu VDus}$$

Relative influences of the acoustic correlates on the variance of the foot duration were investigated with multiple regression analyses using *lm()* on *R* (Ver 4.1.1.). In order to investigate the difference between the proficiency groups, separate models were constructed for the *JP Adv* group and *JP Beg* group. The dependent variable was the CV of the foot duration calculated for each speaker. The predictor variables were the fixed effects of the duration ratio of stressed and unstressed vowels (

Equation 6; hereafter ‘stress contrast’) and that of the number of epenthetic vowels (hereafter ‘vowel epenthesis), as well as their interaction.

The model for the *JP Adv* group yielded a significant main effect of stress contrast (p < .01) but no main effect of vowel epenthesis (p > .5). The interaction was not significant either. (p < .3). To investigate their relative effects, the standardised coefficients of the fixed effect and the interaction were calculated with *lm.beta()* of the *psych* package (Ver 2.1.9.) on *R*. The results are summarised in Table 32. The value was -0.476 for the stress contrast, -0.694 for vowel epenthesis and 0.891 for the interaction between the two. Although the main effect of vowel epenthesis and its interaction with stress contrast both have larger coefficients than that of stress contrast, implying the stronger influences of the former two, neither of their effects is statistically significant.

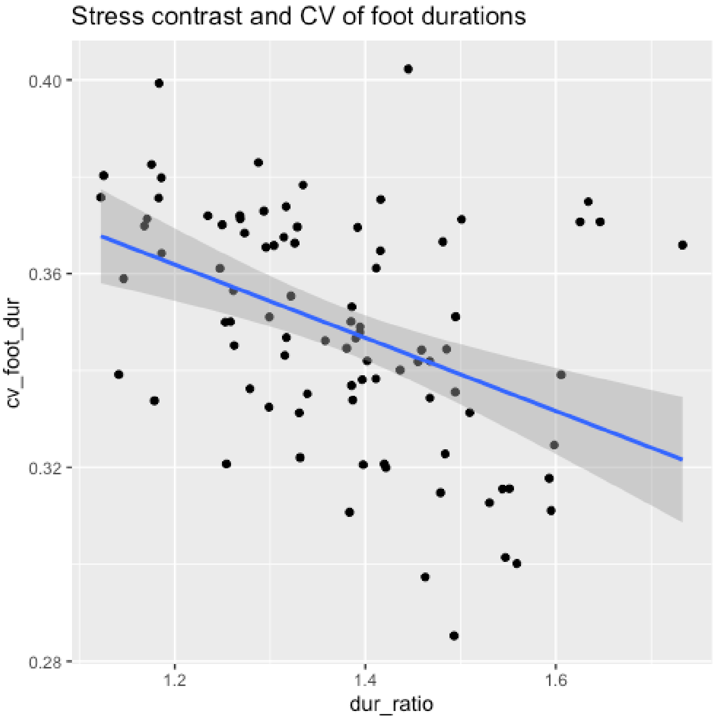


Figure 25 shows the variance of the foot duration (the y-axis) in relation to the degree of stress contrast (the x-axis). It can be seen that the correlation is negative, suggesting smaller variance of the foot duration, i.e. more isochronous rhythm, when greater stress contrasts are made. On the other hand, the negative coefficient of vowel epenthesis is rather difficult to interpret since it implies smaller variance of the foot duration when there is more frequent epenthesis.

Table 32 Standardised coefficients of the predictor variables on the variance of the foot duration (the JP Adv group)

	Stress contrast	*Vowel epenthesis*	*The interaction*
Statistical significance	*p* < .01	*p* > .5 (*n.s.*)	*p* > .3 (*n.s.*)
Standardised coefficient	-0.476	-0.694	0.891

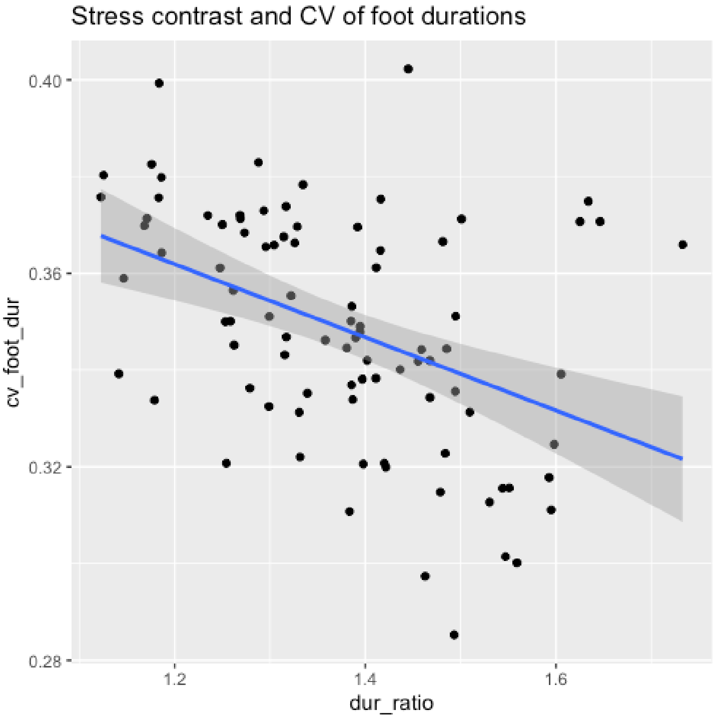


Figure 25 Variance of the foot duration and the degree of the stress contrast (the JP Adv group)

The model for the *JP Beg* group yielded a significant main effect of vowel epenthesis ($p < .01$) and its significant interaction with stress contrast ($p < .05$). The main effect of stress contrast was not significant ($p > .7$). Again, their relative effects were compared by their standardised coefficients. As summarised in Table 33, the value was -0.048 for the stress contrast, whose effect was not statistically significant. The coefficient of vowel epenthesis was 2.644 and that of its interaction with stress contrast was -2.401, implying almost the same degree of influences by the two. The effect of the frequency of vowel epenthesis was positive so that the more frequent epenthesis was, the greater variance (i.e. less isochronous rhythm) the foot duration has. Figure 26 shows the variance of the foot duration (the y-axis) in relation to the number of epenthetic vowels (the x-axis).

Table 33 Standardised coefficients of the predictor variables on the variance of the foot duration (the JP Beg group)

	Stress contrast	*Vowel epenthesis*	*The interaction*
Statistical significance	$p > .7$ (*n.s.*)	$p < .01$	$p < .05$
Standardised coefficient	-0.048	2.644	-2.401

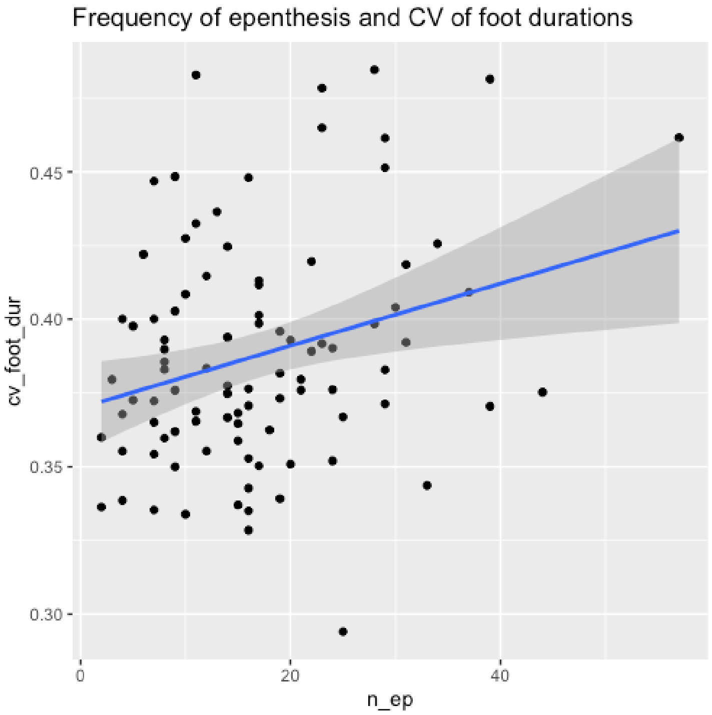


Figure 26 The number of epenthetic vowels and CV of the foot duration (the JP Beg group)

The interaction between the two independent variables is shown in Figure 27. The x-axis shows the degree of stress contrast and the y-axis the variance of the foot duration. The red dots are the data of the speakers who had vowel epenthesis as frequently as or more frequently than the mean. The blue dots are the data of the speakers with less than the mean frequency of vowel epenthesis. The coefficient of the interaction was negative (Table 33) so that, regardless of the frequency of vowel epenthesis, the effect of stress contrast was negative. That is, as observed for the *JP Adv* group, the more isochronous rhythm was manifested when greater stress contrast was made. In addition, it can be interpreted that the effect of stress contrast was greater for speakers with more frequent vowel epenthesis. Since the frequency of vowel epenthesis negatively correlates with the speaker's proficiency (Section 6.3.1), it can be inferred that the effect of stress contrast would have been greater for lower-proficiency speakers among the *JP Beg* group.

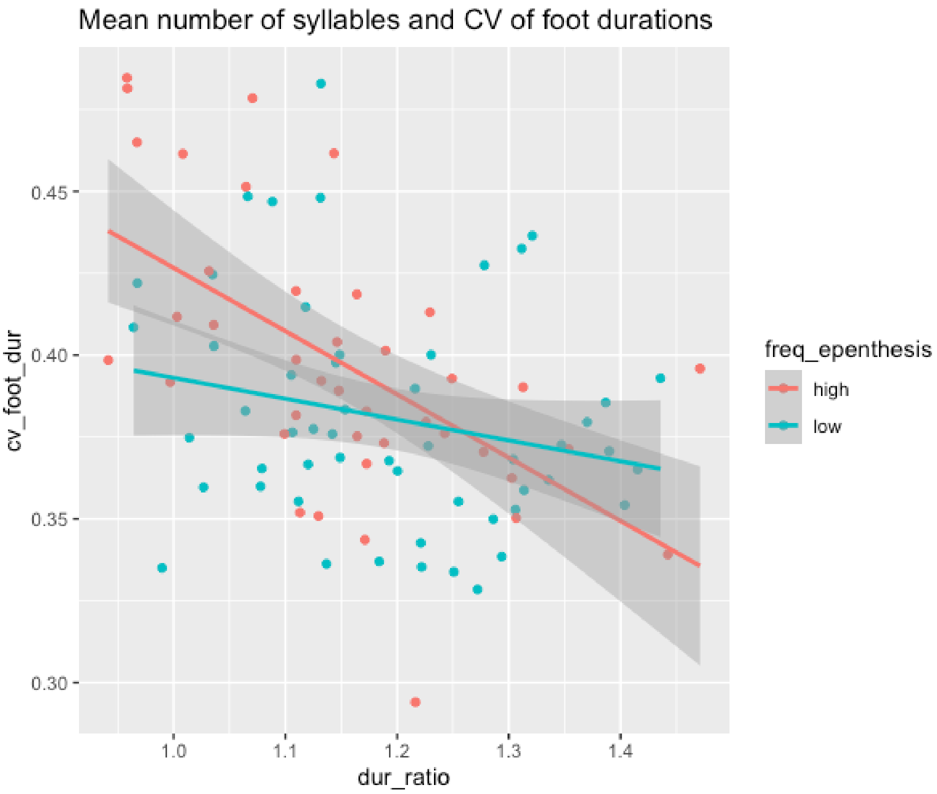


Figure 27 Variance of the foot duration and the degree of the stress contrast (the JP Beg group)

7.4. Discussion

No previous studies have demonstrated the physical isochrony of *NE* feet except in restricted environments (e.g. Tajima 1998). This was perhaps because the *NE* foot duration does correlate with the number of syllables within the feet. Actually, the analysis in the current study yielded a significant strong correlation between the two (*rho*s > 0.7, *p*s < .001 in all groups; Section 7.3.2). However, the results of the analyses also suggest *NE* feet are at least more isochronous than *JE* feet. In fact, the results showed the duration of the *NE* feet had a smaller variance than that of the *JE* feet (Section 7.3.1). The coefficient of variation (CV) of the *NE* foot duration was significantly smaller than that of the *JE* foot duration (*p* < .01), which supports Hypothesis I (Table 34). In addition, the foot duration of the *Adv JE* had a significantly smaller CV than that of the *Beg JE* (*p* < .001), supporting Hypothesis II.

Table 34 The hypotheses and results of the analysis of the foot duration

Hypothesis I	The duration of the *NE* feet has a smaller variation than that of the *JE* feet.	**Supported**
Hypothesis II	The duration of the *Adv JE* feet has a smaller variation than that of the *Beg JE* feet.	**Supported**

Perhaps, the smallest variance of the *NE* foot duration would have been due to the smallest effect of the number of syllables on the foot duration. The one-syllable feet of the *NE* and *Adv JE* had significantly greater duration than those of the *Beg JE* (*p* < .001 and *p* < .05 respectively). On the other hand, the three-syllable feet of the *NE* had significantly smaller duration than those of the *Adv JE* (*p* < .01) and those of the *Beg JE* (*p* < .001). This means the proportional increase in the foot duration in relation to the number of syllables was smallest in the *NE*, followed by the *Adv JE.*

Despite the significant positive correlations, the foot duration was not completely proportional to the number of syllables. The result is consistent with that of Mochizuki-Sudo and Kiritani (1991) which also observed smaller proportional increases in the foot duration with additions of foot-internal unstressed syllables in the speech of more proficient speaker groups[18]. Perhaps reflecting durational contrast of vowels with and without lexical stress (i.e. lexically stressed vowels have significantly greater duration than lexically unstressed vowels; Section 5.3.1), the mean duration of the two-syllable feet, consisting of one stressed and one unstressed syllables, was not twice as long as that of the one-syllable feet, consisting of one stressed syllable (1.483 for the *EN* group, 1.558 for the *JP Adv* group and 1.579 for the *JP Beg* group). Similarly, the mean duration of the three-syllable feet, consisting of one stressed and two unstressed syllables, was not three times as long as that of the one-syllable feet (1.775 for the *EN* group, 1.949 for the *JP Adv* group and 2.056 for the *JP Beg* group).

Contrary to the results of the previous studies (Fowler 1977, 1981, Huggins 1973, Lehiste 1972, Rakerd et al. 1987), compensatory shortening was not observed on the foot-internal stressed syllables except for the unexpectedly greater duration of the stressed syllables of the three-syllable feet in the *Beg JE* (Figure 18; See Section 7.3.3); theoretically, the duration of each syllable should be compressed when there are more syllables within the foot. Hence it seems the duration of foot-internal stressed syllables were independent of the number of syllables at least in the current study. In fact, the target syllables were strictly controlled in the study by Rakerd et al. (1987) and the same syllables were compared with the presence or absence of a following unstressed syllable.

[18] However, statistical significance was not tested in Mochizuki-Sudo and Kiritani (1991).

This might have resulted in the contrasting results between their study and the current study.

It is important to note that previous studies did not use the same method in investigating the foot duration. For example, Lehiste (1972) and Rakerd et al. (1987) measured the duration of the stressed syllables while Mochizuki-Sudo and Kiritani (1991) measured the duration of the vowels. The current study compared the syllable duration and found no significant difference. To further investigate the difference between the results of the previous studies and the current study, the same analysis was conducted on the foot-internal stressed vowels. Figure 28 shows the distributions of the duration of the foot-internal stressed vowels in relation to the number of foot-internal syllables (red: one, green: two, blue: three) and the proficiency groups (left: the *EN* group, centre: the *JP Adv* group, right: the *JP Beg* group). The result showed significantly greater duration of the stressed vowels in the one-syllable feet than the two-syllable and the three-syllable feet in all proficiency groups (*p*s < .001). Now the result is consistent with the previous studies (e.g. Lehiste 1972, Rakerd et al. 1987) which observed compensatory shortening of the stressed vowels. However, no such shortening of stressed vowels was observed when the two-syllable and the three-syllable feet were compared. The duration of the stressed vowels was not significantly different in the *EN* and the *JP Beg* groups (*p*s > .8). In the *Adv JE*, the duration of the stressed vowels in the two-syllable feet had a marginally significant difference from that of the stressed vowels in the three-syllable feet. However, the mean duration of the latter was larger, which does not imply compensatory shortening. This is comparable to Mochizuki-Sudo and Kiritani (1991), in which the shortening of stressed syllables was observed when one-syllable and two-syllable feet were compared in the *NE* and *JE* speech but not when three-syllable or four-syllable feet were compared.

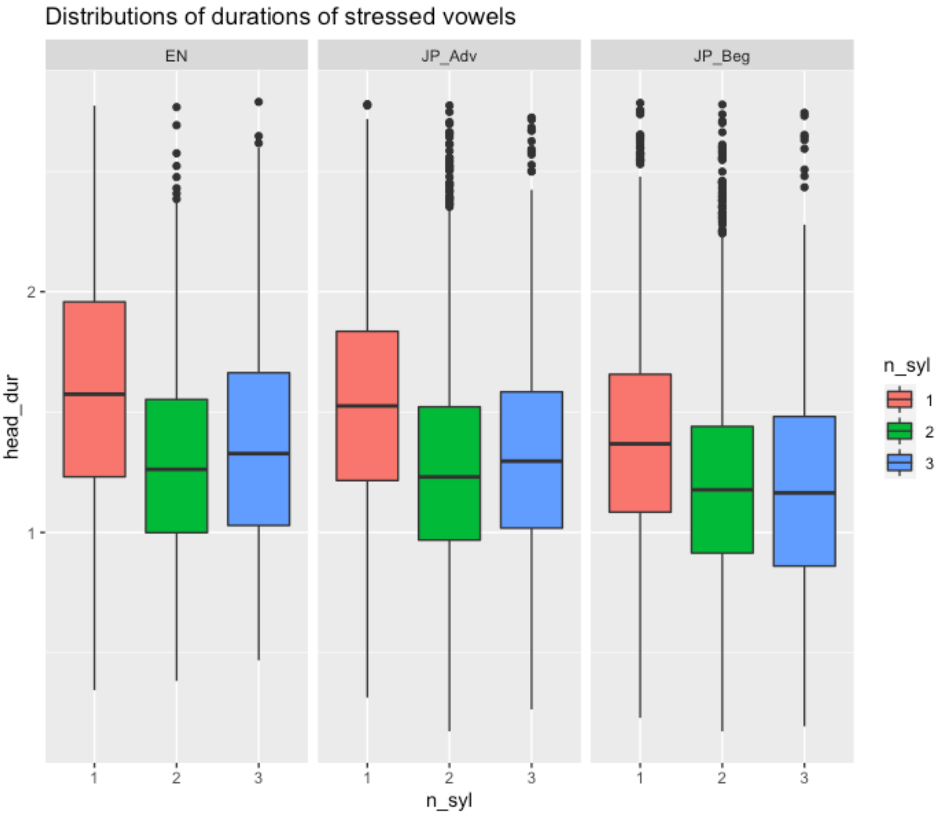


Figure 28 Duration of stressed vowels in relation to the number of syllables in the foot

On the other hand, unlike the result of Mochizuki-Sudo and Kiritani (1991), which observed greater compensatory shortening of stressed vowels in the *NE* and the *Adv JE* than the *Beg JE*, the current data showed no such difference. Figure 29 compares the mean duration of stressed vowels between the speaker groups (red: the *NE* group, green: the *JP Adv* group, blue: the *JP Beg* group). Indeed, the mean duration of the stressed vowels in the *NE* and the *Adv JE* were significantly greater than those in the *Beg JE* regardless of the number of syllables (e.g. the mean duration of the stressed vowels in the one-syllable feet were significantly greater in the *NE* and the *Adv JE* than in the *Beg JE*; *p*s <.001) there was no considerable difference between the speaker groups when the data were converted to ratios. Table 35 summarises the mean duration of stressed vowels in relation to the number of foot-internal syllables. The leftmost column shows the number of syllables in the foot. For the sake of comparison, the mean duration of the stressed vowels was divided by the mean duration of stressed vowels in the one-syllable foot. The proportion of the increase in the mean duration of stressed syllables from the one-syllable feet to the three-syllable feet was 0.856 in the *NE*, 0.861 in the *Adv JE* and 0.852 in the *Beg JE.*

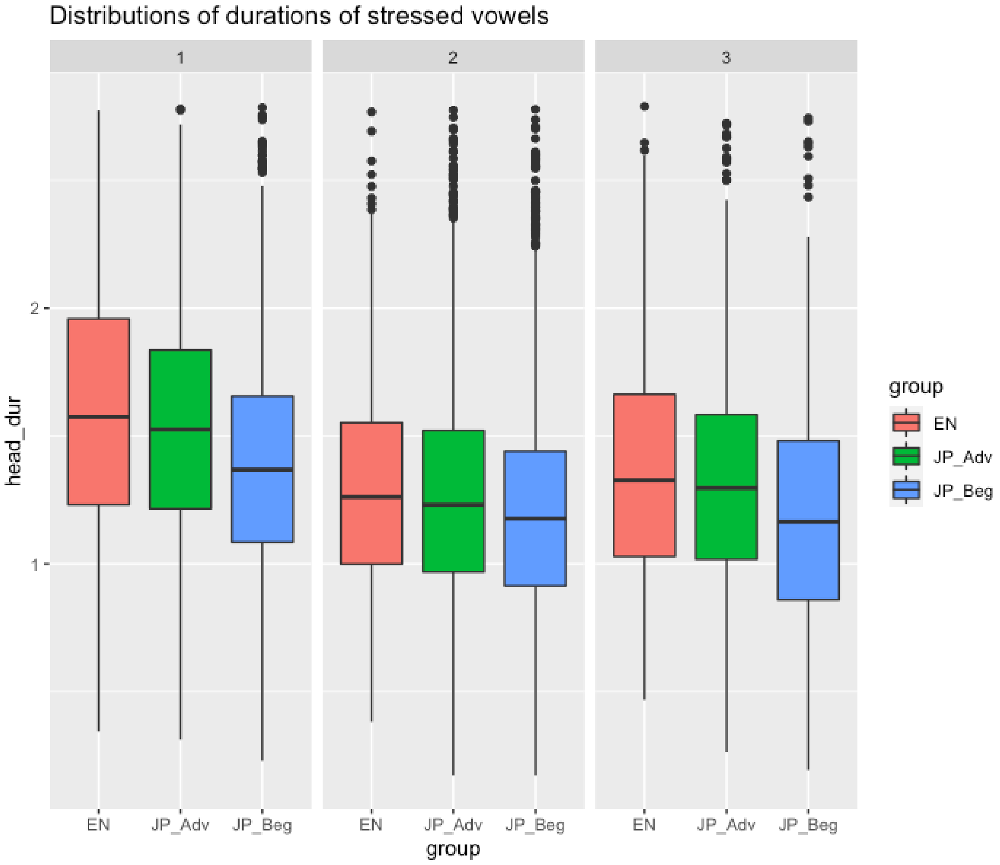


Figure 29 Duration of stressed vowels in relation to the number of syllables in the foot

Table 35 Mean normalised duration of the stressed vowels in relation to the number of foot-internal syllables

	EN (n = 1,399)	**JP Adv (n = 5,109)**	**JP Beg (n = 5,104)**
1	1.000	1.000	1.000
2	0.809	0.826	0.866
3	0.856	0.861	0.852

Although compensatory shortening was observed for foot-internal unstressed syllables, there was no considerable group difference. In fact, the total duration of foot-internal unstressed syllables negatively correlated with the speakers' proficiency (Figure 21; See Section 7.3.4). This seems to have contributed to the smallest increase in the foot duration by additions of extra unstressed syllables in the *NE,* followed by the *Adv JE* (Section 7.3.2). However, when the duration was converted to ratios (Table 31; See Section 7.3.4), no considerable difference was observed between the proficiency groups; the normalised total duration of the unstressed syllables in the three-syllable foot of the *EN* group (1.714) was not notably different from that of the *JP Adv* group (1.691) and the *JP Beg* group (1.712). That is, compensatory shortening was observed even for the *JP Beg* group when the unstressed syllables in two-syllable and three-syllable feet were compared. With the assumption of the transfer of Japanese mora rhythm, the presence of the compensatory shortening in the *JE* unstressed syllables and the absence of any notable group difference were both quite unexpected. Perhaps the *JP Beg* group in the current study was proficient enough to manifest some compensatory shortening when there are more syllables in the foot.

Summing up the results of compensatory shortening s, neither Hypotheses III nor IV was supported (Table 36). Although compensatory shortening was observed on foot-internal stressed vowels, it was not observed on the stressed syllables. Also, the shortening was not consistently observed when vowels in two-syllable and three-syllable feet were compared. On the other hand, the shortening was observed on foot-internal unstressed syllables of all three groups when two-syllable and three-syllable feet were compared. However, there was no considerable difference between the speaker groups either for the stressed or unstressed syllables.

Table 36 The hypotheses and results of the analyses

Hypothesis III	Greater compensatory shortening is observed for the syllables of the *NE* feet than those of the *JE* feet.	**Not supported**
Hypothesis IV	Greater compensatory shortening is observed for the syllables of the *Adv JE* feet than those of the *Beg JE* feet.	**Not supported**

Lastly, the result of the analysis revealed different relative contributions of stress contrast and vowel epenthesis on the variance of the foot duration in the *Adv JE* and *Beg JE* (Research Question I; Table 37 below). Vowel epenthesis affected the foot rhythm only in the *Beg JE* speech (p < .01). Its effect on the variance of the foot duration was positive perhaps because epenthetic vowels would increase the duration of some feet by adding extra duration (Section 7.3.5). In contrast, the degree of stress contrast manifested by the vowel duration was the only correlate of the isochronous rhythm of the *Adv JE* feet (p < .01). Its effect on the variance of the foot duration was negative so that more isochronous foot rhythm was manifested when greater stress contrast was made. The factor also affected the foot rhythm of the *Beg JE* in relation to vowel epenthesis (*p*

< .05). As shown in Figure 27 (Section 7.3.5), the effect of the interaction of the two factors was negative overall, implying more isochronous foot rhythm when the stress contrast was greater, as observed in the *Adv JE*. The effect was assumed to be stronger for lower-proficiency speakers who epenthesise vowels frequently. Previous studies (Konishi et al. 2018, Saito et al. 2016) observed different relative contributions of segmental and suprasegmental factors on the perceived proficiency of L2 English between *Adv JE* and *Beg JE*. The current study also found different relative contributions of the phonetic factors in relation to the learners' proficiencies.

Table 37 The research question and the result of the analysis of the foot duration

Research Question I	What are the phonetic correlates of the *JE* foot rhythm?
Result	Clear lexical stress contrast with the vowel duration positively contributed to the isochrony of the foot rhythm in the *Adv JE,* and in the *Beg JE* when there was more frequent vowel epenthesis. Vowel epenthesis negatively contributed to the isochrony of the foot duration only in the *Beg JE*.

Altogether, the results indicate that vowel epenthesis primarily affects the isochrony of the *JE* foot rhythm of low-proficiency speakers. This suggests the importance of clear stress contrast in manifesting isochronous foot rhythms in both *Adv JE* and *Beg JE.* This is further supported by Figure 30 below, which also shows the distributions of the CV of the foot duration of the *Beg JE*. The x-axis is the number of epenthetic vowels summarised for each speaker and the y-axis is the CV of the foot duration. The red dots are the data of the speakers whose manifested stress contrasts were greater than or as great as the mean. The blue dots are the data of the speakers with smaller than the mean degree of stress contrast. This implies a quite small effect of vowel epenthesis on the variance of the foot duration when stress contrast is made appropriately. That is, even among the *JP Beg* group, the effect of vowel epenthesis seems to be limited to the speech of lower-proficiency speakers, who manifests smaller stress contrasts with the vowel duration.

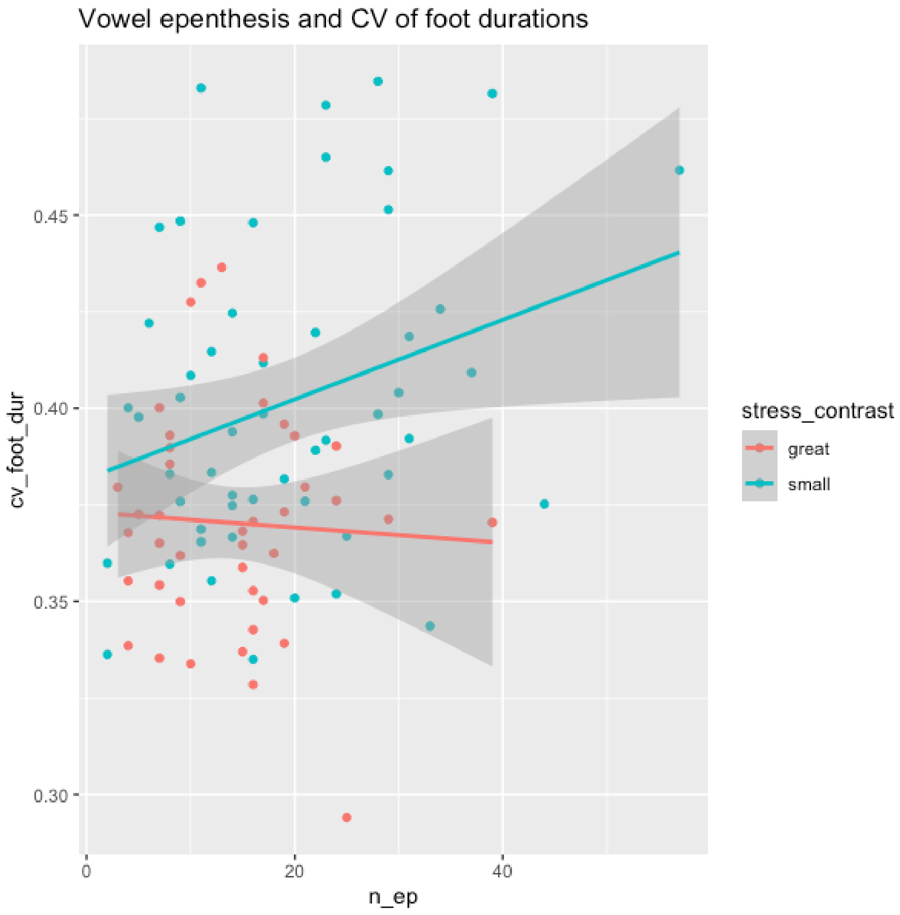


Figure 30 Variance of the foot duration and the number of epenthetic vowels in relation to the stress contrast (the JP Beg group)

7.5. Summary

In this chapter, *JE* foot rhythm was investigated in relation to speakers' proficiency and the results of the other analyses in this dissertation, i.e. lexical stress (Chapter 5) and vowel epenthesis (Chapter 6). With the assumption of trochaic bias (Cutler 1989; See Section 2.2.1), the feet were defined to consist of a stressed syllable followed by zero to three unstressed syllables. The analysis of the variance of the foot duration showed that more isochronous foot rhythm was manifested by the *NE* than *JE*. In addition, the *Adv JE* manifested more isochronous feet than the *Beg JE*.

Foot duration was further investigated in relation to the number of foot-internal syllables. Again, the result showed smaller variance of feet in the *NE*, i.e. greater duration of the one-syllable feet and smaller duration of the three-syllable feet, followed by the *Adv JE.* Unlike previous studies (Fowler 1977, 1981, Huggins 1973, Lehiste 1972, Rakerd et al. 1987) which observed compensatory shortening on the foot-internal stressed syllables in relation to the number of syllables in the foot, the results of the current analyses did not observe such shortening consistently. Although the shortening was observed in foot-internal unstressed syllables, no considerable difference was found in relation to speaker proficiency when the duration was normalised. Overall, speaker proficiency affected variance of the foot duration but not compensatory shortenings on either stressed or unstressed syllables.

Lastly, the effect of lexical stress on the realisations of isochronous foot rhythm was overall larger than that of epenthetic vowels. In the *Adv JE* speech, stress contrast had a statistically significant influence on the foot rhythm. However, no significant effect of the number of epenthetic vowels was observed. In contrast, epenthetic vowels but not stress contrast had a significant main effect on the foot rhythm of the *Beg JE*. Furthermore, the significant interaction of the two factors indicates a greater effect of stress contrast for lower-proficiency speakers in the *JP Beg* group. As speakers in the

JP Beg group become more proficient and manifest greater stress contrasts, the effect of vowel epenthesis will be much smaller.

8. General Discussions

8.1. Perceived proficiency of Japanese English rhythm

In order to investigate *JE* foot rhythm, the current study first aimed at providing an appropriate measure of the speech proficiency of *JE*. Most previous studies of *JE* pronunciations relied on some objective measures of general English proficiency (e.g. scores of some proficiency tests, length of residence in English-speaking countries and ages of arrivals in the country) that do not only represent the goodness of pronunciation. Furthermore, most of the few studies of human ratings of *NNE* speech relied on so-called *native speaker norms* in evaluating pronunciation. With the majority of English speakers in the 21st century being nonnatives (Bolton 2004, Crystal 2003), the current study employed *NNE* speaking raters as well as *EN* raters.

The current study investigated human ratings of the *JE* utterances of *the North Wind and the Sun* extracted from the *J-AESOP* corpus (Chapter 4). It examined three phonetic factors, i.e. *segmental accuracy, prosody,* and *fluency*, that are estimated to affect *NNE* for its intelligibility and *nativelikeness* (the other commonly used measure of *NNE* speech; See Section 4.3 for the reasons of choosing the term over the more commonly used *accentedness*). The raters consisted of four *EN*, four *JP* and eight native speakers of other languages (*OT*), all of whom had completed a postgraduate course in either phonetics, second language acquisition or a related field. With the assumption of graduate-level knowledge of phonetics and phonology among the raters, the current study aimed to avoid potential disagreements on ratings of linguistic categories by experienced and inexperienced raters (Saito et al. 2017).

The results of the analysis of the scores showed high inter-rater consistency for all four measures. The Cronbach's Coefficient Alpha values (Cronbach 1970) for all four categories were higher than .89, indicating high inter-rater consistencies of the ratings. When all 16 raters were compared, excluding any one of the individual raters did not result in a higher alpha value, suggesting no considerable effect of raters' first language. Among the three phonetic factors, i.e. *segmental accuracy, prosody* and *fluency*, the effect of *segmental accuracy* was strongest in the *EN* and *JP* raters' perception of *nativelikeness.* The relative effect of *segmental accuracy* was stronger for the *EN* rater group while that of *fluency* was stronger for the *JP* rater group. Of the four kinds of scores, the *prosody* score was used for the remaining analyses in this dissertation.

Although there were no objective criteria to evaluate the accuracy of the ratings other than investigating the inter-rater consistency, the results of the analyses using the rated scores (Chapters 5, 6 & 7) were mostly consistent with those of the previous studies which adopted different methods of dividing *JE* into different proficiency groups. Potential differences could be in the duration and intensity of vowels in relation to the presence or absence of lexical stress (Sections 5.3.1 and 5.3.2). In contrast to the previous study by Lee et al. (2006), in which no significant difference in the use of duration to manifest lexical stress contrast was observed between the *EN* group and the *JP Adv* group, the result of the current study indicated significantly greater stress contrast with the vowel duration by the *EN* group than by the *JP Adv* group. A difference also existed between the result of the current study and that of Konishi and Kondo (2015); in the current study, the durational contrast of lexical stress was observed for unstressed vowels (e.g. the duration of unstressed vowels produced by the *EN* group were significantly smaller than those produced by the *JP Adv* group) while in Konishi and Kondo (2015), the group difference was observed for stressed vowels (the duration of stressed vowels produced by the *EN* group were significantly greater than those produced by the *JP Adv* group). Although those contrastive results might have been due to different measures used to divide the subjects into proficiency groups, they might as well have been due to different proficiencies of the *JP* subjects, e.g. all *JE* subjects in the Lee et al.'s (2006) study were highly proficient (referred to as early or late "Japanese-English bilinguals"). Despite those differences, general tendencies of the acquisition of *JE* stress are the same across the studies, e.g. more proficient speakers manifest greater stress contrast with the vowel duration and the vowel intensity. In addition, the result of the analysis concerning the frequency of *JE* vowel epenthesis (Section 6.3.1) was also mostly consistent with those of the previous studies. In those respects, the scores used in the current study, which consist of both *NE* and *NNE* ratings, should be valid for phonetic investigations.

8.2. Isochrony in English rhythm

The main goal of this dissertation is to investigate the acquisition of *NE* and *JE* foot rhythms, focusing on isochrony in their speech. The isochrony in the *NE* foot rhythm has been said to be *mental,* rather than *physical*. Although previous studies indicated attempts towards isochronous rhythm in native speakers' utterances (Fowler 1977, 1981, Huggins 1973, Lehiste 1972, Rakerd et al. 1987), as well as native listeners' inclination to hear isochronous rhythm (Lehiste 1975, 1977), no study hitherto has successfully shown the physical realisation of isochrony in the *NE* speech (Cruttenden 2014).

Although the rhythmic unit of Japanese is different from that of English (mora in Japanese and foot in English), its rhythm has also been discussed in relation to isochrony. It is well known that the duration of utterances in Japanese speech tends to be proportional to the number of morae (Bradlow et al. 1995, Port et al. 1987). Despite some phonetic obstacles in the realisations of isochronous mora rhythm (e.g. special morae are shorter than other morae; Arai 1999, Beckman 1982, Campbell and Sagisaka 1991), Japanese mora rhythm is known to be much more isochronous than English foot rhythm. In contrast to English foot rhythm, which is manifested with lexical stress, Japanese rhythm is not affected by the presence or absence of lexical pitch accent either.

The differences in the rhythmic units in English and Japanese, as well as their different syllable structures, would result in the transfer of the Japanese rhythm to the *JE* rhythm. The L1 transfer would make the *JE* rhythm less isochronous since Japanese speech is more proportional to the number of morae than English speech is to the number of syllables. Despite the potential negative transfer, previous studies on *JE* prosody did not focus on the manifestations of isochronous rhythm. Most of them investigated other prosodic factors which influence English foot rhythm, e.g. lexical stress (Lee et al. 2006, Konishi & Kondo 2015) and consonant and vowel intervals (Grenon & White 2008, Kawase et al. 2016, Ozaki et al. 2017). Mochizuki-Sudo and Kiritani (1991) examined the relationship between the foot duration and the number of syllables but they did not investigate the variance of the foot duration regardless of the number of foot-internal syllables.

Since the aim of the current study is to investigate *physical*, rather than *mental*, isochrony in the English foot rhythm, the variance of the foot duration was first investigated not taking into consideration the number of foot-internal syllables (Section 7.3.1). This analysis contrasts with those of the previous studies which examined the mean and the variance of the duration of each foot interval rather than the variance of the duration across different foot intervals. This is perhaps because previous studies investigated a small number of subjects and test sentences. Mochizuki-Sudo and Kiritani (1991) investigated 10 *JP* subjects and 16 types of feet (four types one-syllable to four-syllable feet). Mori et al. (2014) examined 42 *JP* subjects and three types of feet. By contrast, the current study is a corpus-based one that investigated 183 *JP* subjects and 46 different types of feet (16 monosyllabic, 18 disyllabic and 12 trisyllabic feet) (See Table 29 in Section 7.2). The investigation of the variance was possible because of the large data size.

The result of the analysis showed smaller variance of the foot duration in more proficient speech. When the coefficient of variation (CV) of the foot duration was compared between the proficiency groups, the *NE* feet indeed had a smaller variance of the duration than the *JE Adv* feet, and the *JE Adv* feet also had a smaller variance of the duration than the *JE Beg* feet. Hence, it was statistically shown that the *NE* foot is indeed physically more isochronous than the *JE* foot.

In addition, the results in the current study indicated greater proportional increases in the foot duration in relation to the number of foot-internal syllables in less proficient English (Section 7.3.2). The effect of the number of syllables on the increase in the foot duration was smallest in the *NE*, then in the *Adv JE*. The *NE* had the greatest duration of the one-syllable feet but the smallest duration of the three-syllable feet. Hence they had the smallest proportional increase. On the other hand, the *Beg JE* had the smallest duration of the one-syllable feet and the largest duration of the three-syllable feet, hence the largest proportional increase. The increase in the foot duration in the *Adv JE* was in between the two. The result is consistent with that of Mochizuki-Sudo and Kiritani (1991), although statistical significance was not computed in their study. The proportion of the mean duration of the three-syllable feet to that of the one-syllable feet was 1.775 in the *NE*, 1.949 in the *Adv JE* and 2.056 in the *Beg JE*.

These two findings are quite important in that they suggest the *NE* foot rhythm is indeed isochronous when compared to less proficient speech, which has some interference of the L1 rhythm (*JE* in the current study). Previous studies have failed to show physical isochrony in the English foot but attempted to support its existence either by controlling the speech environment (e.g. Tajima 1998) or investigating native speakers' attempts to manifest isochronous rhythm as well as their preferences to hear it (Fowler 1977, 1981, Huggins 1973, Lehiste 1972, 1975, 1977, Rakerd et al. 1987). However, it has not been clear what extent of isochrony proves the existence of an isochronous rhythm in a language. The measures proposed by Ramus et al. (1999), i.e. ∆V (the standard deviation of the vowel duration), ∆C (that of the consonant duration) and %V (the proportion of vowel intervals), and those normalised for the speech rate, i.e. VarcoV (the coefficient of variations of the vowel duration) and VarcoC (that of the consonant duration), have successfully presented cross-linguistic differences of speech rhythm, e.g. English has greater VarcoV and VarcoC than Japanese. Indeed, Kawase et al. (2016) showed VarcoV in the *JE Beg* was significantly smaller than that of the *NE* and the *Adv JE.* Although it suggests greater durational contrast is made in the *NE* and the *Adv JE* than the *Beg JE*, it does not necessarily indicate more isochronous feet in the *NE* and the *Adv JE.* By actually comparing the duration of the *NE* feet to that of the *JE* feet, the current study has demonstrated that the *NE* feet are indeed more isochronous than the *JE* feet.

Perhaps, the smaller variance of the *NE* foot duration has been due to the smaller duration of the lexically unstressed syllables in the *NE* compared to the *JE*. Both *NE* and *JE* foot duration correlated with the number of syllables (Section 7.3.2). The results of the analyses of lexical stress showed significantly smaller duration of lexically unstressed vowels in the *NE* than the *Adv JE* and in the *Adv JE* than the *Beg JE* (Sections 5.3.1). This means the increase in foot duration when an extra syllable is added to the foot is statistically smallest in the *NE*, followed by the *Adv JE.*

Compared to vowel epenthesis, lexical stress contrast was shown to have a greater influence on the manifestations of the isochronous foot rhythm of *JE*. It was the only parameter that had a statistically significant influence on the variance of the foot duration in the *Adv JE*, i.e. the variance of the foot duration was smaller when there was a greater degree of shortening of unstressed vowels. In contrast, the effect of epenthetic vowels seems to be limited to low-proficiency speakers. Although they influenced the manifestation of isochronous feet in the speech of lower-proficiency speakers in the *JP Beg* group, their effect was much smaller for speakers who manifested greater stress contrasts (Section 7.3.5).

The current study also examined the most well-known phenomenon among the attempts for physical isochrony, i.e. compensatory shortening. It is known to be observed especially on lexically stressed syllables (e.g. Fowler 1977, 1981, Huggins 1973, Lehiste 1972, Rakerd et al. 1987), i.e. the duration of the foot-internal stressed syllables negatively correlates with the number of syllables in the foot. The results of the current study indicated an absence of any considerable effect of the number of foot-internal syllables on the duration of the stressed syllables (Section 7.3.4), which contrasts with the results of the previous studies. The inconsistency might be due to different experimental settings. For example, Rakerd et al. (1987) found compensatory shortening of foot-internal stressed syllables under a more controlled setting; the duration of the same syllable was compared with or without following unstressed syllables. Similarly, Fowler (1981) used the metronome beat, encouraging the subjects to compress the syllable duration. In addition, the current study found that, when the shortening was investigated on foot-internal stressed vowels, rather than syllables, significant differences were observed between the duration of the one-syllable and two-syllable feet in all proficiency groups (Section 7.4). This result is consistent with Mochizuki-Sudo and Kiritani (1991). However, neither in Rakerd et al. (1987) nor in the current study was the shortening observed when the two-syllable feet and the three-syllable feet were compared for any of the proficiency groups.

The compensatory shortening was also investigated on foot-internal unstressed syllables and vowels. Although the result indicated shortening of foot-internal unstressed syllables in relation to the number of syllables, there was no considerable difference between the proficiency groups (Sections 7.3.4). If the rhythm of Japanese, whose duration is proportional to the number of morae (Bradlow et al. 1995, Port et al. 1987, Sato 1995), had transferred to the *JE* rhythm, the duration of the *JE* unstressed syllables would have been more proportional to the number of syllables. Therefore, the absence of difference even between the *NE* and *Beg JE* was quite unexpected. Perhaps the majority of the *JP Beg* in the current study had already overcome the L1 interference.

To sum up, the results concerning compensatory shortening are rather difficult to interpret. Although the shortening was observed on the foot-internal unstressed syllables, there was no developmental change observed on the shortening. In addition, no shortening was observed on the foot-internal stressed syllables in the current study but was observed in Rakerd et al. (1987), who investigated the phenomena in a more phonetically restricted environment. Regarding foot-internal stressed vowels, both the current study and Mochizuki-Sudo and Kiritani (1991) observed shortening when the one-syllable and the two-syllable feet were compared. However, in neither of the studies was the shortening consistently observed when the two-syllable and the three-syllable feet were compared.

8.3. Implications for teaching English rhythm to Japanese learners

The results of the current study have some important implications for teaching English rhythm to *JP* learners. Firstly, characteristics of *JE* prosody appear to vary depending on the learner's proficiency level. Especially, vowel epenthesis had an influence on the isochronous foot rhythm only when the proficiency of the *JP* is quite low. (In the current study, the influence was observed for lower-proficiency speakers in the *JP Beg* group.) As the *JP* learner acquires greater command of the vowel duration to manifest stress contrast, the effect of epenthetic vowels will not be observed. An implication for teaching English to *JP* learners is that pronunciation instructions to *JP Beg* should focus more on the vowel duration to properly manifest lexical stress contrast than on how to avoid vowel epenthesis. The instruction of stress contrast should also be effective on *JP Adv* since it was the only correlate of the foot rhythm.

Another important finding is that the effect of lexically unstressed vowels on *JE* foot rhythm seems to be greater than that of lexically stressed vowels. The speech of more proficient speakers had smaller variance of the foot duration. Since the effect of proficiency was observed in the duration of lexically unstressed, rather than stressed, vowels, what seems to have contributed to the isochrony in the foot rhythms is how much durational reduction was made in unstressed vowels. The duration and intensity of the unstressed vowels were both significantly smaller in the *NE* than the *Adv JE* and in the *Adv JE* than the *Beg JE*. This implies smaller positive increases of the foot duration in relation to the number of foot-internal syllables in more proficient speech. Although there was a difference in the intensity of stressed vowels between the proficiency groups, no group difference was observed for the duration.

Lastly, the scores rated by *NE* and *NNE* raters in the current study properly predicted proficiency of *JE* speech. Although not all results were completely consistent with those of previous studies, most results in the current study are at least explicable by existing theories of phonetics and second language acquisitions. It is hoped that more future studies will adopt measures to evaluate L2 English from the perspective of *English for International Communication* rather than aiming at the acquisition of *nativelike* English.

9. Conclusions

This dissertation investigated the foot rhythm in Japanese English based on proficiency scores rated by native and nonnative English-speaking raters. It investigated the effects of two phonetic correlates of Japanese English, i.e. lexical stress and vowel epenthesis, on perceived proficiency as well as on manifestations of isochrony in the foot rhythm. The most important finding of the dissertation is that English foot manifested by native speakers is indeed isochronous especially when compared to the foot rhythm of less proficient speech, i.e. Japanese-accented English. This contrasts with previous studies, none of which supported the existence of physical isochrony in English speech except in some limited environments. The current study demonstrated a statistically less isochronous rhythm in the Japanese English compared to the native-speaker English, perhaps due to the transfer of the Japanese phonotactics and the mora-timed rhythm. The two primary phonetic correlates of the isochronous foot rhythm in Japanese English were the durational contrast of lexical stress and vowel epenthesis. Overall, the importance of the former seems to be greater in the speech of both advanced and beginner speakers of Japanese English.

The results of the investigations in the dissertation have two important implications for future research. Firstly, since the current study focused on Japanese learners' acquisition of English, the results may not apply to the acquisition of English rhythm by native speakers of other languages. It is hoped that future studies will investigate manifestations of English foot rhythm by native speakers of other languages, especially those with different rhythmic structures and phonotactics from Japanese. Another implication is for the field of second language teaching. The dissertation found the importance of the durational reductions of lexically unstressed vowels in manifesting isochronous foot rhythm, especially in lower-proficiency speech. It is hoped that future research will investigate whether instructing proper manifestations of the isochronous rhythm or the durational reduction of unstressed vowels is more effective in manifesting more isochronous rhythm.

References

Abercrombie, D. (1967). *Elements of general phonetics*. Aldine Publishing Company.

Aldrich, A. (2020). Adult Early-Bilingual Speech Rhythm: Evidence from Spanish and English. In *Proceedings of the 10th International Conference on Speech Prosody 2020*, 528-532.

Anderson, P. J. (1993). *The interstress interval as an indicator of perceived intelligibility among nonnative speakers of English.* Doctoral dissertation, Wichita State University, Kansas.

Arai, T. (1999). A case study of spontaneous speech in Japanese. In *Proceedings of the international congress of phonetic sciences (ICPhS)*, 615-618. Berkeley, CA: Department of Linguistics, University of California.

Archibald, J. (1997). The acquisition of English stress by speakers of nonaccentual languages: Lexical storage versus computation of stress. *Linguistics, 35,* 167-181.

Beckman, M. E. (1982). Segment duration and the 'mora' in Japanese. *Phonetica*, *39*(2-3), 113-135.

Beckman, M. E. (1986). Stress and Non-Stress Accent. In Van den Broecke, M.P.R and van Heuven., VJ (Eds). Netherlands Phonetic Archives; 7. Foris publications, USA.

Beckman, M. E. (1992). Evidence for speech rhythms across languages. *Speech perception, production and linguistic structure*, 457-463.

Beckman, M. E., & Ayers, G. (1997). Guidelines for ToBI labelling. The OSU Research Foundation, 3.

Beckman, M. E., & Edwards, J. (1994). Articulatory evidence for differentiating stress categories. In P. Keating (Ed.), *Phonological structure and phonetic form: Papers in laboratory phonology III* (pp. 7–33). Cambridge: Cambridge University Press.

Beckman, M. E., & Pierrehumbert, J. B. (1986). Intonational Structure in Japanese and English. Phonology, 3(01), 255-309.

Bekku, S. (1977) *Nihongo no Rizumu* [The rhythm of Japanese]. Tokyo: Kodansha.

Bolinger, D. L. (1958). A theory of pitch accent in English. *Word*, *14*(2-3), 109-149.

Bolinger, D. (1981). *Two kinds of vowels, two kinds of rhythm* (Vol. 289). Indiana University Linguistics Club.

Bolton, K. (2004). World Englishes. In A. Davies & C. Elder (eds), *Handbook of Applied Linguistics*. Oxford: Blackwell, 367-396.

Bradlow, A., Port, R. F. & Tajima, K. (1995). The combined effects of prosodic variations on Japanese mora timing. In *Proceedings of XIIIth International Congress of Phonetic Sciences, Stockholm, Sweden, 1995*.

Burzio, L., & Luigi, B. (1994). *Principles of English stress* (No. 72). Cambridge University Press.

Campbell, N. & Sagisaka, Y. (1991). *Onsei taimingu ni mirareru moora to onsetsu no eikyou ni tsuite* [Moraic and Syllable-Level Effects on Speech Timing]. *IEICT Technical Report, SP 90-107.* 35-40.

Cole, D., & Miyashita, M. (2008). The function of pauses in metrical studies: acoustic evidence from Japanese verse. In *Formal Approaches to Poetry*. 173-192. De Gruyter Mouton.

Cronbach, L. J. (1970). *Essentials of psychological testing* (3rd ed.). New York: Harper & Row.

Cruttenden, A. (1997). *Intonation.* Cambridge University Press.

Cruttenden, A. (2014). *Gimson's Pronunciation of English* (8th ed.). New York: Routledge.

Cutler, A. (1989). Auditory lexical access: where do we start? In *Lexical representation and process*, 342-356. MIT Press.

Cutler, A., & Butterfield, S. (1990). Durational cues to word boundaries in clear speech. *Speech Communication*, *9*(5-6), 485-495.

Cutler, A., & Butterfield, S. (1991). Rhythmic cues to speech segmentation: Evidence from juncture misperception. *Journal of memory and language*, *31*(2), 218-236.

Cutler, A., & Carter, D. M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech & Language*, *2*(3-4), 133-142.

Dallyn, T. D. (2018). *Hokkaido hougen no accent tokutyo ni kansuru kizyututeki kenkyuu* [Descriptive Study on Characteristics of Hokkaido-Accented Japanese]. Doctoral dissertation, Hokkaido University

Dauer, R. M. (1983). Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics*, *11*(1), 51-62.

Davidson, L. (2011). Phonetic, phonemic, and phonological factors in cross-language discrimination of phonotactic contrasts. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(1), 270-282.

Davidson, L., Martin, S., & Wilson, C. (2015). Stabilizing the production of nonnative consonant clusters with acoustic variability. *The Journal of the Acoustical Society of America*, *137*(2), 856-872.

de Jong, K. (2004). Stress, lexical focus, and segmental focus in English: Patterns of variation in vowel duration. *Journal of Phonetics*, *32*, 493–516.

de Jong, K., Beckman, M. E., & Edwards, J. (1993). The interplay between prosodic structure and coarticulation. Language and Speech, 36(2,3), 197–212.

Dellwo, V. & P. Wagner. (2003). Relations between language rhythm and speech rate. *Proceedings of 15th International Congress of Phonetic Sciences,* 471-474.

Derwing, T. M. (2008). Curriculum issues in teaching pronunciation to second language learners. In Teoksessa & Zampini (Eds). *Phonology and Second Language Acquisition.* John Benjamins publishing company.

DeVellis, R. F. (2005). Inter-rater reliability. *Encyclopedia of Social Measurement.* 317-322.

Dörnyei, Z. (2014). *The psychology of the language learner: Individual differences in second language acquisition*. Routledge.

Dupoux, E., Kakehi, K., Hirose, Y., Pallier, C., & Mehler, J. (1999). Epenthetic vowels in Japanese: A perceptual illusion? *Journal of experimental psychology: human perception and performance*, *25*(6), 1568.

Erickson, D., Suemitsu, A., Shibuya, Y., & Tiede, M. (2012). Metrical structure and production of English rhythm. *Phonetica*, *69*(3), 180-190.

Fear, B. D., Cutler, A., & Butterfield, S. (1995). The strong/weak syllable distinction in English. *The Journal of the Acoustical Society of America, 97*(3), 1893–1904.

Féry, C. (2018). *Intonation and prosodic structure.* Cambridge University Press.

Fowler, C. A. (1977) *Timing control in speech production.* Ph.D. thesis, University of Connecticut.

Fowler, C. A. (1981). A relationship between coarticulation and compensatory shortening. *Phonetica*, *38*(1-3), 35-50.

Fry, D. B. (1955). Duration and Intensity as Physical Correlates of Linguistic Stress. *The Journal of the Acoustical Society of America*, 27(4), 765-768.

Fry, D. B. (1958). Experiments in the perception of stress. *Language and speech*, *1*(2), 126-152.

Funatsu, S., Imaizumi, S., Fujimoto, M., Hashizume, A., & Kurisu, K. (2008). Do Japanese speakers perceive nonexistent vowels in non-native consonant clusters? *Journal of the Acoustical Society of America*, *123*(5), 3072.

Graham, C., & Post, B. (2018). Second language acquisition of intonation: Peak alignment in American English. *Journal of Phonetics*, *66*, 1-14.

Greenberg, S., & Fosler-Lussier, E. (2000, May). The uninvited guest: Information's role in guiding the production of spontaneous speech. In *Proceedings of the Crest Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling*, 129-132.

Grenon, I., & White, L. (2008). Acquiring rhythm: A comparison of L1 and L2 speakers of Canadian English and Japanese. In *Proceedings of the 32nd Boston University conference on language development*, 155-166.

Guion, S. G., Harada, T., & Clark, J. J. (2004). Early and Late Spanish–English Bilinguals' Acquisition of English Word Stress Patterns. *Bilingualism: Language and Cognition, 7*(3), 207-226.

Gussenhoven, C. (2004). *The phonology of tone and intonation*. Cambridge University Press.

Hayes, B. (1995). *Metrical Stress Theory: Principles and Case Studies.* Chicago: The University of Chicago Press.

Hall, N. (2003). *Gestures and segments: Vowel Intrusion as Overlap* (Unpublished doctoral dissertation). University of Massachusetts.

Homma, Y. (1991). The rhythm of tanka, short Japanese poems: Read in prose style and contest style. In *Proceedings of the XIIth International Congress of Phonetic Sciences*, *2*, 314-317.

Huggins, A. W. F. (1973). Some Within-and Between-Word Timing Effects. *The Journal of the Acoustical Society of America*, *54*(1), 312-312.

Ikoma, M. (1993). Der Betonungszählende Rhythmus im Deutschen: Eine Akustisch-phonetische Untersuchung [The stress-counting rhythm in German: An acoustic-phonetic investigation]. *Sophia linguistica: working papers in linguistics, 33*, 197-216.

Ikoma, M. (1998). Doitsugo ni okeru rizumu no tojisei ni kansuru ichikosatsu: foot nai no onsetsucho no jikanhosho [On the rhythmic features in German: temporal factors of foot-internal syllables]. *Lingua 9 (1998), 53-66.*

Kawase, S., Kim, J., & Davis, C. (2016). The influence of second language experience on Japanese-accented English rhythm. *Proceedings of Speech Prosody 2016*, 746-750.

Knight, R-A. (2012). *Phonetics: A Coursebook*. New York: Cambridge University Press.

Kochanski, G., Grabe, E., Coleman, J., & Rosner, B. (2005). Loudness Predicts Prominence: Fundamental Frequency Lends Little. *The Journal of the Acoustical Society of America, 118*(2), 1038-1054.

Kohler, K. J. (2009). Rhythm in speech and language. *Phonetica*, *66*(1-2), 29-45.

Komatsu, M., & Aoyagi, M. (2005). Vowel Devoicing vs. Mora-Timed Rhythm in Spontaneous Japanese-Inspection of Phonetic Labels of OGI_TS. In *Ninth European Conference on Speech Communication and Technology*.

Kondo, M. (2012). Design and Analysis of Asian English Speech Corpus — How to Elicit L1 Phonology in L2 English Data —. In *Developmental and Crosslinguistic Perspectives in Learner Corpus Research, 4*, 251-278.

Kondo, Y. (2000). Production of schwa by Japanese speakers of English: An acoustic study of shifts in coarticulatory strategies from L1 to L2. In Broe & Pierrehumbert (Eds) *Papers in laboratory phonology V. Acquisition and the lexicon*, 29-39. New York: Cambridge University Press.

Konishi, T. & Kondo, M. (2015). Developmental Change in English Stress Manifestation by Japanese Speakers. In *Proceedings of the International Congress of Phonetic Sciences XVIII.* Glasgow, UK.

Konishi, T., Yun, J. & Kondo, M. (2018) Acoustic correlates of L2 English stress — Comparison of Japanese English and Korean English. In *Phonetics and Speech Sciences, 10*(1). Seoul: Korean Society of Speech Sciences Press, 9-14.

Kubozono, H. (1998). *Onseigaku, Oninron* [Phonetics and Phonology]. Kuroshio-shuppan.

Kubozono, H. (1999). *Nihongo no onsei – Gendai gengogaku nyumon* [Sounds of Japanese – Introduction to Contemporary Linguistics]. Iwanami-shoten. Tokyo, Japan.

Labrune, L. (2012). *The phonology of Japanese*. Oxford University Press.

Ladd, D. R. (2008). *Intonational Phonology.* Cambridge University Press.

Ladefoged, P. (1999). American English. In International Phonetic Association (Ed.), *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet,* 41-44. Cambridge University Press.

Ladefoged, P., & Johnson, K. (2011). *A Course in Phonetics. (6th Ed.)* Wadsworth, Cengage learning.

Lea, W. A., Medress, M. F. & Skinner, T. E. (1972). Prosodic aids to speech recognition: I. Basic Algorithms and Stress Studies. *Univac DSD* (Report No. PX7940). St Paul, Minesota.

Lee, B., Guion, S. G., & Harada, T. (2006). Acoustic Analysis of the Production of Unstressed English Vowels by Early and Late Korean and Japanese Bilinguals. *Studies in Second Language Acquisition, 28*(3), 487-513.

Lehiste, I. (1972). The timing of utterances and linguistic boundaries. *The Journal of the Acoustical Society of America*, *51*(6B), 2018-2024.

Lehiste, I. (1973). Rhythmic units and syntactic units in production and perception. *The Journal of the Acoustical Society of America*, *54*(5), 1228-1234.

Lehiste, I. (1974). The timing of utterances and linguistic boundaries. *Speech and Hearing Science: Selected Readings*, *51*(6), 20.

Lehiste, I. (1975). The perception of duration within sequences of four intervals. In *Proceedings of the 8th International Congress of Phonetic Sciences.* Leeds, UK.

Lehiste, I. (1977). Isochrony reconsidered. *Journal of Phonetics, 5*(3), 253-263.

Lieberman, P. (1960). Some acoustic correlates of word stress in American English. *The Journal of the Acoustical Society of America*, *32*(4), 451-454.

Masuda, H. & Arai, T. (2010). Processing of consonant clusters by Japanese native speakers: Influence of English learning backgrounds. *Acoustical Science and Technology, 31*(5). 320- 327.

Mattys, S. L. (2000). The perception of primary and secondary stress in English. *Perception & Psychophysics*, *62*(2), 253-265.

Mazuka, R., Cao, Y., Dupoux, E. & Christophe, A. (2011). The development of phonological illusion: A cross-linguistic study with Japanese and French infants. *Developmental Science* 14(4), 693-699.

McCarthy, J. J. (2006). Prosodic morphology. In K. Brown (Ed), *Encyclopedia of Language and Linguistics. 62*.

Meng, H., Tseng, C. Y., Kondo, M., Harrison, A., & Viscelgia, T. (2009). Studying L2 Suprasegmental Features in Asian Englishes: a Position Paper. In *INTERSPEECH 2009, 10th Annual Conference of the International Speech Communication Association*, 1715-1718.

Mochizuki-Sudo, M., & Kiritani, S. (1991). Production and perception of stress-related durational patterns in Japanese learners of English. *Journal of Phonetics*, *19*(2), 231-248.

Mori, Y., Hori, T., & Erickson, D. (2014). Acoustic correlates of English rhythmic patterns for American versus Japanese speakers. *Phonetica*, *71*(2), 83-108.

Mott, B. L. (2011). *English phonetics and phonology for Spanish speakers*. 2nd ed. Edicions de la Universitat de Barcelona.

Munro, M. J. (2016). Pronunciation learning and teaching: What can phonetics research tell us. In *Proceedings of the International Symposium of Applied Phonetics 2016.* 26-29.

Munro, M. J., & Derwing, T. M. (1999). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, *49*, 285–310.

Nespor, M., Shukla, M., & Mehler, J. (2011). Stress-timed vs. syllable-timed languages. In *the Blackwell companion to phonology*, 1-13.

Okada, H. (1999). Japanese. In International Phonetic Association (Ed.), *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet,* 117-119. Cambridge University Press.

Okobi, A. O. (2006). *Acoustic correlates of word stress in American English.* Doctoral dissertation, Massachusetts Institute of Technology.

Ota, M., Ladd, D. R., & Tsuchiya, M. (2003). Effects of foot structure on mora duration in Japanese? In *Proceedings of the 15th International Conference on Phonetic Sciences*, 459-462.

Otake, T. (1988). A temporal compensation effect in Arabic and Japanese. *Onsei gakkai kaiho.* The Phonetic Society of Japan, *189*. 19-24.

Otake, T. (1989). A cross linguistic contrast in the temporal compensation effect. *The Journal of the Acoustical Society of America*, *85*(S1), S96-S96.

Ozaki, Y., Yazawa, K., & Kondo, M. (2017). L2 English speech rhythm of Japanese speakers: an alternative implementation of the Varco metrics. In *Proceedings of the Phonetics Teaching and Learning Conference UCL*, 84-88.

Plag, I., Kunter, G., & Schramm, M. (2011). Acoustic correlates of primary and secondary stress in North American English. *Journal of Phonetics*, *39*(3), 362-374.

Pierrehumbert, J., & Beckman, M. (1988). Japanese Tone Structure. Linguistic Inquiry Monographs, (15), 1-282.

Pike, K. L. (1945). *The Intonation of American English*. Ann Arbor, Michigan: University of Michigan Press, 1945.

Port, R. F. (2003). Meter and speech. *Journal of phonetics*, *31*(3-4), 599-611.

Port, R. F., Al-Ani, S., & Maeda, S. (1980). Temporal compensation and universal phonetics. *Phonetica*, *37*(4), 235-252.

Port, R. F., Dalby, J., & O'Dell, M. (1987). Evidence for mora timing in Japanese. *The Journal of the Acoustical Society of America*, *81*(5), 1574-1585.

Poser, W. J. (1990). Evidence for foot structure in Japanese. *Language, 66*(1), 78-105.

Rakerd, B., Sennett, W., & Fowler, C. A. (1987). Domain-Final Lengthening and Foot-Level Shortening in Spoken English. *Phonetica*, *44*, 147-155.

Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, *73*(3), 265-292.

Roach, P. (1982). On the distinction between ‘stress-timed’ and ‘syllable-timed’ languages. *Linguistic controversies*, *73*, 79.

Roach, P. (2009) *English Phonetics and Phonology: A Practical Course* (4th Ed.) Cambridge University Press.

Roberts, L., & Meyer, A. S. (2012). Individual differences in second language learning: Introduction. *Language Learning*, *62*(Supplement S2), 1-4.

Saito, K. (2019). Individual differences in second language speech learning in classroom settings: Roles of awareness in the longitudinal development of Japanese learners’ English/ɹ/pronunciation. *Second Language Research*, *35*(2), 149-172.

Saito, K., Trofimovich, P., & Isaacs, T. (2015). Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels. *Applied Psycholinguistics*, *37*(2), 217-240.

Saito, K., Trofimovich, P., & Isaacs, T. (2017). Using listener judgments to investigate linguistic influences on L2 comprehensibility and accentedness: A validation and generalization study. *Applied Linguistics*, *38*(4), 439-462.

Sato, Y. (1995). The Mora Timing in Japanese: A Positive Linear Correlation between the Syllable Count and Word Duration. *Onsei gakkai kaiho.* The Phonetic Society of Japan, *209, 40-53.*

Shaw, J. A., & Kawahara, S. (2018). The lingual articulation of devoiced /u/ in Tokyo Japanese. *Journal of Phonetics*, *66*, 100-119.

Shibuya, Y., & Erickson, D. (2010). Consonant cluster production in Japanese learners of English. In *Proceedings of the 10th annual conference of the International Speech Communication Association.* Tokyo: Waseda University.

Shockey, L. (2003). *Sound patterns of spoken English*. Blackwell Publishing Ltd.

Skehan, P. (1991). Individual differences in second language learning. *Studies in second language acquisition, 13*(2), 275-298.

Slevc, L. R., & Miyake, A. (2006). Individual differences in second-language proficiency: Does musical ability matter?. *Psychological science*, *17*(8), 675-681.

Sluijter, A. M., & van Heuven, V. J. (1996a). Acoustic Correlates of Linguistic Stress and Accent in Dutch and American English. In *Spoken Language, 1996. ICSLP 96. Proceedings, Fourth International Conference on*, *2*. 630-633. IEEE.

Sluijter, A. M., & van Heuven, V. J. (1996b). Spectral balance as an acoustic correlate of linguistic stress. *The Journal of the Acoustical society of America*, *100*(4), 2471-2485.

Smith, L. E., & Nelson, C. L. (2019). World Englishes and issues of intelligibility. In Nelson, Proshina & Davis (Eds). *The handbook of world Englishes*, 430-446.

Jun, SA. (2005). Prosodic Typology. In Jun, SA. (ed). *Prosodic Typology: The Phonology of Intonation and Phrasing*. 430-458.

Taft, L. (1984). *Prosodic Constraints and Lexical Parsing Strategies.* Doctoral dissertation, University of Massachusetts.

Tajima, K. (1998). *Speech Rhythm in English and Japanese: Experiments in Speech Cycling.* Doctoral Dissertation, Indiana University.

Tajima, K., Erickson, D., & Nagao, K. (2000). Factors affecting native Japanese speakers' production of intrusive (epenthetic) vowels in English words. In *Proceedings of the 6th International Conference on Spoken Language Processing*, 558–561.

Tajima, K., & Port, R. F. (2003). Speech rhythm in English and Japanese. *Phonetic interpretation: Papers in laboratory phonology VI*, 317-334.

Teranishi, R. (1980). Two-moras-cluster as a rhythm unit in spoken Japanese sentence or verse. *The Journal of the Acoustical Society of America*, *67*(S1), S40-S40.

Ueyama, M., & Jun, S. A. (1996). Focus realization of Japanese English and Korean English intonation. *UCLA Working Papers in Phonetics*, 110-125.

Vance, T. J. (2008). *The sounds of Japanese*. Cambridge University Press.

Venditti, J. J. (2005). The J_ToBI Model of Japanese Intonation. In Jun, SA. (ed). *Prosodic Typology: The Phonology of Intonation and Phrasing*. 172-200.

Visceglia, T., Tseng, C. Y., Kondo, M., Meng, H., & Sagisaka, Y. (2009). Phonetic Aspects of Content Design in AESOP (Asian English Speech cOrpus Project). In

Speech Database and Assessments, 2009 Oriental COCOSDA International Conference on. 60-65. IEEE.

Visceglia, T., Tseng, C. Y., Su, Z. Y., & Huang, C. F. (2010). Interaction of lexical and sentence prosody in Taiwan L2 English. In *SLaTE Workshop*, Interspeech.

Volín, J., & Zimmermann, J. (2011). Spectral slope parameters and detection of word stress. *Proceedings of Technical Computing Prague*, 125-129.

Warner, N., & Arai, T. (2001a). Japanese mora-timing: A review. *Phonetica*, *58*(1-2), 1-25.

Warner, N., & Arai, T. (2001b). The role of the mora in the timing of spontaneous Japanese speech. *The Journal of the Acoustical Society of America*, *109*(3), 1144-1156.

Whalen, D. H., & Levitt, A. G. (1995). The universality of intrinsic F0 of vowels. *Journal of phonetics*, *23*(3), 349-366.

Yazawa, K., Konishi, T., Hanzawa, K., Short, G. & Kondo, M. (2015) Vowel Epenthesis in Japanese Speakers' L2 English. *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS XVIII)*, Glasgow, Scotland

Zhang, Y., Nissen, S. L., & Francis, A. L. (2008). Acoustic Characteristics of English Lexical Stress Produced by Native Mandarin Speakers. *The Journal of the Acoustical Society of America, 123*(6), 4498-4513.

Appendix: List of recordings in the *J-AESOP* corpus

The sentences have been designed so that italicised words and sentences will be investigated.

1) Target words in career sentences (20 sentences)

Instruction: Read each sentence once at a natural speaking rate and volume. Try not to stress or emphasize any particular word or phrase.

I said *apartment* five times.

I said *overnight* five times.

I said *misunderstand* five times.

I said *supermarket* five times.

I said *money* ten times.

I said *hospital* ten times.

I said *white wine* ten times.

I said *elevator* ten times.

I said *available* ten times.

I said *information* ten times.

I say *January* ten times.

I said *experience* ten times.

I say *California* ten times.

I said *Vietnamese* ten times.

I say *department store* ten times.

I said *morning* ten times.

I said *video* ten times.

I say *tomorrow* ten times.

I say *Japanese* ten times.

I said *afternoon* ten times.

2) Target words at prosodic boundaries (15 sentences)

Instruction: Read each sentence once at a natural speaking rate and volume. Try not to stress or emphasize any particular word or phrase.

Do you need any *money*?

Did he go to the *hospital*?

Has Jane found an *apartment*?

Can packages be shipped *overnight*?

Would you like a glass of *white wine*?

Where is the *elevator*?

When will Bill be *available*?

Who can give me the *information*?

Why do you always *misunderstand*?

Where is the nearest *supermarket*?

In December and *January*, the sun rises at seven in the *morning.*

Although Fred didn't have any *experience*, he had no trouble learning how to make a *video*.

When Sue left this evening for *California*, she said she would call me *tomorrow*.

If you want to learn *Vietnamese*, I think it will be easier than *Japanese*.

If you want to check out the new *department store*, we can go this *afternoon*.

3) Target words in narrow focus (20 sentences)

Instruction: Read each sentence once at a natural speaking rate and volume. Emphasize the word or phrase that seems appropriate for each context.

(Context: Did Bill lose everything in the robbery?)

No. His *MONEY* was taken, but they didn't take his computer.

(Context: Can doctors give blood tests at this clinic?)

No. You should go to a *HOSPITAL* for blood tests.

(Context: Can we open a branch of our office in this building?)

No. This is an *APARTMENT* building, not a commercial building.

(Context: Will 3-day delivery be fast enough?)

No. We need *OVERNIGHT* delivery.

(Context: Did you order a Coke?)

No. I ordered *WHITE WINE*, not Coke.

(Context: How will I carry all these boxes up to the fifth floor?)

You should take the *ELEVATOR* instead of the stairs.

(Context: Would you like a table by the window?)

Someone is already sitting there. Are there any *AVAILABLE* table by the window?

(Context: Why couldn't anyone help me at the service desk?)

You should have gone to the *INFORMATION* desk.

(Context: Did you misunderstand the question?)

I didn't *MISUNDERSTAND* the question; I just chose not to answer it.

(Context: Do you buy fruit at the farmer's market?)

No. I usually buy fruit at the *SUPERMARKET* because they stay open later.

(Context: Have you been trained to do this job?)

No. But I think *EXPERIENCE* is more important than training.

(Context: Do people speak Chinese in Vietnam?)

No. they speak *VIETNAMESE* in Vietnam.

(Context: Is Teresa still living in Texas?)

No. Teresa lives in *CALIFORNIA* now.

(Context: Is Lunar New Year in February?)

No. It's in *JANUARY* this year.

(Context: Does Mary's flight arrive in the evening?)

No. Mary is taking a *MORNING* flight.

4) Reduced and unreduced function words (5 sentences)

Instruction: Read each sentence once at a natural speaking rate and volume. Emphasize the word or phrase that seems appropriate for each context.

If the birthday party wasn't *for* Mary, then who was it *FOR*?

Jane saw a picture *of* the boy she was fond *OF*.

John went *to* visit the woman he had written *TO*.

I *can* run faster than you *CAN*.

He went to a fancy dress party *as* a guest, but what did he dress *AS*?

5) Prosodic disambiguation (10 sentences)

Instruction: You will see five sets of two similar sentences appearing in two different contexts. Try to read the two versions of each sentence in such a way that you make the difference between them clear to a listener.

(Context: Alice is going to divorce Tom.)

When Alice leaves, Tom will be upset.

(Context: We think Alice and Tom should stay together.)

When Alice leaves Tom, we'll be upset.

(Context: Is anyone available to baby-sit today?)

I'll look after the children until lunchtime.

(Context: Can you help me find the ring I lost at the kindergarten this morning?)

I'll look after the children have left.

(Context: Fred and John are arguing. They both want Mary to be in their team.)

The fight is over Mary.

(Context: Mary doesn't know why everyone else has already left the boxing arena.)

The fight is over, Mary.

(Context: I'm not sure if I should let Peter into my English class.)

He's a good body, isn't he?

(Context: James always helps the younger children with their homework.)

He's a good body, isn't he?

(Context: Wherever Sarah goes, everyone stops and talks to her.)

She knows everyone, doesn't she?

(Context: Should I introduce Lucy to the team? I think she's met everyone before.)

She knows everyone, doesn't she?

6) The North Wind and the Sun

The North Wind and the Sun were disputing which was the stronger when a traveler came along wrapped in a warm cloak. They agreed that the one who first succeeded in making *the traveler take his cloak off should be considered stronger than the other. Then the North Wind blew as hard as he could. But the more he blew, the more closely did the traveler fold his cloak around him. And at last, the North Wind gave up the attempt. Then the Sun shone out warmly. And immediately the traveler took off his cloak. And so the North Wind was obliged to confess that the Sun was the stronger of the two.*

7) Computer-prompted dialogue

Instruction: In this task, you will play the role of a reservation agent at EVA airlines, helping a customer to make a reservation over the phone. She wants to book a ticket from Taipei to New York City. Please respond to her questions and requests until you have completed the reservation process. During the phone call, you will hear the customer's voice through your headset; at the same time, everything the customer says will appear on the screen. Written prompts indicating how you should respond to each question or request will also appear on the screen.

(Customer: Good morning. I'd like to reserve a ticket from Taipei to JFK airport in New York.)

When would you like to travel?

(Customer: November twenty-second)

Did you say the twenty-second, or the twenty-seventh?

(Customer: That would be a Tuesday, the twenty-second.)

Would you like a window seat or an aisle seat?

(Customer: An aisle seat, please.)

Would you like a special dinner?

(Customer: No, thank you. I don't think so.)

When would you like to reserve your returning flight?

(Customer: I'm not sure when I'll be returning. So I'll call from New York to reserve the date.)

Your flight, BR 317 will depart from CKS airport at 11:15 AM on November 22nd. You will arrive at Narita Airport at 2:50 PM. You will transfer to Flight 809 to New York JFK Airport, which departs at 7:08 PM from Gate 13F. You will land at JFK Airport at 4:30 PM on November 22nd.

(Customer: Um... wait! Did you say the flight from Narita to New York leaves from Gate 30F?)

The flight leaves from Gate 13F, not Gate 30F.

(Customer: Oh, sorry. Got it. Gate 13F.)

May I have your name?

(Customer: My name is Lucy Hasegawa-Johnson. That's L-U-C-Y H-A-S-E-G-A-W-A J-O-H-N-S-O-N.)

Is that L-U-C-Y H-A-S-E-G-A-W-A J-O-H-N-S-O-N?

(Customer: That's right.)

May I have your credit card number?

(Customer: It's VISA 5924-8013-6702-3516. Expiration date 09/2012.)

So that's VISA 5924-8013-6702-3516. Expiration date 09/2012.

(Customer: That's correct.)

May I have your billing address?

(Customer: It's 1425 Lakeshore Drive, Apartment 47B, Chicago, Illinois 60195.)

Let me repeat that back to you. 1425 Lakeshore Drive, Apartment 47B, Chicago, Illinois 60195.

(Customer: Yes.)

May I have your contact phone number?

(Customer: It's 609-472-1358.)

Is that 609-472-1358?

(Customer: Yes, that's right.)

Is there anything else I can help you with this morning?

(Customer: No, thank you. That's all I need today.)

Goodbye and thank you for calling EVA Airlines.

8) Picture description task

Instruction: In a few moments, a picture will appear at the top half of the screen. In the middle of the screen, you will see a series of five questions, presented one at a time. Please use the information you see in the picture to answer each of the questions in the order.

Question 1: What items are on the man's shopping list?

Question 2: Where in the supermarket will the man find the items on the list?

Question 3: What will the man do after he finds everything on the list?

Question 4: What will the man do after he leaves the supermarket?

Question 5: Describe everything you can see in the picture.