Short-term Rhythmic Training for L2 Pronunciation Enhancement

Yanhao Li

MSc in Applied Linguistics and Second Language Acquisition, 2025

Note that some graphs/tables/images have been removed in order to comply with copyright restrictions.

DECLARATION BY THE CANDIDATE AS AUTHOR OF THE DISSERTATION

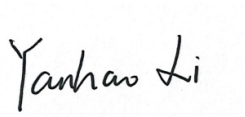
1. I understand that I am the owner of this dissertation and that the copyright rests with me unless I specifically transfer it to another person.

2. I allow the Department to deposit on my behalf a copy of this dissertation in the Oxford University Research Archive ('ORA') where it shall be freely available online for use in accordance with ORA's Terms and Conditions of Use [https://ora.ox.ac.uk/terms_of_use].

3. I understand that this dissertation should not contain material that can be used to personally identify individuals or specific groups of individuals (unless permission has been obtained from the individuals) and that such material should be removed before this dissertation is deposited in ORA.

4. I agree to be bound by the terms of the ORA Grant of Non-exclusive Licence [hS ps://ora.ox.ac.uk/deposit_agreements] and I warrant that to the best of my knowledge, making my thesis available on the internet will not infringe copyright or any other rights of any other person or party, nor contain defamatory material.

5. I agree that my dissertation shall be available for download in ORA in accordance with paragraphs 2, 3 and 4 above.

Signed [an electronic signature is sufficient]:	
Date:	14 August 2025

Short-term Rhythmic Training for L2 Pronunciation Enhancement

Yanhao Li

St Catherine's College

University of Oxford

A thesis submitted for the degree of
Master of Science

2025

Table of Contents

Acknowledgements

This research topic has been in my mind since my senior high school days, and now, I am so proud to have developed it from a simple idea into a comprehensive dissertation. What an incredible journey it has been!

I would like to express my deepest gratitude to my supervisor, Dr Faidra Faitaki, for her invaluable guidance, insightful feedback, and warm encouragement throughout this year. Her unwavering belief in my abilities and her thoughtful suggestions have helped me overcome obstacles whenever I doubted myself or faced challenges. I am also grateful to her and the CreATE research team for the many moments of musical inspiration and stimulating discussions that enriched my work.

My sincere thanks go to all the participants who generously volunteered their time to take part in this study. I also wish to acknowledge the Department of Education at the University of Oxford for providing the resources and academic environment that made this research possible.

Finally, I would like to thank my family and friends for their greatest and unconditional support for me. Your encouragement has been my greatest motivation and strength throughout this journey.

Abstract

Speech Rhythm plays a crucial role in language pronunciation, contributing to more natural prosody and fluency. However, it remains an underexplored area in second language (L2) learning and is often overlooked by both educators and researchers. This gap is particularly problematic when learners' first language (L1) has rhythmic patterns that contrast sharply with their target L2, as few structured pedagogical approaches are available to address such differences. This study investigated the effectiveness of an original short-term rhythmic training on L2 English (a stress-timed language) speech rhythm in L1 Mandarin (a syllable-timed language) speakers and examined the influence of individual differences and text type on learning outcomes. Groups of Mandarin-speaking English learners (n = 38) were randomly assigned to either an experiment group (n = 19), which received rhythmic training, or a control group (n = 19) without intervention. All participants completed pre- and post-tests with conversational texts, and an additional rhythmic text test was administered after the post-test. Speech rhythm was measured using two acoustics metrics, standard deviation and Normalized Pairwise variability Index of vocalic intervals (ΔV and nPVI-V), and data were analysed using linear mixed-effects regression models. Results showed that rhythmic training significantly improved L2 stress-timed rhythmic patterns, with greater rhythmic perception skills predicting better learning outcomes. The text type also influenced their speech production, with rhythmic texts facilitating more stress-timed patterns than conversational texts. These findings provide empirical support for incorporating short-term rhythmic training and rhythm-rich materials into L2 pronunciation teaching. Additionally, this study also highlights the need for future longitudinal and replicable studies concerning rhythmic training evaluation.

Keywords: speech rhythm, rhythmic training, second language learning, stress-timed speech patterns, rhythmic perception, text type

List of Abbreviations

CUREC	Central University Research Ethics Committee
DREC	Departmental Research Ethics Committee
ESL	English as a Second Language
L1	Native/first language
L2	Second language
ΔV	Standard Deviation of vocalic intervals
PVI	Pairwise variability Index
nPVI	Normalized Pairwise variability Index
nPVI-V	Normalized Pairwise variability Index of vocalic intervals
CALL	Computer-assisted language learning
DS	Digit Span
LexTALE	Lexical Test for Advanced Learners of English
MET	Music Ear Test
LMER	Linear Mixed-Effects Model
Rhy	Rhythmic texts
Con	Conversational texts

List of Figures

List of Tables

1. Introduction

In second language (L2) learning, mastering speech rhythm often poses a persistent challenge, especially when L2 learners intend to achieve a native-like level (Gilbert, 2008). However, in L2 pronunciation teaching, rhythm has typically received less emphasis than segmental accuracy, such as the articulation of specific vowels and consonants. This is particularly relevant for Mandarin-speaking learners of English, as their first language (L1) and target language (L2) differ substantially in rhythmic patterns. English is generally considered as a stress-timed language, characterized by alternating strong–weak patterns of stressed and unstressed syllables. In contrast, languages such as Mandarin Chinese are often described as more syllable-timed, with syllables occurring at more regular time intervals. Such rhythmic differences present a great challenge in L1-L2 transfer, resulting in less target-like rhythmic patterns in L2 production.

Drawing on the established link between rhythm in both music and language, prior research has explored the relationship between musical aptitude and L2 performance (Milovanov et al., 2010; Picciotti et al., 2018). Some studies have incorporated musical elements such as melody and rhythm into language teaching, showing that music-based tasks can enhance L2 pronunciation (Wang et al., 2010.; Wiener & Bradley, 2020). However, empirical studies specifically examining rhythmic training interventions remain limited, especially those using structured rhythm-based materials and evaluating outcomes through measurable acoustic changes.

The present study aimed to address this gap by designing and implementing a short-term rhythmic training for Mandarin-speaking learners of English. Two acoustic metrics were used to quantify rhythmic patterns: the standard deviation of vocalic intervals (ΔV) and the normalized Pairwise Variability Index of vocalic intervals (nPVI-V). This study investigated whether rhythmic training improves learners' production of L2 stress-timed rhythmic patterns. In addition, it explored how individual differences in rhythmic perception, working memory, and English proficiency level influence training outcomes, and examined whether types of text affect rhythmic production. Overall, the findings of this study aim to provide empirical evidence on the effectiveness of rhythm-based training for L2 pronunciation and offer pedagogical implications for integrating rhythm-focused strategies into L2 language teaching.

2. Literature Review

This experimental study aimed to explore whether short-term rhythmic training is effective in enhancing L2 English learners' pronunciation, as well as how other factors, such as linguistic and cognitive abilities and the textual environment, influence this immediate uptake. While the intervention was short-term and the observed improvements were more appropriately described as "learning" or "uptake" rather than long-term "acquisition", the study drew on literature framed within the acquisition paradigm to situate its findings within the broader research context. This chapter begins with a general introduction to speech rhythm and its classifications. Then, the acoustic features of different types of speech rhythm will be discussed, drawing on previous empirical studies that examine the validity of various metrics for quantifying rhythmic patterns. The two main metrics employed in this study, standard deviation of the vocalic intervals (ΔV) and the normalized Pairwise Variability Index of the vocalic intervals (nPVI-V), will be explained in detail. The latter half of this chapter provides a more focused overview of research on the acquisition of speech rhythm in both L1 and L2. This is followed by a pedagogical discussion on the teaching and learning of rhythm, based on the limited number of existing studies. The final section identifies the research gaps in the field of L2 rhythm acquisition and presents the motivation and aims of the current experimental study.

2.1. Speech Rhythm

Rhythm refers to the senses relating to a regular, repeated pattern of sound or movement, such as stress and beat in music (Oxford University Press, 2024). More specifically, rhythm in speech functions as an organizational tool, structuring the flow of language into predictable units such as syllables, words, and phrases (Allen & Hawkins, 1980). Studies exploring rhythm (e.g., Allen & Hawkins, 1980; Lehiste, 1977) mentioned that this organizing function of rhythm helps structure temporal sequences of speech and increase utterances' predictability and intelligibility. For example, Allen & Hawkins (1980) argued that this rhythm-generated structure "produces useful perceptual redundancy in speech by constraining the time when (important) articulatory events may occur" (p. 229). Similarly, Lehiste (1977) suggested that rhythm creates expectations for upcoming stress patterns, allowing listeners to pay attention and reduce their cognitive load when processing perceptual information.

In the last century, a classification of speech rhythm was proposed by Pike (1946) and refined by Abercrombie (1965). Pike (1946) defined rhythm as the isochronous (approximately equal temporal) recurrence of different types of speech units. The term "isochrony" in speech refers to the tendency for certain elements of speech (such as syllables or stresses) to occur at regular time intervals, creating some perceived rhythmic patterns. Based on the idea of speech isochrony, all spoken languages can be broadly categorized into two rhythmic types: stress-timed (e.g., English, Russian, Arabic) and syllable-timed (e.g., Mandarin, French, Spanish). For a detailed summary of representative languages and classification sources, see Dauer (1983). Pike (1946) argued that, in syllable-timed languages, each syllable tends to take up roughly the same amount of time. The rhythm is based on the regular recurrence of syllables, making speech sound steadier—like a metronome. For example, in Mandarin, a sentence such as "**巧克力很好吃**" (Qiǎo kè lì hěn hǎo chī, "Chocolate is delicious") consists of six syllables, each produced with relatively equal duration, regardless of stress or grammatical role. In contrast, for stress-timed languages, it is the intervals between one stressed syllable and the next, rather than every syllable, that are approximately equal. The stress-timed pattern is structured around stressed syllables, with unstressed sounds compressed or stretched to maintain the regular stress intervals. In English, for instance, the sentence "Chocolate is delicious" also has six syllables (CHO-co-late DE-li-cious), but the stress falls primarily on "CHO" and "DE". The intervening syllables (those in lower case) are reduced in duration to preserve the timing between stressed beats. Abercrombie (1965) further explained the mechanism behind these two language patterns using physiological principles. He argued that speech is driven by a pulmonic air-stream mechanism, where the respiratory muscles produce regular chest-pulses corresponding to syllables in syllable-timed languages. But in stress-timed languages, some chest-pulses are reinforced as stress-pulses, making stressed syllables.

However, the existence of isochrony in speech, together with this dichotomy of rhythmic types, has been challenged by a series of studies (Allen & Hawkins, 1980; Dauer, 1983, 1987; Lehiste, 1977). Lehiste (1977) reviewed several instrumental studies on rhythmic patterns and concluded that perfect physical isochrony (i.e., strict, equal time intervals between speech elements) does not occur in natural speech production. Nonetheless, rhythmic regularity can still be perceived by listeners, largely due to cognitive processes such as normalization and rhythmic expectancy. According to Lehiste (1977),

normalization refers to the perceptual tendency to adjust for durational variation by averaging interval lengths (i.e., overestimating short intervals and underestimating long ones), making irregular timing perceptually regular. Similarly, rhythmic expectancy allows listeners to anticipate the timing of upcoming events in speech based on prior rhythmic patterns. Supporting this perspective, Allen and Hawkins (1980) argued that listeners exhibit a perceptual bias such that they hear syllables as lasting the same amount of time even if the actual temporal intervals between them are irregular. Under this light, isochrony exists more as a perceptual and structural tendency, rather than as a strict physical constraint. Consequently, the strict dichotomy between stress-timed and syllable-timed languages has also been called into question.

Building on this debate on strict dichotomy, several contrastive studies have investigated rhythmic patterns across different language types (e.g., Roach, 1982; Dauer, 1983). Roach (1982) designed a carefully controlled instrumental study to determine whether languages could be objectively classified as stress-timed or syllable-timed based on measurable timing patterns in speech. His result did not show consistent patterns for both types, providing little support for a strict binary classification of speech rhythm. Similarly, Dauer (1983) examined among several languages such as English, Thai, Spanish and Greek, reaching the same conclusion as Roach (1982): there were no significant differences in the isochrony of inter-stress intervals across languages traditionally classified into these two rhythmic types.

Instead, Dauer (1983) proposed that speech rhythm should be viewed as a continuum, with languages differing in the degree to which they exhibit stress-based rhythm. These differences emerge from specific phonological and phonetic properties of languages. Based on cross-linguistic analysis, Dauer (1983) proposed three key features that distinguish more stress-timed from more syllable-timed languages: (1) syllable structure, (2) stress pattern, and (3) the presence or absence of vowel reduction.

These features are again evident in the comparison between the English sentence “Chocolate is delicious” (/ˈtʃɒk.lət ɪz dɪˈlɪʃ.əs/) and its Mandarin equivalent “**巧克力很好吃**” (Qiǎo kè lì hěn hǎo chī). For the first syllable-related feature, English permits complex syllable structures, as seen in “chocolate” /ˈtʃɒklət/, where the first syllable contains both a consonant cluster onset /tʃ/ and a coda /k/, and the second syllable includes a reduced vowel /ə/ (schwa). Regarding the second stress-related feature, English also exhibits strong stress contrast, with primary stress falling on the first syllable

of “chocolate” (/ˈtʃɒklət/) and the second syllable of “delicious” (/dɪˈlɪʃəs/), while function words like “is” are typically unstressed and phonetically reduced. As a result, the third feature—vowel reduction is common in those unstressed syllables, as those in -late and -cious reduced to schwa /ə/. In contrast, Mandarin, a more syllable-timed language, follows a simple (C)V(N)[1] structure, lacks strong stress contrast, and maintains full vowel articulation in most cases. These differences contribute to the perception of more regular and syllable-based timing in Mandarin, as opposed to the stress-based rhythmic patterns observed in English.

Following her examinations across different languages, Dauer (1987) further illustrated the existence of mixed-rhythmic languages, which incorporate both stress-timing and syllable-timing features. One such example is Polish, which occupies an intermediate position on the rhythmic continuum. Polish permits complex syllable structures, including clusters in both onset and coda positions (e.g., in the word wstrząs /fstʂɔ̃s/ “shock”), a feature typical of stress-timed languages. It also exhibits alternating rhythmic stress, with a strong tendency for penultimate stress across words. However, unlike prototypical stress-timed languages such as English, Polish does not display systematic vowel reduction at a normal speech rate. This case highlights the limitations of a binary rhythm typology and supports a more gradient view of rhythmic classification.

2.2. Acoustic Features of Speech Rhythm

Dauer's (1983) conclusion that rhythmic characteristics can be inferred from three main phonological indices (syllable, stress and vowel) prompted interest in different language rhythmic patterns. Many studies have since sought to explore the acoustic correlates of different rhythm categories, such as syllable-timed and stress-timed languages (Grabe & Low, 2008; Ling et al., 2000; Nolan & Asu, 2009; Ramus et al., 1999).

In English, vowel-related features have received the most attention in rhythm studies. According to Ling et al. (2000), this focus is largely due to the difficulty in determining syllable and stress boundaries in English, which often leads to a high degree of subjectivity and inconsistency across researchers. For example, the pronunciation and the syllabification of the word ‘fire’ are different across dialectal areas and speaker variations. Some say /faɪ.ər/ (two syllables) while others say /faɪr/ (one syllable). In this case, it is hard to determine a standard criterion for phonetic segmentation.

[1] (C)V(N) indicates a syllable structure consisting of an optional consonant onset (C), a vowel nucleus (V), and an optional nasal coda (N), which is typical of many Mandarin syllables.

This ambiguity is also further supported by Ramus et al. (1999), who highlighted the challenges of using abstract categories like syllables and metrical feet when doing rhythm analysis. To address these challenges, they proposed using more surface-accessible and quantifiable features, like vocalic and consonantal intervals. The key metrics in their measurement include %V (the percentage of utterance duration occupied by vocalic intervals), ΔC (the standard deviation of consonantal interval durations), and ΔV (the standard deviation of vocalic interval durations). These metrics are based on segmental durations and enable a more gradient comparison between language rhythmic types. For instance, %V, which represents the proportion of vocalic intervals in an utterance, is typically higher in syllable-timed languages such as Mandarin, where each syllable is realized with a full, unreduced vowel. In contrast, stress-timed languages, such as English, exhibit vowel reduction in unstressed syllables, resulting in a lower overall proportion of vowels. The "chocolate" example[2] illustrates this contrast (see Table 1): in English, vowel reduction (e.g., the schwa in "chocolate" and "delicious") leads to a lower %V (30%). In Mandarin's equivalent phrase (Qiǎo kè lì hěn hǎo chī), vowels are fully articulated in each syllable, producing a higher %V of around 60%.

Language	Utterance	Total Duration (ms[3])	Vowel Duration (ms)	%V
English (stress-timed)	"Chocolate is delicious" /ˈtʃɒk.lət ɪz dɪˈlɪʃ.əs/	1200	360	30%
Mandarin (syllable-timed)	"巧克力很好吃" (Qiǎo kè lì hěn hǎo chī)	1200	720	60%

Table 1 Illustration of %V

The second metric, ΔC represents the standard deviation of consonantal interval durations, reflecting the extent of temporal variability among consonant segments. Stress-timed languages, such as English, which often have complex syllable structures and consonant clusters, tend to show higher ΔC values due to greater variation in consonant durations. But for syllable-timed languages like Mandarin, they feature simpler syllable structures and exhibit more consistent consonant timing. This distinction is illustrated in Table 2 using the same "chocolate" example.

[2] All figures in the "chocolate" example (Tables 1-3) are adapted from Crystal and House (1988) on English vowel length and Aoyama and Reid (2006) study on English consonant length.
3 All acoustic durations in this study are measured in milliseconds (ms).

Language	Consonant Durations (ms)	ΔC
English (stress-timed)	100 (ʧ), 40 (k), 20 (l), 60 (d), 90 (ʃ)	High (~32 ms)
Mandarin (syllable-timed)	50 (q), 60 (k), 50 (l), 60 (h), 55 (ch)	Low (~4.5 ms)

Table 2 Illustration of ΔC

Finally, ΔV, the standard deviation of vocalic interval durations, captures how much vowel lengths fluctuate within an utterance. Stress-timed languages like English exhibit greater ΔV values due to alternating stress patterns and vowel reduction. In contrast, Mandarin maintains relatively even vowel durations across syllables, contributing to a lower ΔV. As shown in Table 3, English vowel durations fluctuate, while Mandarin maintains a more syllable-timed pattern with more even and stable vowel durations.

Language	Vowel Durations (ms)	ΔV
English (stress-timed)	[180 (ɒ), 60 (ə), 120 (ɪ), 90 (ə)]	High (~32 ms)
Mandarin (syllable-timed)	[120 (iǎo), 110 (è), 115 (ì), 120 (ī)]	Low (~4.5 ms)

Table 3 Illustration of ΔV

As shown in Table 3, ΔV captures differences in vowel duration variability across languages. This metric is particularly relevant because vowel reduction—a defining feature of stress-timed languages—results in uneven vowel lengths, with unstressed vowels often shortened or centralized (Dauer, 1983). In contrast, syllable-timed languages tend to maintain more uniform vowel durations. Since this durational variability stems directly from the presence or absence of vowel reduction (one of the features to distinguish between language stress patterns), ΔV offers a quantitative means of indexing rhythmic type. As Grabe and Low (2008) observed, “so-called stress- and syllable-timed languages differ in the durational variability encountered in vowels” (p.3). Consequently, many studies (e.g., Grabe & Low, 2008; Ling et al., 2000; Low, 1998; Ramus et al., 1999) have employed ΔV to compare rhythmic patterns, particularly between syllable-timed and stress-timed patterns.

However, relying solely on standard deviation to represent the variability of vowel duration is inadequate for capturing the full complexity of speech rhythm. Standard deviation measures the overall dispersion of vowel durations from the mean, but it fails to account for the temporal patterning of those durations. This limitation introduces two key

problems. First, standard deviation is not sensitive to the temporal ordering in speech. For instance, imagine two hypothetical languages: one in which three long vowels are followed by three short vowels, and another in which long and short vowels alternate regularly. Although the rhythmic patterns of these two languages are distinct, their standard deviation values may be similar. This is because the metric does not reflect how vowel durations are sequenced. Second, the standard deviation is highly susceptible to variations in speaking rate, both across utterances and between speakers. In natural speech, fluctuations in tempo might inflate or reduce the standard deviation values, thereby obscuring the actual rhythm characteristics of a language.

To address the limitations of standard deviation, researchers suggested a pairwise measurement for vocalic intervals: the Pairwise Variability Index (PVI). This measurement was initially proposed by Low (1994; 1998) and has been widely adopted in rhythm research (e.g., Grabe & Low, 2008; Ling et al., 2000; Nolan & Asu, 2009). Unlike standard deviation, which provides a general measure of dispersion, the PVI captures the degree of variability between successive vowel durations. This makes it particularly well-suited for capturing the temporal alternation patterns that distinguish rhythmic types. For example, stress-timed languages often show alternation between longer and shorter vowels, whereas syllable-timed languages typically exhibit more uniform vowel durations. Two main versions of PVI have been proposed: raw PVI (rPVI) and normalized PVI (nPVI). The rPVI measures absolute durational differences between successive intervals, while the nPVI adjusts for speech rate by expressing these differences relative to the mean duration of each pair.

The rPVI is calculated using the following formula (d_k = duration of the k^{th} vocalic intervals; m = the total number of intervals):

$$\text{rPVI} = \frac{1}{m-1} \sum_{k=1}^{m-1} |d_k - d_{k+1}|$$

This formula computes the average absolute difference in duration between each pair of adjacent segments (e.g., vocalic intervals). Higher values indicate greater durational variability of vowels, which aligns with stress-timed rhythmic patterns.

However, this formula does not account for speaking rate. When speech accelerates or slows down, the rPVI may be distorted by these tempo fluctuations. To resolve this issue, Deterding (1994) introduced a normalization procedure of PVI. The procedure involves

calculating the mean duration of all intervals in an utterance, and dividing each individual duration by this mean, creating a normalized value for subsequent analysis. Building on this idea, Low (2002) suggested an improved metric in her doctoral thesis: the Normalised Pairwise Variability Index (nPVI). This updated version reflects not only alternations of long and short vowels but also is independent of different speech rates. The formula is expressed as:

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

Each pairwise difference is represented as a proportion of the mean of the two intervals being compared. The result is then scaled by 100 for readability. But it should be noted that this calculation is not suitable for all speech segment types. According to Gay (2009), it is obvious that the segmental duration will be compressed in faster speech and lengthened in slower speech. This is important in the study of vowels, since speech rate affects vowel durations more significantly than consonant durations. As a result, nPVI is particularly appropriate for analyzing vowels, which exhibit more variation in duration and play a critical role in rhythmic distinctions.

In this study, both the ΔV and the nPVI of vocalic intervals (nPVI-V) are used to quantify speech rhythm. While ΔV captures overall dispersion in vowel duration, nPVI-V reflects local variability between adjacent vowels and controls for speech rate. Because each metric highlights a different aspect of vowel duration variability, using both allows for cross-validation and a more robust assessment of rhythmic patterns. Higher values on both metrics, especially the nPVI-V, are typically associated with stress-timed patterns, while lower values correspond to syllable-timed ones.

2.3. Acquisition of Rhythm in L1 and L2

According to the seminal work by Abercrombie (1967) on speech rhythm, the rhythmic structure of a language plays a foundational role in language acquisition. Rhythm is considered one of the earliest prosodic features acquired in infancy, forming a scaffolding for children's language processing (Ramus et al., 1999; White & Mattys, 2007; Whitworth, 2002). According to Aoyama and Guion (2007), by the age of one, the L1 rhythm has been deeply possessed by infants and will be unconsciously applied to any L2 they learn. In other words, once acquired, rhythmic patterns become deeply ingrained and are particularly resistant to change. This makes rhythm one of the most persistent

challenges for adult learners when acquiring a second language (e.g., Adams, 1979; Gut, 2003; Low et al., 2000; Ling & Wang, 2005).

Adams (1979) provided one of the earliest and relatively comprehensive studies into this issue, highlighting the difficulties L2 learners face because of the influence of their L1. Her study involved 30 non-native graduate teachers of English from Asian, Southeast Asia, and the Pacific Islands, whose native languages were predominantly syllable-timed or mora-timed (e.g., Malay, Thai, Japanese). Drawing on typological descriptions of different rhythmic types, Adams anticipated that the participants' L1 backgrounds (especially syllable- or mora-timed) would transfer into their L2 English. This prediction was confirmed by participants' pre-test reading results, which revealed even timed syllables, minimal vowel reduction, and an absence of the strong–weak alternation characteristic of English stress-timed rhythm. She then randomly allocated these participants into an experimental group (n = 15) receiving a 20-hour rhythm-focused training programme, or a control group (n = 15) receiving no specific rhythm instruction. Both groups improved, but the experimental group's gain was twice that of the control group. These results underline that, regarding the rhythm acquisition, natural exposure alone is insufficient for even experienced L2 learners from different L1 rhythmic backgrounds. Based on the instruction content, Adam also highlighted the importance of explicit instruction in speech rhythm patterns—particularly stress placement, vowel reduction, and timing—for overcoming ingrained L1 influences. However, a key limitation of Adam's study is that speech rhythm (i.e., rhythmic naturalness) was assessed solely through perceptual judgments by native speakers. Although the raters were blind to group membership and testing stage, this method is inherently subjective and susceptible to individual differences in judgment. Furthermore, because such assessments depend on the specific raters involved, they lack the replicability and precision compared to instrumental acoustic analysis.

Recognizing this limitation, subsequent studies moved toward more quantitative and replicable approaches, employing objective metrics such as the nPVI, consonantal interval variability (ΔC), and the proportion of vocalic intervals (%V) to examine rhythmic timing (Gut, 2003; Low et al., 2000; Ling & Wang, 2005). These metrics provided empirical insight into how L1 and L2 speech differed.

For instance, Low et al. (2000) applied such quantitative methods in a comparative study of Singapore English (SE) and British English (BE), using the normalized Pairwise Variability Index (nPVI) and vowel formant analysis to capture rhythmic timing

differences. Participants in the SE group were 10 ethnically Chinese undergraduates aged 20-22. All of them were bilingual in English and Mandarin Chinese, with Mandarin as their L1 and English acquired as their L2 through schooling. The BE group also consisted of 10 undergraduates from the same age group, all monolingual native English (L1) speakers raised and educated in southern England. As Mandarin is generally classified as a syllable-timed language, SE speakers were expected to show more regular timing and weaker vowel reduction than the BE group. To test this, Low et al. (2000) employed two reading materials: a Full Vowel Set, containing only full vowels, and a Reduced Vowel Set, containing a mixture of full and potentially reduced vowels. Results showed that BE speakers had higher nPVI values in the Reduced Vowel Set than in the Full Vowel Set, reflecting greater durational variability when reduced vowels were present. In contrast, SE speakers showed little difference in nPVI values between the two sets, indicating minimal durational vowel reduction in SE speech. Furthermore, the authors analyzed the vowel formants using spectra within both SE and BE speakers' speech. They found that BE reduced vowels were centralized toward the schwa region, whereas SE reduced vowels remained relatively peripheral, indicating weaker spectral reduction. Together, all findings provide strong acoustic evidence that objective metrics such as nPVI and vowel formant analysis can effectively distinguish stress-timed from syllable-timed patterns.
Another important study by Lin and Wang (2005) provides a more targeted investigation into the rhythmic characteristics of Mandarin Chinese and their influence on L2 English production. Participants were adults (n = 12): six native Canadian English speakers and six native Mandarin speakers who had been learning English as their L2. The authors analyzed their speech using two key rhythm metrics: %V and ΔC. They compared two pairs of speech produced from these participants: (1) L1 Mandarin vs. L1 English, and (2) Mandarin speakers' L2 English vs. native English. Their findings confirmed the traditional auditory classification of Mandarin Chinese as a more syllable-timed language, showing significantly different rhythmic characteristics from the stress-timed pattern of English. Furthermore, the authors reported that, in L2 production, learners exhibited intermediate rhythmic properties, positioning between those of the L1 and the L2. As learners progressed to a more advanced proficiency level, their rhythmic patterns in English increasingly resembled those of native speakers, but never became fully identical. This progression indicates that while L1 influence is strong, it may gradually diminish with increased exposure and practice, but rarely disappears entirely. In other words, L2 experiences and proficiency level would influence the L2 pronunciation

learning process. Moreover, Lin and Wang found that %V, which focuses on vowel segments, was a more robust and reliable acoustic metric than ΔC, particularly when speech rate varied.

While Lin and Wang's (2005) study found that L2 proficiency level affects L2 pronunciation learning outcomes, later research has highlighted the role of additional individual difference factors in this language learning process, such as working memory and music aptitude (e.g., Darcy et al., 2015; Mattys & Baddeley, 2019; Sun et al., 2024; Suzukida, 2021; Yalçın et al., 2016). Higher working memory capacity, especially complex span capacity (e.g., L2 sentence recall tasks) and storage capacity (e.g., digit span task), has been shown to support more native-like phonological processing and reduce L1 interference (Darcy et al., 2015; Mattys & Baddeley, 2019). Likewise, musical aptitude, especially the sensitivity to pitch, tone and rhythm, was found to significantly enhance learners' abilities in perceiving and learning accurate L2 sounds (Sun et al., 2024; Suzukida, 2021). However, although these studies have established correlations between individual differences and overall L2 pronunciation performance, most adopt a broad focus on pronunciation and do not directly address how those factors influence the learning of L2 speech rhythm. Thus, this gap highlights the need for the current study, which aims to measure how English proficiency level, working memory and musical aptitude (rhythmic perception) contribute to L2 speech rhythm learning.

Taken together, all studies mentioned in this section reveal the complex interplay in L2 rhythm learning process: rhythm is both resistant and modifiable. On the one hand, L1 rhythmic structures are deeply embedded and persist into L2 use; on the other hand, individual different factors (such as better L2 proficiency level, stronger working memory capacity and better rhythmic perception ability) can facilitate a more native-live speech production. This dual nature of persistence and plasticity forms the theoretical basis for the present study. Specifically, this thesis explores whether rhythmic training can accelerate the development of English stress-timed patterns for Mandarin-speaking learners. Given the rhythmic divergence between Mandarin and English (syllable-timed vs. stress-timed), this pairing offers a compelling context to explore the effectiveness of rhythm-based pedagogical interventions.

2.4. Teaching English Rhythm

While research has reviewed the theoretical foundations of speech rhythm, the acoustic features of rhythmic patterns, and the acquisition of rhythm, comparatively little attention

has been given to how rhythm can be effectively learned and taught in L2 contexts. This section begins by discussing the role of English rhythm in communication, then outlines a range of instructional approaches used in L2 pedagogy and finally explores the challenges and emerging methods in rhythm instruction.

2.4.1. Role of English Rhythm in Communication

Rhythm is widely recognized as a key component of English pronunciation, playing a crucial role in both speech production and perception (Avery, 1992; Brown, 2017; Wong, 1987). In other words, rhythm not only structures the delivery of information but also facilitates listeners' comprehension during communication. Gilbert (2008) metaphorically described rhythmic signals as "road signs" that guide listeners through spoken discourse (p. 2).

For humans, it is impossible to speak without rhythm, intonation and tone. For example, Wong (1987) mentioned that if a non-native speaker lacks knowledge of English rhythmic patterns, they may struggle to be understood or risk being misinterpreted. This is particularly true for speakers whose L1 is syllable-timed. O'Connor (1980) observed that these speakers may produce English with a mechanical rhythm by assigning equal duration to every syllable. This timing pattern disrupts the stress-timed rhythm of English and makes an utterance difficult to understand. Celce-Murcia et al. (1996) further supported this view, noting that incorrect rhythm can lead not only to humorous errors but also to serious communication breakdowns. From the listener's perspective, understanding English rhythm goes beyond simply counting syllables and requires sensitivity to the English stress-timed pattern (i.e., the alternation of strong and weak syllables). In this sense, L2 English learners need to at least possess an awareness of the rhythmic structure to achieve comprehensive and effective spoken English (Gilbert, 2008).

2.4.2. Ways to Teaching and Learning English Rhythm

As Avery (1992) suggested, the primary goal of spoken language training is to help ESL (English as a Second Language) learners develop fluent and comprehensive speech. In this regard, mastering rhythmic patterns would significantly enhance learners' fluency and intelligibility.

Tuan and An (2010) outlined several pedagogical approaches for introducing English rhythm in language teaching contexts. One widely recommended approach is the use of

physical gestures to embody rhythm. Supported and explored by numbers of research (Avery, 1992; Gower, 1983; Llanes-Coromina et al., 2018; O'Connor, 1980; Wong, 1987), this method encourages learners to use body movement such as clapping, tapping, or nodding to feel and express rhythmic timing. Llanes-Coromina et al. (2018) further confirmed that brief training with rhythmic beat gestures could enhance L2 pronunciation. These physical gestures offer a multisensory learning experience, making abstract rhythmic patterns more accessible and memorable.

Another method involves the use of visual representations of rhythm, referred to as 'notion' by Tuan and An (2010). Teachers can use visual cues such as capital letters, underlining, or symbolic markings to highlight stress in sentences (Gower, 1983; Kenworthy, 1987; Ur, 1996). For example, Cuisenaire rods are used to represent stressed and unstressed segments with taller and shorter rods respectively (Gower, 1983). Such visualization helps learners map auditory features onto visual forms, facilitating a clearer understanding of rhythm.

A third method involves the use of metrical materials, such as chants, songs and nursery rhymes. These forms are inherently stress-timed and align well with the rhythmic structure of English (Celce-Murcia et al., 1996; Fischler, 2005; Graham, 1993). Musical genres such as jazz (Graham, 1993) and rap (Fischler, 2005) have also been explored for their potential to support rhythmic training in L2 learning. Both Graham's (1993) and Fischler's (2005) studies showed that learners benefit from music-based training, particularly in developing more natural stress and intonation patterns. Additionally, inspired by musical form, rhythm can also be taught using 'nonsense syllables' (e.g., da) to represent segments within words or sentences (Avery, 1992). For example, the word 'pepperoni' can be rhythmically represented as 'da da DA da', highlighting the primary stress on the third syllable. Similarly, the sentence 'to the bus station from here…' can be practiced using rhythmic syllables as 'da da DA da da da da…', emphasizing the natural stress patterns of English.

In addition to metrical inputs, non-metrical materials, such as recordings of natural conversations, can also be valuable. Though this method is more implicit and less structured, it exposes learners to a more authentic speech style. Through strategies such as shadowing, repetition and rhythmic reading of real-life discourse, learners can gradually develop an intuitive sense of timing that extends beyond isolated words or syllables (Kenworthy, 1987; Wong, 1987).

2.4.3. Challenges and New Approaches to Teaching English Rhythm

Teaching English pronunciation, especially rhythm, includes a variety of challenges from both learners' and teachers' perspectives (Gilbert, 2008; Tuan & An, 2010). For learners, acquiring rhythm in a second language is inherently difficult due to its dual nature of persistence and plasticity, as discussed in Section 2.3. Once acquired in early L1 development, rhythmic patterns tend to be deeply ingrained and resistant to change, making them one of the most persistent features to transfer into L2 speech.

From the teacher's perspective, two main obstacles often arise. First, there is a lack of systematic, accessible materials and techniques specifically designed for rhythm instruction in classroom contexts (Wei, 2006). Second, non-native English teachers may lack confidence in their linguistic identity and ability to model English rhythm accurately (Gilbert, 2008). One way to address both challenges is to develop structured, evidence-based instructional models that teachers can adopt with clear guidance.

Zhang and Yuan (2020) took a step in this direction by comparing two explicit pronunciation teaching approaches: segmental-based instruction, which focuses on the accurate production of individual consonants and vowels; and suprasegmental-based instruction, which targets broader prosodic features such as rhythm, stress, and intonation. In their study, 90 Chinese university students were divided into three groups: a segmental-based instruction group (n = 30), a suprasegmental-based instruction group (n = 31), and a control group (n = 29). Over 18 weeks, the two experimental groups received targeted lessons, while the control group followed the regular curriculum. Participants' speech was recorded at pre-, post-, and delayed post-test stages from both controlled reading tasks (reading aloud pre-selected sentences and short passages) and spontaneous speech tasks (free speech prompted by picture descriptions or discussion topics). Those recordings were rated by trained native-speaker judges for overall comprehensibility on a 9-point scale. The authors found that both experimental groups (with targeted instructions) improved in reading-task comprehensibility, but only the suprasegmental group indicated significant and lasting gains in spontaneous speech. These results suggest that explicit suprasegmental instruction, particularly rhythm, has a more durable effect than segmental-only training. However, because the suprasegmental intervention combined rhythm, stress, and intonation, the study could not isolate rhythm's specific contribution. In addition, though reasonable, native-speaker ratings cannot serve as the definitive gold standard for evaluating L2 acquisition outcomes, as they are

inherently subjective and dependent on raters. These limitations highlight the need for research that targets rhythm alone and employs more objective acoustic measures to assess L2 pronunciation improvement.

While Zhang and Yuan's (2020) study stated the effectiveness of including rhythm intervention in the classroom, other researchers have explored technology-assisted approaches to introduce training in a more structured and engaging way. One such innovation is the MusicSpeak system, a computer-assisted language learning (CALL) tool developed by Wang and colleagues (2010, 2016). MusicSpeak was designed to improve English rhythm perception and production, particularly among Chinese L1 learners, who are considered to possess more syllable-time patterns (Mok & Dellwo, 2008; Ramus et al., 1999). This system automatically generates the musical rhythm (represented by beats) of English sentences based on the text input. Users are then able to listen to and practise speaking in synchrony with the corresponding beats.

Wang et al. (2016) evaluated the effectiveness of MusicSpeak by recruiting 20 Chinese university students (L1 Mandarin) and six native American English speakers, who served as the L1 English reference model. All participants recorded the same 15 English sentences (adapted from song lyrics) in two speaking styles: natural (normal reading without MusicSpeak aids) and rhythmic (following MusicSpeak's beat-aligned prompts). Performance was assessed through both objective acoustic analysis and subjective perceptual judgement by native English speakers. Acoustic analysis used the normalised Pairwise Variability Index for syllabic intervals (nPVI-S) and for vocalic intervals (nPVI-V). Perceptual ratings were provided by 10 native English judges, who rated the nativeness of the rhythm in each recording on a 7-point scale. For analysis, the non-native learners were categorized into a "good rhythm" group and a "weak rhythm" group based on expert assessment of their natural-style speech. For learners with initially weak rhythm, rhythmic-style recordings produced with MusicSpeak showed higher nPVI-S values than their natural-style speech and were rated as more rhythmically native-like. But for those in the "good rhythm" group, there was no significant improvement shown. This indicates that MusicSpeak can lead to measurable improvement in both objective acoustic metrics and perceptual ratings, especially for participants with weak rhythm perception. However, this study focused exclusively on immediate, in-training effects and did not assess whether rhythmic gains persisted beyond the imitation context.

Taken together, the findings from Zhang and Yuan (2020) and Wang et al. (2010, 2016) highlight the potential of both classroom-based and technology-assisted approaches for

enhancing L2 English pronunciation. Nonetheless, each leaves important questions unanswered: Zhang and Yuan (2020) provided a general suprasegmental instruction but could not isolate the specific contribution of rhythm from other features; Wang et al. (2016) only examined the immediate and in-training effects, without examining whether rhythmic gains persisted beyond the imitation context. As a result, the present study addressed both limitations by isolating rhythmic-focused training and evaluating its effects not only during training, but also through a post-test.

In addition to those structured approaches and technology-assisted tools explored by Zhang and Yuan (2020) and Wang et al. (2010, 2016), there is another underexplored dimension in rhythm teaching—the nature of the text itself used in the pronunciation training process.

It is noteworthy that all aforementioned teaching models are based on daily conversational text, rather than texts with inherently rhythmic or poetic structures. Although some studies have mentioned the use of poetry to enhance pronunciation, most are theoretically proposed and lack empirical validation (Ariningsih, 2023; Finch, 2003). For instance, Finch (2003) argued that simple poetic forms, such as picture poems (visually shaped texts), haiku (short Japanese poems with a 5-7-5 syllable structure), and song lyrics, can raise learners' awareness of pronunciation, stress and sentence flow. Finch also suggested that such texts reduce anxiety and promote learners' engagement in language learning. However, these claims were based on classroom practice rather than empirical research. Nonetheless, Finch's work offers valuable pedagogical insight into how poetic forms may support rhythmic awareness, leading to a broader question of how text type could influence pronunciation learning.

Later, Ariningsih (2023) conducted a systematic literature review to evaluate the relationship between poetic styles used in teaching and pronunciation improvement. This review screened 113 studies from data sources and finally included only one empirical study that met all inclusion criteria. Though this systematic review lacked detailed synthesis and a thorough methodological justification, it still highlighted the inherent nature of poetry. Poetry is rooted in oral tradition and naturally embeds features such as rhyme and rhythm. These features align closely with the suprasegmental aspects of language learning, including stress and timing. And for learners who are exposed to poetry-related rhythmic texts, they may more easily perceive and reproduce those poetic features in their speech. Despite the methodological limitations and inadequate argument,

Ariningsih's (2023) work pointed out a notable gap in the empirical literature concerning the role of poetic texts (or in a broader sense, texts with explicit rhythmic patterns) in pronunciation development. This gap motivated part of the present study, which aimed to investigate whether the text type (conversational vs. rhythmic) influences learners' ability to produce more native-like and stress-timed English rhythmic patterns.

2.5. Research Aims

As discussed in previous sections, various instructional approaches—such as gesture, music, and visual representation—have been proposed for teaching English rhythm. However, few empirical studies have rigorously evaluated the effectiveness and validity of these methods. To address this gap, the present study introduced a short-term rhythmic training, which was originally designed for advanced Mandarin-speaking learners of English. Unlike previous research, this study aimed not only to implement a rhythm training method but also to empirically evaluate its impact using pre- and post-training acoustic metrics: standard deviation of vocalic intervals (ΔV) and normalized Pairwise Variability Index of vocalic intervals (nPVI-V).

Additionally, the study examined whether individual differences (working memory, English proficiency, and rhythmic perception) and the external textual environments (conversational vs. rhythmic) influence learners' L2 speech patterns. Through this design, the study intended to contribute a validated, evidence-based training intervention for L2 English rhythm production.

2.6. Research Questions and Predictions

This study addresses three main questions:

RQ1: Does short-term rhythmic training improve advanced Mandarin-speaking L2 English learners' production of stress-timed patterns, as measured by increased ΔV and nPVI-V values?

RQ2: Is the extent of this improvement influenced by individual learner variables—namely working memory (Digit Span), English proficiency level (LexTALE), and rhythmic perception ability (Music Ear Test)?

RQ3: Do the types of text read (conversational vs. rhythmic) affect learners' production of English rhythmic patterns?

For RQ1, it was hypothesized that short-term training would improve learners' production of stress-timed patterns in L2 English. Specifically, post-test speech was expected to exhibit higher ΔV and nPVI-V values, indicating greater durational variability of vowels, which is the key feature of a more stress-timed speech pattern of English (Grabe & Low, 2008; Low, 1998). However, it was anticipated that nPVI-V would show more robust changes than ΔV, since ΔV is highly sensitive to speech rate and inter-speaker variation. This greater sensitivity means that any training-related improvement represented by ΔV could be masked by fluctuations in speech tempo or individual speaking styles, making ΔV a less reliable measuring metric than nPVI-V

For RQ2, all three individual difference variables were expected to moderate the extent of improvement:

Working Memory (Digit Span): Learners with higher working memory capacity were expected to have better improvement after training compared to those with lower scores. They may be more effective in encoding and retaining rhythmic information, leading to better integration of stress patterns in their performance after training (Mattys & Baddeley, 2019; Suzukida, 2021; Woods et al., 2011)

English Proficiency (LexTALE): Learners with higher English proficiency, as measured by the LexTALE receptive vocabulary test (Lemhöfer & Broersma, 2012), were expected to show greater improvement after the training intervention. As shown by Lin and Wang (2005), higher-proficiency Mandarin learners of English produced speech that more closely resembled native-like rhythm than that of their lower-proficiency peers. This suggests that more proficient learners tend to be more attuned to target-like speech patterns and may therefore be better able to incorporate learnt rhythm features into their L2 production.

Rhythmic Perception (MET): Learners with better MET performance were expected to benefit more from the rhythmic training, as musical aptitude has been linked to better sensitivity to rhythmic features in speech (Gordon, 2015; Milovanov et al., 2008, 2010; Posedel et al., 2012; Sun et al., 2024; Suzukida, 2021).

For RQ3, rhythmic texts were predicted to facilitate greater stress-timed rhythmic patterns in learners' speech production than conversational texts. Leaners may find it easier to detect and produce stress-timed patterns when given a more explicitly rhythmic text type (Ariningsih, 2023; Finch, 2003).

3. Methodology

The study was conducted online through the Gorilla Experiment Builder (www.gorilla.sc), an online platform for creating and hosting experiments (Anwyl-Irvine et al., 2020). Two versions of the experiment were employed during the process: one for the piloting phase and another for the formal data collection. As only minor adjustments were made after the piloting phase, such as modifications to the wording of the task instructions, the pilot data were included in the final analysis. Data collection took place between 10 May 2025 and 5 June 2025, with the piloting phase in the first 5 days. Participants were recruited via personal contacts and social media platforms.

Following data collection, the dataset was processed using Microsoft Excel for data management, Praat (Boersma & Weenink, 2025) for phonetic segmentation, and R (R Core Team, 2023) for statistical analysis. This study was reviewed and approved by the University of Oxford's Education Departmental Research Ethics Committee (Ethics Reference: EDUC-DREC–1351018).

3.1. Participants

A total of 42 adult participants (31 females, 11 males) took part in the experiment. All participants were native speakers of Mandarin Chinese and had been learning English (L2) for more than 10 years (M = 17 years, SD = 3.9). Approximately 85% of the participants reported beginning learning English during primary school (i.e., before the age of 10). Most (more than 80%) participants reported regular daily exposure to English across the four skills (reading, listening, speaking, and writing).

Participants were randomly assigned to either the experiment group (n = 21) or the control group (n = 21) using Gorilla's balanced randomiser prior to the main task. The key difference between the two groups was that the experiment group received short-term rhythmic training, while the control group did not receive the intervention. All participants provided informed consent before beginning the experiment and agreed to the collection and use of their data. However, due to unexpected dropouts and technical issues (two participants withdrew, one pre-test recording was lost, and one post-test recording was unusable), only 38 complete datasets were included in the final analysis (n = 19 for each group).

3.2. Experiment materials

3.2.1. Basic Ability Tests

All participants completed three basic assessments to measure their linguistic and cognitive abilities as part of the experiment: the Digit Span (DS) task, the LexTALE, and the Rhythm part of the Musical Ear Test (MET).

The DS task, originally designed to test working memory and attention, is a well-established component of the Wechsler memory scales and Wechsler intelligence scales (Wechsler, 1981, 2012). Traditionally, the test involves an examiner reading aloud a sequence of digits, which participants are asked to repeat either in the same (forward) or reverse (backward) order (Blackburn & Benton, 1957). In the present study, a computerized version of the DS test was implemented via Gorilla, where digits were visually presented on the screen and participants typed in their responses. This computerized version significantly enhances the reliability and precision of the assessment (Woods et al., 2011). The backward DS task imposes a greater cognitive load than the forward version, as it demands not only stronger short-term retention but also active manipulation of information (Lezak et al., 2012; Woods et al., 2011). Given these qualities, and the time constraints of the experiment, only the backward DS was selected as a more robust measure for this study.

The Lexical Test for Advanced Learners of English (LexTALE; Lemhöfer & Broersma, 2012) was administered to assess participants' English vocabulary knowledge. This quick and validated tool is particularly well suited for medium to highly proficient L2 English speakers, making it appropriate for the present study's participants. There are only 60 trials in this test, making it efficient and feasible to incorporate alongside other experimental tasks. According to a large-scale study by Lemhöfer and Broersma (2012), LexTALE functions not only as a reliable measure of vocabulary size but also as an indicator of general English proficiency. Furthermore, it offers a more objective alternative to self-rated proficiency measures. In this study, LexTALE scores were used to ensure that participants in both groups were comparable in terms of L2 proficiency.

The Musical Ear Test (MET; Wallentin et al., 2010) is a standardized instrument designed to measure musical perceptual ability, comprising Melody and Rhythm subsets. Given the focus of this study on speech rhythm, only the Rhythm subtest was administered. The online version of the MET has been validated by Correia et al. (2022) and is available as open-access materials on Gorilla and was implemented through the

Gorilla platform (https://app.gorilla.sc/openmaterials/218554). In the Rhythm subtest, participants completed 52 trials, each consisting of two rhythmical phrases presented using a woodblock sound. Half of the pairs were identical, while the other half differed. Participants judged whether each pair was the same or different within a fixed time limit. The trial order was randomized to reduce order effects. Scores on this task served as an index of participants' rhythmic perception ability, which was considered relevant to their potential for acquiring stress-based timing patterns in their L2 English.

3.2.2. Rhythmic Training Materials and Test Instruments

3.2.2.1. Training Materials and Pre/Post-test Instruments

The rhythmic training materials used in this study were designed by the author, drawing inspiration from the MusicSpeak (Wang et al., 2016), which maps English stress patterns onto musical beats to support rhythm acquisition. Although the system itself is not publicly accessible, its mapping mechanism was implemented using the following procedure:

1. Each sentence was transcribed using the CMU Pronunciation Dictionary (Weide, 2008).
2. Each word was assigned one of three stress levels —primary, secondary, or unstressed— based on dictionary markings.
3. Content words were prioritized for stress assignment, while function words were generally unstressed.
4. These stress levels were then mapped onto rhythmic beats in the training audio, with higher stress corresponding to longer and more prominent beats.

The training consisted of 32 utterances, each written in a conversational speech style and designed to feature distinct rhythmic patterns. For each sentence, the author recorded three audio tracks using GarageBand:

1. A drumbeat-only track representing the rhythmic pattern based on stress assignment.
2. A model reading of the sentence in natural English.
3. A combined track integrating the drumbeats with the model reading.

This three-track structure was intended to facilitate rhythm acquisition by combining rhythmic cues with natural speech imitation, in line with previous research that beat-

based activities could enhance the acquisition of rhythmic features (Llanes-Coromina et al., 2018; Picavet et al., 2012.; Wang et al., 2016).

The 32 utterances used in the training materials were organized into three sections, each corresponding to a different conversational topic and increasing levels of rhythmic complexity. Section 1 consisted of eight declarative utterances, including four utterances with four beats, two with five beats, and two with six beats. Section 2 included eight interrogative utterances, with four containing six beats and four with seven beats. Section 3 comprised 16 utterances with much greater rhythmic complexity, ranging from seven to 14 beats. In total, the training set included nine distinct beat types. This progression was designed to gradually enhance participants' sensitivity to rhythmic patterns in English speech. For example, a simple utterance such as "I'll have sushi" contains four evenly spaced beats, reflecting a regular rhythm. In contrast, a more complex one, such as "Tell me how to get to the library from here…", contains 10 beats with more varied inter-beat intervals. By following this progression order in training, participants were expected to build perceptual and productive control of speech rhythm. The complete list of training text materials is provided in Appendix 2 and the recording materials can be found in the online[4] OSF profile.

To evaluate the effectiveness of the rhythmic training intervention (for RQ1 and RQ2) and ensure the validity for statistical comparisons, half of the training utterances (16 out of 32) were selected for use in both the pre-test and post-test (see Appendix 1). These 16 utterances were randomly chosen but still represented all eight beat types, ensuring full rhythmic coverage. This design has two advantages: first, by reusing training materials in the training phase, any changes observed in rhythmic patterns from pre- to post-test can be more directly attributed to the training itself, despite the unavoidable influence of practice effects; second, using only half of the utterances helped reduce the likelihood that participants would fully memorize the training materials and simply recall them during the post-test. In both pre-test and post-test phases, utterances were presented in a randomized order to minimize memorization effects. Moreover, this design also tried to reduce the chance of learners following previously learned rhythm patterns in a fixed and mechanical manner.

[4] All recording materials can be found in the file folder via this view-only link in OSF: https://osf.io/y8rg3/?view_only=1a12284b3bd64addacad2f2790bf0381

3.2.2.2. Rhy-test Instruments

In addition to these repeated measures used for RQ1 and RQ2, a supplementary test (for RQ3), which is referred to as the rhy-test, was administered after the post-test. This test contained a separate set of six rhythmically structured utterances that were not part of the training or earlier testing phases. As discussed in the literature review, both Ariningsih (2023) and Finch (2003) suggested that poetic texts contain explicit rhythmic structures that are more easily detected by learners. Thus, the rhy-test materials were designed to assess learners' sensitivity to English stress-timed patterns without prior exposure. To avoid confounding effects from training or pre-test phase, these rhythmic items were only administered after the post-test, preventing participants from anticipating or rehearsing the rhythmic structures in advance. Unlike the 32 conversational style utterances, these new items were selected for their explicit metrical patterns in a poetic style, containing trochaic (strong–weak) and iambic (weak–strong) structures. For example, *"Sticks and stones will never gonna shake me"* demonstrates a primarily trochaic pattern, while *"They shine so soft across the sky"* follows an iambic rhythm. Although some lines may have slight adjustments for semantic and grammatical correctness, the underlying metrical structures were still explicit and relatively regular (see Appendix 3 for the full list of rhy-test utterances).

3.2.2.3. Follow-up Questionnaire after Training

As mentioned in the experiment procedure, a short follow-up questionnaire was designed to explore what strategies participants used during the training intervention (see Appendix 4 for the detailed question). Previous research has identified a range of effective pedagogical approaches for teaching rhythm, especially the use of physical gestures (Avery, 1992; Gower, 1983; Llanes-Coromina et al., 2018; O'Connor, 1980; Wong, 1987). Strategies such as tapping the table, clapping hands may also be employed directly by learners without explicit instruction, particularly when learners are exposed to a rhythm-focused learning environment. But this strategy-related question was not part of the main research aims. It was only included as a supplement to the primary study. Consequently, the results of this questionnaire are presented in the Discussion section as contextual information to complement the main findings.

3.3. Experiment Procedure

The experiment was designed and conducted online in Gorilla (www.gorilla.sc). All participants were encouraged to use a PC or laptop and to wear earphones throughout the experiment. They began by reading the participant information sheet and providing consent to participate, before completing a brief background questionnaire (see Appendix 4 for the beginning questionnaire) covering demographic information and English learning experience.

Participants were randomly assigned to either the experiment or the control group through a balanced randomizer embedded in Gorilla. The sequence of tasks differed slightly between groups to accommodate the training intervention in the experimental condition. For the control group, they first took the pre-test, followed by three basic ability tests (DS, LexTALE and MET). After completing these assessments, participants proceeded to complete the post-test and the rhythmic speech test (rhy-test). In contrast, Participants in the experiment group completed the pre-test first, followed by a rhythmic training session, after which they answered a short question on their training experience. They then completed the same three ability tests (DS, LexTALE, MET), followed by the post-test and rhy-test. This sequence arrangement was designed to create a time interval between the pre- and post-tests, minimizing the influence of memorization and reducing potential practice effects.

The pre-test, post-test, and rhy-test followed the same procedure. In each case, a sentence was presented on the screen. Participants were instructed to click the ‘Start’ button to initiate audio recording and the ‘Stop’ button to end it. They could record and self-correct their speech within a designated time limit. If the time expired before they clicked ‘Stop’, the task automatically progressed to the next item. The time limit was set at 15 seconds per sentence.

For the rhythmic training, each sentence was accompanied by three audio cues: a rhythmic beat track (recorded using GarageBand), a model spoken version (read and recorded by the author), and a combined version of both tracks. Participants were also encouraged to practice producing the rhythm and speaking simultaneously with the beat during the training phase.

Due to the complexity of the experiment, a pilot study involving four participants (one in the experimental group and three in the control group) was conducted before the formal launch. The pilot confirmed the technical functionality and feasibility of the experimental

design. Several adjustments were made based on participant feedback to improve clarity and user experience. First, a welcome screen was added with clearer task instructions and an overview of the experiment structure. Second, in response to concerns about task fatigue during MET, a five-minute optional break was added before its start. Although some participants suggested the possibility of shortening the test, this suggestion was not adopted as the MET is only validated in its full form. However, a progress countdown bar and motivational messages were added to sustain engagement. Additionally, the time limit for recording tasks (pre/post and rhy-tests) was extended from 10 seconds to 15 seconds, allowing more time for natural reading and reducing time pressure.

3.4. Acoustic Analysis and Segmentation

All audio recordings from the pre-test, post-test and rhy-tests were downloaded via Gorilla and converted to .wav format for acoustic analysis in Praat. The vowel segments in each sentence were manually annotated by the author, and their durations (in milliseconds) were extracted and recorded in an Excel spreadsheet for subsequent statistical analysis.

The most essential step in this acoustic analysis process was segmentation—identifying the start and end of each vowel for the measurement of the vocalic interval. Vowels are generally characterized by their clear and regular periodic waveform, reflecting consistent vocal fold vibration (Ladefoged, 2015).

Figure *1* illustrates how vowels appear in Praat in both waveform and spectrogram. In the waveform view (top panel), vowels appear as smooth, repeating cycles with relatively high amplitude, setting them apart from less sonorous consonants. For the spectrum view, vowels show prominent formants. The formant refers to a concentration of acoustic energy at particular frequencies that correspond to resonances in the vocal tract. There are usually several formants, each at a different frequency, roughly one in each 1,000Hz band. The first formant (F1) is inversely related to vowel height (the openness of the mouth), while the second formant (F2) relates to the vowel backness (the front-back position of the tongue). In the wideband spectrogram (see the bottom panel), vowels, represented by formants, show as horizontal dark bands that persist throughout the vowel's duration. These stable and identifiable acoustic features facilitated the identification of vowel boundaries in most cases.

Figure 1 Waveform and spectrogram of a vowel in Praat

In this study, vocalic intervals were measured from the vowel onset to the vowel offset, regardless of whether the segment contained a monophthong or a diphthong. While most boundaries could be identified using acoustic cues, some cases presented ambiguity, especially those transitions between consonants and vowels. In such instances, a consistent set of segmentation criteria was applied to ensure standardization. These criteria were based on the widely accepted guidelines proposed by Peterson and Lehiste (1960), with minor adaptations made to suit the specific aims and constraints of the present study. These are discussed below.

3.4.1. Initial Plosive

For vowels following initial plosives, segmentation procedures varied slightly depending on whether the plosive was voiceless or voiced. In the case of voiceless plosives, an initial burst is typically visible as a spike in the spectrogram, followed by a brief period of aspiration. This aspiration period appears as a low-energy region, particularly around the frequency of the first formant (F1), as seen in the /p/ of "pants". In contrast, voiced plosives generally lack the aspirated interval and instead present a more prominent frication phase, such as the /b/ in “base”.

Due to the variability and difficulty in consistently measuring the aspiration phase of plosives, vowel duration in these contexts was measured from the end of voice onset time (VOT). This point is identified in the waveform as when strong periodic striations become visible and in the spectrogram as the moment when F1 stabilizes. Figure 2 illustrates this process for the vowel in “pants” (following a voiceless plosive), while Figure 3 shows the vowel in “base” (following a voiced plosive). In both figures, arrows

mark the periodic striations and the stabilization point of F1, which together define the onset of the vowel. And the red zone indicates the vowel area.

Figure 2 Illustration of the segmentation of the vowel in “pants” from Praat

Figure 3 Illustration of the segmentation of the vowel in “base” from Praat

3.4.2. Initial and Final Nasals

Identifying the vowel onset following an initial nasal is generally straightforward. This boundary is usually marked by an abrupt acoustic shift—from the relatively stable, low-energy formant structure of the nasal segment to the rapid formant transitions characteristic of the vowel onset. In the waveform, this change is observed as the

emergence of regular vocal-fold periodicity, while in the spectrogram, it is indicated by a rise in F1 energy

In contrast, it is quite challenging to differentiate between a vowel and a following nasal consonant (e.g., /m/ in *dreams*, /n/ in *stones*) due to the nasalization. In the present study, the vowel offset was marked at the point where a sudden decrease in harmonic energy occurred. This change is particularly visible in the spectrogram: the strong harmonic bands of the vowel, which appear as dark horizontal lines, begin to fade as nasalization begins. Figure 4 displays the segmentation example of the vowel in “dreams”, with the red zone indicating the vowel area.

Figure 4 Illustration of the segmentation of the vowel in “dreams” from Praat

3.4.3. Initial Glide (/w/)

In most cases within the present data, the transition from an initial glide to the following vowel was acoustically smooth and continuous. However, in some tokens, the glide segment could be clearly distinguished from the vowel. In such cases, the glide was excluded from vowel duration measurements, as the analysis focused on the vowel alone. For example, when /w/ is followed by a front vowel such as /i/, the /w/ segment often exhibits a perceptible steady state with distinct acoustic characteristics. In the waveform, this appears as a lower-amplitude segment prior to the more prominent vowel part, represented by periodic striations. In the spectrogram, /w/ is characterized by low F1 and F2 starting frequencies that rise sharply during the transition into the vowel, providing a clear separation point (see Figure 5, and the red zone is the vowel area).

Figure 5 Illustration of the clear separation between the glide /w/ and vowel /i/ from Praat

By contrast, when /w/ is followed by a back vowel such as /ɑ/, the transition tends to be acoustically seamless, making it difficult to determine a clear boundary. In such cases, glides were included as part of the vocalic interval (see Figure 6 for an example of this type, with the red zone indicating the vowel area).

Figure 6 Illustration of the conjoining part of the glide /w/ and vowel /ɑ/ from Praat

3.5. Statistical Analysis

The primary quantitative analyses were conducted using R and R Studio. Participants' data, including scores from individual ability measures and acoustic indices of speech rhythm, were systematically cleaned, formatted, and prepared for statistical modelling. This section outlines the procedures used to quantify the participants' scores on the cognitive and linguistic tests, introduces an additional index of speech production, and explains the rationale for the statistical modelling approach adopted in this study.

3.5.1. Quantification of Individual Abilities

Three individual ability variables were measured to account for variability in participants' cognitive and linguistic abilities: DS, LexTALE and MET. Each was scored according to established protocols (Blackburn & Benton, 1957; Lemhöfer & Broersma, 2012; Wallentin et al., 2010; Woods et al., 2011).

The DS score was calculated as the proportion of correct responses from a set of 12 backward DS trials (Woods et al., 2011). Similarly, for MET, the proportion of correctly identified same/different rhythm pairs was calculated from the rhythm subset of the MET, which comprised 52 trials. This approach aligns with the scoring procedures reported in the original MET studies (Correia et al., 2022; Wallentin et al., 2010).

It is worth mentioning that the LexTALE test was scored using the "$\%correct_{av}$" (averaged % correct) proposed by Lemhöfer and Broersma (2012). This method adjusts

for the unequal number of words (n = 40) and non-words (n = 20) presented in the test. The final score was calculated using the following formula:

*((number of words correct/40*100) + (number of nonwords correct/20*100)) / 2*

All three scores were treated as continuous variables and were later entered as covariates in the main statistical models. This allowed the analysis to control for individual differences when evaluating the effectiveness of rhythmic training.

3.5.2. Indices of Speech Rhythm

As discussed in the literature review, this study quantified speech rhythm using two key indices: standard deviation (ΔV) and the normalized Pairwise Variability Index of vocalic intervals (nPVI-V), both calculated from the vowel durations in participants' speech recordings. These indices were chosen to capture rhythmic variation. Specifically, the durational variability of vowels is the central feature to determine whether speech is more syllable-timed or stress-timed.

The first metric, ΔV, provides a general measure of the dispersion of vowel durations within an utterance. It reflects how much the vocalic interval deviates from the mean and serves as a traditional representation of durational variability. However, as highlighted in the literature review, ΔV does not account for the temporal ordering of vowels and will be affected by overall speech rate. To address these limitations, the nPVI-V was included as another metric. The nPVI-V quantifies the relative difference in duration between each pair of adjacent vowels, providing information about the rhythmic alternation between longer and shorter segments. Unlike ΔV, the nPVI-V captures temporal patterns by analyzing how vowel durations vary from one to the next, rather than focusing on the overall dispersion. Additionally, it adjusts for speech rate by comparing each vowel pair with its average length. By employing both ΔV and nPVI-V, the study provides a comprehensive and reliable representation of the durational variability of vocalic intervals.

During the piloting phase, it was noted that some participants omitted or mispronounced words, which disrupted the vowel count and affected the calculation of ΔV and nPVI-V. To account for this situation, an additional variable, accuracy, was introduced to track the reliability of the measurement.

In this study, accuracy was operationalized in a narrow and objective manner, distinct from broader definitions used in L2 speech research. Traditionally, accuracy refers to

how closely a learner's spoken output matches native-like norms of accentedness, intelligibility and comprehensibility, normally evaluated by native speakers' perceptual judgements (Munro & Derwing, 1995; Thomson & Derwing, 2015). Instead, in this study, accuracy means whether the number of vowels produced in each sentence matched the expected number based on standard pronunciation provided in the Oxford English Dictionary IPA notation (Oxford University Press, 2024). For example, in the sentence *'how to get to the library from here'*, there are 10 expected vowels. If all 10 vowels were produced and visible in Praat (i.e., not devoiced, omitted, or merged), the sentence received an accuracy score of "1". If any vowel was missing or mispronounced to the extent that it could not be segmented, the accuracy was scored as "0". And these vowels were excluded from the ΔV and nPVI-V calculations to ensure consistency across participants. Although this narrow sense of accuracy and its binary coding system are restrictive and not representative of the accuracy construct as a whole, they were deemed appropriate for this study based on their objectivity and the alignment with this study's analytical focus.

3.5.3. Modelling Approach

To investigate the effectiveness of the rhythmic training intervention, a Linear Mixed Effects Regression (LMER) model was employed for the primary statistical analyses (Bates et al., 2015). This approach is particularly well-suited for experimental study data where repeated measures are taken from individual participants at multiple time points—such as during pre-test, post-test, and delayed post-test phases (Baayen et al., 2008). In this study, the R packages *lme4* (Bates et al., 2015) and *lmerTest* (Kuznetsova et al., 2017) were employed to fit LMER models and obtain p-values for fixed effects.

One major advantage of LMER is its flexibility: unlike ANOVA or other linear models, it does not require predictor or outcome variables to be normally distributed. Instead, it only assumes that the model residuals are normally distributed and homoscedastic. This allows researchers to model real-world data, which often does not meet normality assumptions. Additionally, the LMER can handle unbalanced designs and missing data better. Since data collection for the present study was conducted fully online, it was crucial to mitigate situations of data loss due to participant dropout or technical issues such as recording errors.

In the present design, each participant completed the pre-test, post-test and the additional rhy-test, creating a hierarchical data structure within subjects. Using the LMER model,

both fixed effects (e.g., time, group, MET, DS and LexTALE scores) and random effects (i.e., individual participant variability) could be measured. The primary dependent variables (ΔV and nPVI-V) were analysed as continuous outcomes, while time (pre/post), group (experiment/control) and their interaction served as the main predictors. Simultaneously, to control for inter-individual differences in three basic ability tests, the scores in the DS, MET and LexTALE were included as covariates. Also, the random intercept of each participant was added, acknowledging the dependence between repeated observations from the same person and capturing individual variability. Although the rhy-test involved a new set of materials and thus was not a repeated measure, an LMER was still applied because of its ability to model individual variability through random effects.

4. Results

4.1. Overview of Data Quality and Accuracy

Each participant was expected to produce 38 utterances across three test phases, containing 314 vocalic intervals to be measured per participant. This large sample size provided a robust basis for reliable speech rhythm quantification and allowed for a small margin of data loss.

During the data inspection, a small number of recordings were identified as invalid due to technical issues such as an empty file. Three of the participants (ID number: 6, 32 and 45) had more than five invalid recordings (8, 10, and 8 errors, respectively). Despite this, the overall proportion of usable data remained sufficient, and thus these participants were still included in the analysis.

In terms of speech accuracy, this was measured as the proportion of correctly produced vowels relative to the expected total. And the results are shown in Table 4. Overall, most participants produced the expected number of vowels. However, five participants (ID numbers: 5, 7, 16, 33 and 42) showed accuracy below 0.8 in at least one test phase. It is important to clarify that a low accuracy score does not necessarily reflect incorrect pronunciation. For example, the phrase ‘pair of’ was sometimes pronounced as ‘pairof’ due to vowel elision. In such cases, only one vowel could be segmented in Praat, resulting in an accuracy score of “0”. However, vowel elision is common in interconnected speech, especially when the speech style is fast and smooth (Celce-Murcia et al., 1996; Kenworthy, 1987). This highlights that the accuracy metric used here was to flag any major mismatches or omissions that could interfere with the main calculation and should not be interpreted as reflective of the participants’ pronunciation accuracy in a broader phonological sense.

Test Phase	Mean	SD
Pre-test	0.94	0.10
Post-test	0.92	0.10
Rhy-test	0.97	0.10

Table 4 Speech accuracy proportion in each test phase

4.2. Pre-intervention Group Comparability

Before addressing the main research questions, it is crucial to ensure that the experiment group and the control group were comparable in terms of three basic abilities. Checking the groups' comparability is important to exclude any alternative explanations for observed training effects. Although participants were randomly assigned to groups, this does not guarantee balanced allocation, especially in a relatively small sample size (n = 38). This section compares participants' basic abilities on three measures: working memory (DS), English proficiency (LexTALE), and rhythmic perception ability (MET). Descriptive statistics were first calculated for the three abilities, which are shown in Table 5.

Group	DS(Mean)	DS(SD)	Lex(Mean)	Lex(SD)	MET(Mean)	MET(SD)
Control	10.10	2.66	69.20	11.80	34.20	7.05
Experiment	11.20	1.07	73.50	13.60	33.50	7.06

Table 5 Descriptive statistics of individual ability tests by groups

Then, the normality of the distributions was assessed using Shapiro–Wilk tests. DS scores significantly deviated from normality in both the control group ($W = 0.71$, $p < .001$) and experiment group ($W = 0.70$, $p < .001$). MET scores were normally distributed in the control group ($W = 0.95$, $p = .456$), but not in the experiment group ($W = 0.86$, $p = .009$). LexTALE scores were approximately normally distributed in both groups (control: $W = 0.97$, $p = .709$; experiment: $W = 0.93$, $p = .185$).

Based on these findings, non-parametric Wilcoxon rank-sum tests were applied to DS and MET, and Welch's *t*-test was used for LexTALE. As reported in Table 6, no significant group differences were found in DS, LexTALE, or MET scores. These results confirm that the experiment and control groups were comparable in their cognitive, linguistic and rhythmic abilities before the intervention. Therefore, any observed post-training improvement can be more reliably attributed to the intervention rather than to pre-existing group differences.

Measure	Test Used	Test Statistic	p-value	Difference
DS	Wilcoxon rank-sum test	W = 140.5	.224	ns
LexTALE	Welch's t-test	t(35.31) = –1.03	.308	ns
MET	Wilcoxon rank-sum test	W = 174	.861	ns

Table 6 Group comparison on individual ability tests

4.3. The Effectiveness of Rhythmic Training (RQ1) and the Influence of Basic Abilities (RQ2)

To evaluate the impact of rhythmic training on learners' production of English stress-timed patterns (RQ1), and to examine whether individual abilities moderate the acquisition process (RQ2), a series of descriptive and statistical analyses was conducted. Before evaluating the training effect, independent-samples *t*-tests were performed on the pre-test values for ΔV and nPVI-V to assess the baseline comparability between the two groups. Results showed no significant group differences at pre-test for ΔV, $t(43.7) = -1.73$, $p = .09$, or nPVI-V, $t(44.7) = 0.08$, $p = .935$, indicating that the two groups were statistically equivalent in their initial rhythmic performance.

4.3.1. Descriptive Trends in Speech Rhythm (pre-post)

Descriptive statistics summarizing the pre- and post-test values for both ΔV and nPVI-V are presented in Table 7 and Table 8 and visualized in Figure 7 and Figure 8.

For ΔV, the control group demonstrated a slight decrease from pre-test to post-test, while the experiment group maintained relatively stable values from pre-test to post-test. This pattern suggests neither group exhibited substantial changes in overall vowel durational variability when measured by standard deviation.

For nPVI-V values, the control group also showed a small decrease from pre-test to post-test. In contrast, the experiment group exhibited an increasing trend from pre-test to post-test, which indicates a potential improvement after the training. These descriptive findings offer a preliminary insight into a possible training effect and set the stage for the formal statistical test using the LMER model.

Group	mean (pre)	sd (pre)	mean (post)	sd (post)
Control	53	11.2	51.3	14.1
Experiment	60.1	13.8	60.4	14.8

Table 7 ΔV by group in two test phases (pre-post)

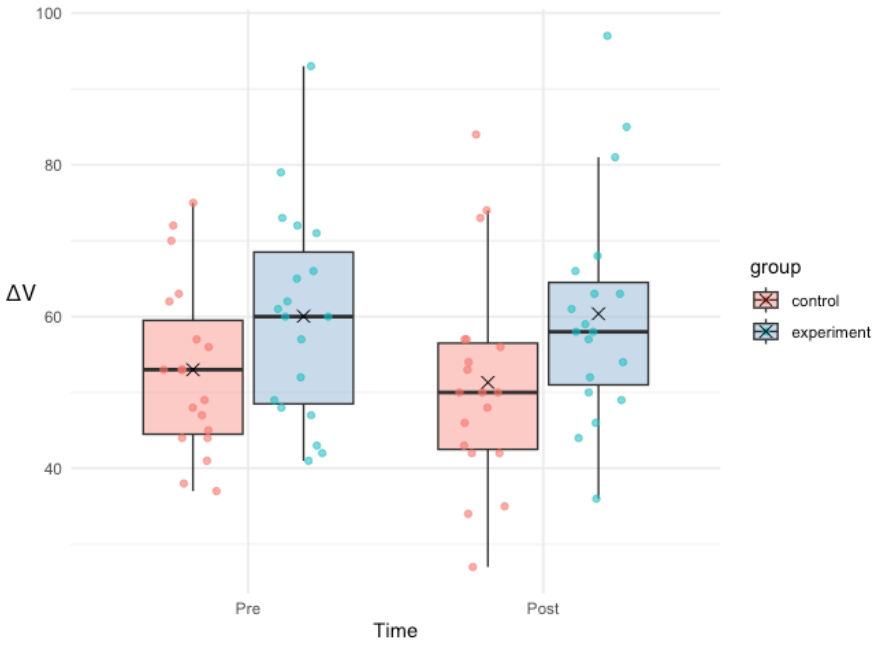


Figure 7 Boxplot of ΔV by group in two test phases (pre-post)

Group	mean (pre)	sd (pre)	mean (post)	sd (post)
Control	50.3	7.1	48	6.3
Experiment	50.2	4.5	52.5	6.7

Table 8 nPVI-V by group in two test phases (pre-post)

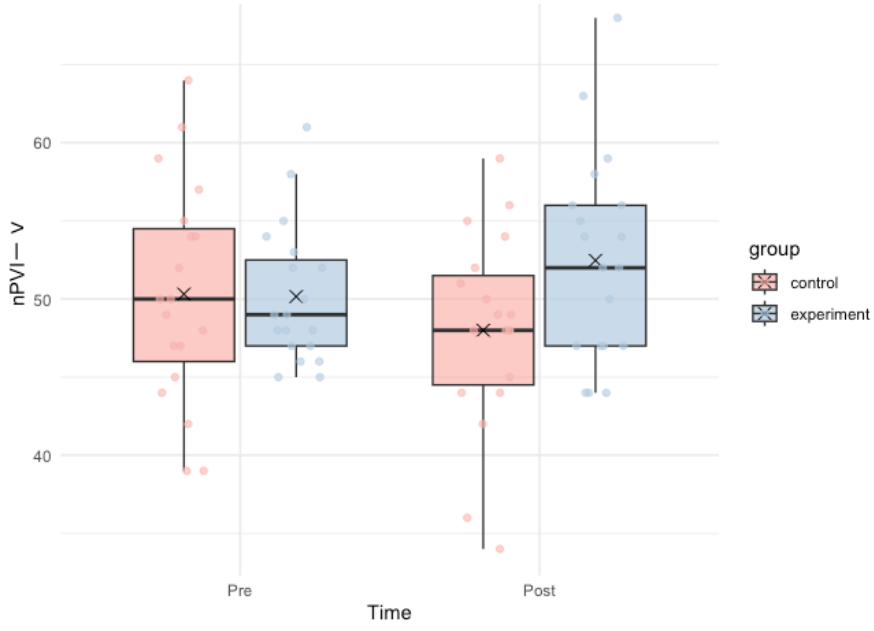


Figure 8 Boxplot of nPVI-V by group in two test phases (pre-post)

4.3.2. Statistical Modelling for Speech Rhythm (pre-post)

As discussed in section 3.5.3, an LEMR was employed for the primary data analysis to investigate the effectiveness of the training intervention. This model incorporates all fixed effects, such as time (pre-test or post-test), group (experiment or control), three ability scores, and random effects (i.e., individual participant variability). As a result, both RQ1 and RQ2 can be addressed simultaneously. Since there are two dependent variables, ΔV and nPVI-V, separate LEMR models were fitted for each:

```
Model 1 <- lmer(ΔV ~ time * group + MET + DS + lex_tale + (1 |
ID), data = ΔV)
Model 2<- lmer(nPVI-V ~ time * group + MET + DS + lex_tale + (1 |
ID), data = nPVI-V)
```

Prior to interpreting the results, key assumptions were assessed. Though the LMER model does not require the predictor or outcome variables to be normally distributed, it does assume that the residuals are approximately normally distributed and homoscedastic. By visual inspection of residual plots and Q–Q plots, the residuals appeared generally normally distributed and showed no clear pattern of heteroscedasticity. These diagnostic checks suggest that the LMER assumptions were reasonably met, supporting the validity of the following model inferences.

ΔV (Model 1):

Model 1 was fitted to investigate the effects of rhythmic training on the standard deviation of vocalic intervals (ΔV), which represents the overall variability in speech rhythm. The fixed effects included `time` (pre-test vs. post-test), `group` (control vs. experiment), and their interaction (`time * group`), which directly addresses RQ1 (i.e., whether changes in ΔV over time differ between groups). To address RQ2, three individual difference variables, musical rhythmic perception (MET), working memory (DS), and English proficiency (LexTALE), were entered as covariates. Most importantly, a random intercept for participant `(1 | ID)` was included to account for individual baseline variability in the repeated measures. As a result, any changes observed over time can be more confidently attributed to the training intervention, rather than to participants' specific characteristics.

The model results are summarized in Table 9. The `time * group` interaction was not statistically significant, indicating no significant difference from pre- to post-test between the experiment and control groups. The main effect of `time` and of `group` did not reach

statistical significance either, indicating that both groups exhibited similar pronunciation patterns across the pre- and post-test phases. But it is noticeable that the experiment group showed higher overall ΔV values than the control group across both time points. Moreover, none of the individual difference variables were significant predictors of ΔV changes.

Random effects revealed that individual variability accounted for a substantial portion of the total variance (*intercept variance* = 122.70, *SD* = 11.08), with residual variance at 65.06 (*SD* = 8.07).

Predictor	Estimate (b)	SE	df	t	p
(Intercept)	64.0219	16.346	33.4257	3.917	< .001***
Time (post)	-1.6842	2.6169	36	-0.644	0.524
Group (experiment)	8.323	4.6063	45.382	1.807	0.077
MET	0.1676	0.2976	33	0.563	0.577
DS	-0.2825	1.0788	33	-0.262	0.795
LexTALE	-0.2007	0.1702	33	-1.179	0.247
Time × Group	2	3.7009	36	0.54	0.592

Table 9 Results of Model 1 for ΔV

nPVI-V (Model 2):

Model 2 was fitted to investigate the effects of rhythmic training on the normalized Pairwise Variability Index of vocalic intervals (nPVI-V). Model 2 used the same fixed effects and covariates, as well as the same random intercept for participants as Model 1. The only change was the outcome variable—nPVI-V.

Results are shown in Table 10. The interaction between `time` and `group` was statistically significant, indicating that the experiment group demonstrated a greater pre-to-post increase than the control group in nPVI-V values. But the main effect of neither `time` nor `group` was statistically significant. This suggests that the training effect was specific to the experiment group and only emerged when considering how each group changed over time.

Among the covariates, MET scores emerged as a significant predictor of nPVI-V. This result suggests that participants with stronger rhythmic perception abilities tended to exhibit greater durational variability in their vowel production, reflecting more obvious stress-timed patterns when speaking L2 English. However, the other two abilities —

working memory (DS) and English proficiency level (LexTALE)— were not significant predictors.

Moreover, random effects revealed substantial individual variability in baseline nPVI-V scores (*intercept variance* = 13.33, *SD* = 3.65), with residual variance at 19.45 (*SD* = 4.41).

Predictor	Estimate (b)	SE	t	df	p
Intercept	36.23	6.32	5.73	33.86	< .001***
Time (post)	-2.32	1.43	-1.62	36	.114
Group (experiment)	-0.2	1.91	-0.11	55.84	.916
MET	0.36	0.11	3.13	33	**.004****
DS	0.42	0.42	1.01	33	.322
LexTALE	-0.03	0.07	-0.53	33	.598
Time × Group	4.63	2.02	2.29	36	**.028***

Table 10 Results of Model 2 for nPVI

4.4. Comparison between Text Types (RQ3)

In addition to the conversational texts designed for pre-test, post-test and training phases, a separate set of rhythmic texts was shown to participants during the rhy-test phase. These rhythmic texts were designed to contain more explicit poetic and metrical cues (e.g., iambic or trochaic patterns), offering an opportunity to explore whether different textual environments influence learners' production of English speech rhythm.

Since the rhy-test was introduced immediately after the post-test, comparison between these two datasets offered the most controlled conditions in terms of timing and task familiarity. Although the experiment group had received training before the rhy-test (unlike the control group), both groups completed the rhy-test under comparable conditions: at the same point in the study timeline, immediately after the post-test, and without any additional instructions or feedback in between. By this stage, both groups had become familiar with the task structure and had completed the MET task, potentially gaining their rhythmic awareness. In contrast, comparing the rhy-test to the pre-test would introduce confounding factors, such as order effects, task familiarity and practice effects, which may have complicated the interpretation. Nevertheless, when interpreting

the outcomes, it is important to consider the potential effect of the training received by the experiment group.

In the following analysis, post-test data are labelled as Con (Conversational texts) and rhy-test data are labelled as Rhy (Rhythmic texts), representing their text types. As in earlier analyses, the two acoustic metrics—ΔV and nPVI-V—were again employed to help us assess the rhythmic patterns of learners' L2 speech production.

4.4.1. Descriptive Statistics for the Influence of Text Types (Con vs. Rhy)

Descriptive statistics were calculated to provide a general insight into participants' rhythmic performance across text types. For ΔV (see Table 11 for data summary and Figure 9 for visualization), the control group had lower scores in the conversational condition than the rhythmic condition. The findings were similar for the experiment group.

For the nPVI-V metric (see Table 12 and Figure 10), the same pattern was observed. The nPVI-V values of both the control group and the experiment group were lower in the conversational condition than in the rhythmic condition. Interestingly, though both groups had similar mean nPVI-V values in the rhythmic condition, the control group's scores were a lot higher in the rhythmic condition as compared to their conversational baseline (i.e., their level of gains across two text types) than their experiment counterparts.

It is also noteworthy that, across both conditions, the experiment group consistently demonstrated higher ΔV values than the control group. This trend may reflect initial group differences in speech rhythm production, which will be further explored in the Discussion section.

Group	Mean (Con)	SD (Con)	Mean (Rhy)	SD (Rhy)
Control	51.3	14.1	73.9	15.0
Experiment	60.4	14.8	80.6	17.3

Table 11 ΔV by group in two text types (Con vs. Rhy)

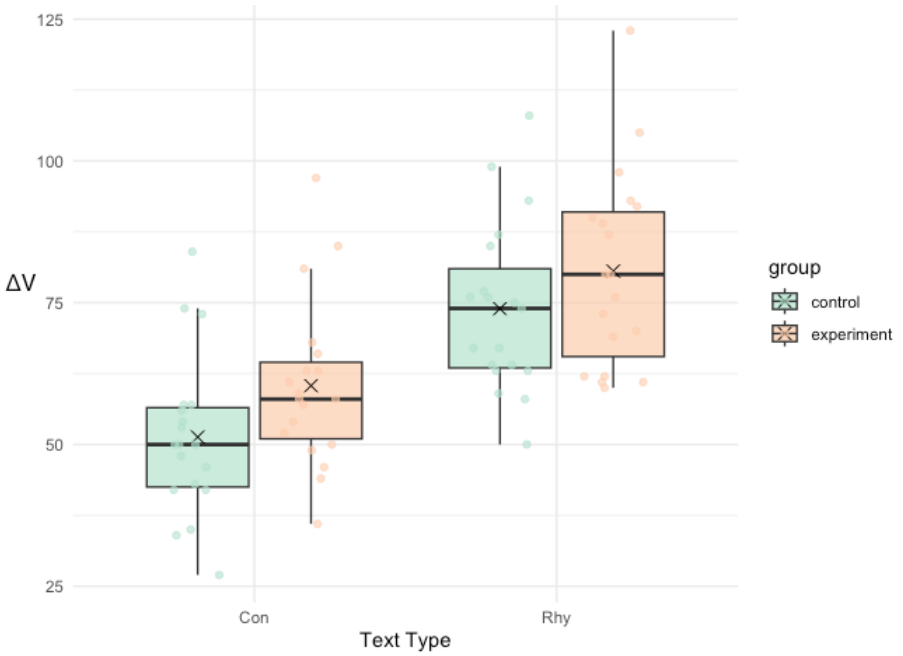


Figure 9 Boxplot of ΔV by group in two text types (Con vs. Rhy)

Group	Mean (Con)	SD (Con)	Mean (Rhy)	SD (Rhy)
Control	48.0	6.3	66.9	7.0
Experiment	52.5	6.7	66.7	10.4

Table 12 nPVI-V by group in two text types (Con vs. Rhy)

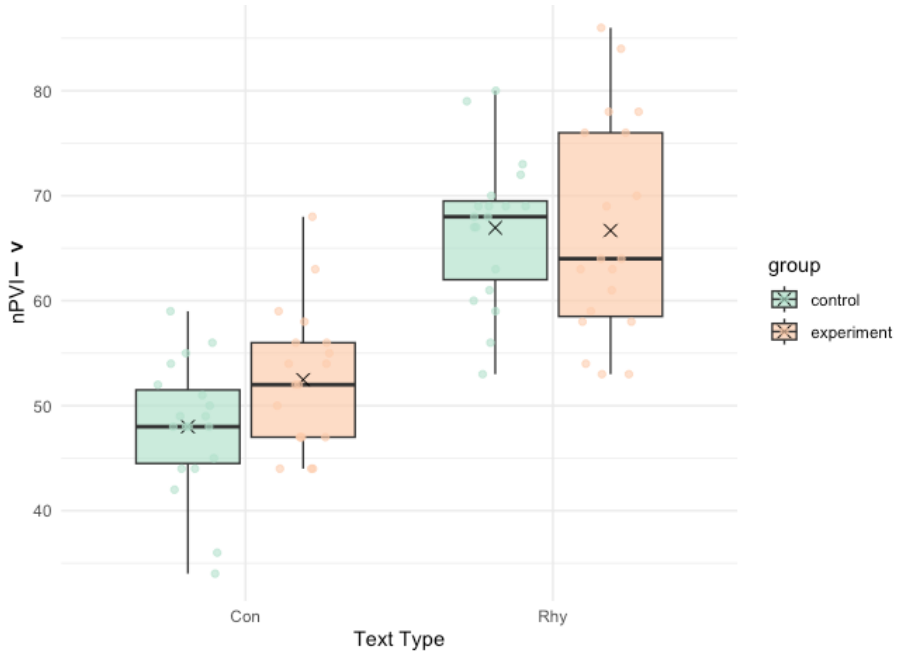


Figure 10 Boxplot of nPVI-V by group in two text types (Con vs. Rhy)

4.4.2. Statistical Modelling for the Influence of Text Types (Con vs. Rhy)

As discussed in section 3.5.3, an LMER was also employed to analyse any possible effect on the given text types. The two models (one for ΔV and one for nPVI-V) were as follows:

```
Model 3 <- lmer(△V ~ text_type * group + MET + DS + lex_tale + (1 | ID), data = △V _text)
Model 4 <- lmer(nPVI-V ~ text_type * group + MET + DS + lex_tale + (1 | ID), data = nPVI-V_text)
```

Before interpreting the model results, key assumptions of the LMER analysis were checked. Residual and Q–Q plots indicated approximate normality and homoscedasticity of residuals, suggesting that model assumptions were reasonably met.

ΔV (Model 3):

To explore whether text types read by participants influenced their speech rhythm, Model 3 was fitted with ΔV as the dependent variable. Fixed effects included `text_type` (Con vs. Rhy), `group` (experiment vs. control), and their interaction (`text_type * group`). Three individual difference variables: rhythmic perception ability (`MET`), working memory (`DS`), and English proficiency (`LexTALE`) were included as covariates to account for inter-individual variability in rhythmic performance. A random intercept for participant `(1 | ID)` was included to control for repeated measures and individual baseline differences.

Model results are presented in Table 13. No significant interaction was found between `text type` and `group`, suggesting that the effect of text types on ΔV did not differ meaningfully across groups. However, there were significant main effects of both the `text type` and `group` separately. These results suggest that, regardless of group differences, participants produced significantly greater ΔV values when reading rhythmic texts compared to conversational ones, indicating greater durational variability in vowel timing. Additionally, the experiment group consistently exhibited higher ΔV values across both text types compared to the control group. Among the covariates, none of the individual abilities emerged as significant predictors of ΔV values.

Predictor	Estimate (b)	SE	df	t	p
(Intercept)	83.82	17.73	33.35	4.73	< .001***
Text Type (Rhythmic)	22.63	2.56	36	8.84	**< .001*** **
Group (Experimental)	12.13	4.92	43.15	2.46	**.018***
MET	0.18	0.32	33	0.57	.573
DS	-1.31	1.17	33	-1.12	.273
LexTALE	-0.37	0.18	33	-2	.054
Text Type × Group	-2.42	3.62	36	-0.67	.508

Table 13 Results of Model 3 for ΔV

nPVI-V (Model 4):

Model 4 was fitted to examine the effects of text type and individual difference variables on nPVI-V values. As shown in Table 14, the interaction between `text type` and `group` was not significant, suggesting that the influence of text type on nPVI-V did not differ significantly between the experiment and control groups. But it is noteworthy that the coefficient was negative, indicating that the control group may have shown slightly larger gains in rhythmic texts compared to the experiment group.

For the main effects, a significant main effect of text type was observed. This indicates that participants in both groups produced significantly greater durational variability between adjacent vowels when reading rhythmic texts compared to conversational texts. Additionally, there was a significant main effect of group, with the experiment group showing higher overall nPVI-V values across both text types. This suggests that, regardless of the text condition, the experiment group produced speech with more stress-timed characteristics, potentially reflecting the impacts of the training intervention they received.

Among the covariates, the MET score emerged as a significant positive predictor of nPVI-V, suggesting that learners with better rhythmic perception skills tended to produce more stress-timed patterns (higher nPVI-V) regardless of text type. Neither working memory (`DS`) nor English proficiency (`LexTALE`) showed significant effects.

Predictor	Estimate (b)	SE	df	t	p
(Intercept)	46.77	8.18	33.93	5.72	< .001***
text_type (Rhythmic)	18.95	1.93	36	9.82	**< .001*** **
group (Experiment)	5.27	2.51	57.33	2.1	**.040***
MET	0.31	0.15	33	2.08	**.045***
DS	-0.02	0.54	33	-0.04	.968
LexTALE	-0.13	0.08	33	-1.55	.131
text_type × group	-4.74	2.73	36	-1.74	.091

Table 14 Results of Model 4 for nPVI-V

5. Discussion

The present study investigated the effectiveness of short-term rhythmic training on L2 English speech rhythm, with a particular focus on whether the intervention enables participants to use the more stress-timed pronunciation pattern which is appropriate in the L2. Two acoustic metrics (ΔV and nPVI-V) were used to quantify rhythmic variation. The first aim (RQ1) was to determine whether rhythmic training would enhance learners' stress-timed rhythm in L2 English. The second aim (RQ2) explored whether individual cognitive and linguistic abilities moderated speech rhythm performance. Additionally, the influence of text types on L2 rhythmic performance was examined as a RQ3, addressing an underexplored area due to limited prior empirical research.

Results showed that the experiment group had a statistically significant post-training increase in nPVI-V values compared to the control group, which partially supports our prediction for RQ1. Among individual abilities, the participants' rhythmic perception ability (measured by MET) emerged as a significant predictor of the performance of stress-timed patterns (higher ΔV and nPVI-V), highlighting its role in speech rhythm acquisition as predicted in RQ2. Interestingly, both groups demonstrated markedly obvious stress-timed pronunciations when reading rhythmic materials, suggesting the importance of explicit textual cues in achieving target-like speech rhythm. This result greatly supports the prediction of RQ3, showing that the type of text used in the pronunciation task can meaningfully shape learners' speech rhythmic patterns.

The discussion part is organized as follows: First, the effect of the training intervention is addressed, with further consideration of learners' reported strategies and the role of guided instruction versus rote practice. Then, individual differences, especially the impact of rhythmic perception skills, are illustrated. Following individual differences, the role of text types on L2 rhythmic production is discussed, providing insights into how rhythmic texts function in L2 learning contexts. The final two parts point out some limitations and pedagogical implications of this study.

5.1. The Effectiveness of Short-term Rhythmic Training (RQ1)

Models 1 and 2 addressed RQ1 by comparing pre- and post-test results between the control and experiment groups, using two acoustic metrics of vocalic interval variability: ΔV (Model 1) and nPVI-V (Model 2). These two models tested whether short-term rhythmic training led to more stress-timed rhythmic patterns in L2 English speech.

The findings of Model 1 do not support the hypothesis. The interaction between `time` and `group` was not statistically significant, suggesting that the training did not lead to measurable changes in ΔV values. In contrast, Model 2 revealed a significant `time*group` interaction, indicating that the experiment group exhibited a greater pre-to-post increase in nPVI-V values than the control group. In other words, participants in the experiment group (who received training) showed a marked increase in nPVI-V values from pre- to post-test, suggesting improved durational variability in vocalic intervals, which supports the effectiveness of rhythmic training intervention. As stated in previous research (e.g., Grabe & Low, 2008; Ling et al., 2000; Mok & Dellwo, 2008), increased durational variability between adjacent vowels is a defining characteristic of more stress-timed rhythmic patterns (such as English) compared to those syllable-timed ones (such as Mandarin). And the post-training increase in nPVI-V among participants who received training suggests a shift from L1-like syllable-timed patterns toward more L2 target-like, stress-timed pronunciation.

These findings not only reinforce existing literature but also extend previous research in L2 rhythm acquisition. For example, Wang et al. (2016) found that learners produced more stress-timed patterns when reading aloud synchronously with rhythmic beats, which were automatically generated to match the rhythm of the given text via the MusicSpeak system. However, their study focused exclusively on immediate, in-training effects and did not assess whether such rhythmic gains persisted beyond the imitation (training) context. By incorporating a post-test, the current study addresses this gap and demonstrates the persistence of this rhythm-based improvement. This suggests that short-term rhythmic training has the potential to produce effects beyond the intervention.

Furthermore, the focus of rhythm and rhythm-based training in the current study complements existing research on instructional approaches to pronunciation improvement. For example, Zhang and Yuan (2020) compared the effect of segmental-based and suprasegmental-based pronunciation instruction, stating that suprasegmental features, such as rhythm, stress, and intonation, significantly enhanced learners' overall speech comprehensibility. However, the broad scope and general instruction in their study made it difficult to isolate the contribution of individual features. The current study addressed this limitation by targeting rhythm specifically and evaluating its impact independently. The significant results not only support the beneficial role of

suprasegmental instruction but also offer more detailed evidence of how the rhythm-based training can be implemented in L2 pronunciation learning.

In addition, the findings align with the pedagogical arguments made by Tuan and An (2010), who advocated for the use of songs in English teaching. Their review emphasized that songs promote both the perception and the production of rhythm, potentially leading to developments in pronunciation. Since rhythm is a core component of songs, the effectiveness of rhythmic training observed in the current study echoes their claim that "the use of songs in the classroom stimulates the individual sounds and also promotes (speech) rhythm as well" (Tuan & An, 2010, p. 24).

However, though two acoustic indices were applied in this study, only the nPVI-V showed a statistically significant time and group interaction and supported the effectiveness of the training intervention. While both ΔV and nPVI-V quantify variability in vowel durations, ΔV failed to reveal any significant effects. This could be due to its limitations as a metric of speech rhythm. As previously noted, ΔV captures overall variability in vowel duration but does not reflect the temporal sequencing between adjacent vowel intervals. Together with its susceptibility to fluctuations in speech rate, these limitations may jointly reduce its reliability and stability as a measure of speech rhythm (White & Mattys, 2007). Reflected in the descriptive data shown in Table 7 and 8, the standard deviations for ΔV were notably larger than those for nPVI-V. While this could partly suggest greater measurement instability in ΔV, it may also reflect overall variability in participants' performance captured by this metric. This observation aligns with previous research assessing different rhythmic metrics that highlights two drawbacks of the ΔV measure (Barry et al., 2003.; Dellwo & Wagner, 2003; White & Mattys, 2007). The first key issue is that ΔV measures the raw variability in the duration of vocalic intervals without accounting for overall speech rate (Barry et al., 2003.; Dellwo & Wagner, 2003; White & Mattys, 2007). As White and Mattys (2007) pointed out, there is "an inverse correlation between speech rate and the score for standard deviation" (p. 14). This inverse correlation means that faster speech might have lower ΔV values compared to the slower style due to the overall lower duration. Given that participants in this current study were exposed to identical materials in both test phases (pre- and post-test) and were likely to get familiar with the task over time, they were expected to read faster in the post-test. This was especially true for the control group, who did not experience training.

The second limitation of the ΔV measure is its insensitivity to the temporal ordering of duration. In other words, ΔV ignores the order in which vowel sounds appear in speech

but just focuses on how varied the lengths of all the vowels are. This drawback is especially problematic when employing ΔV to distinguish between stress-timed and syllable-timed rhythm languages. This is because stress-timed rhythm is perceived through alternating strong–weak timing relations between adjacent segments (refers to vocalic intervals in this study), which might not be captured by the ΔV measurement. To illustrate this, consider one of the test utterances from the present study: *"Tell me how to get to the library from here."* Imagine two participants, A and B. Participant A possesses stress-timed rhythmic awareness and produces the utterance with alternating strong–weak patterns: *TELL me HOW to GET to the LIbrary from HERE*. Participant B, lacking such rhythmic sensitivity, produces the first half of the sentence with stressed vowels of longer duration, followed by a sequence of unstressed vowels: *TELL ME HOW TO GET to the library from here.* Even though these two pronunciation patterns are perceptually distinct, their ΔV values could be the same if the overall variability in vowel durations is coincidentally similar.

Unlike ΔV, nPVI-V considers both the timing between each pair of vowels and differences in speech rate. In simpler terms, nPVI-V measures how those vowel lengths change from one to the next, making it more efficient in capturing the alternating strong-weak patterns that are typically observed in more stress-timed languages like English. Because of this sensitivity to temporal sequences and rate normalization process, nPVI-V is considered a more robust and informative measure of speech rhythm than ΔV. Therefore, for researchers examining the effects of rhythm-based training or changes in speech rhythm, pairwise metrics such as nPVI are recommended over measures like standard deviation.

5.1.1. Observations on Guided Training vs. Rote Practice

An interesting pattern emerged from the descriptive data from the pre- and post-test (see Table 7 and Table 8). For both ΔV and nPVI-V metrics, the control group exhibited a slight decrease in mean values from the pre-test to the post-test phase (ΔV: from M = 53.0 to M = 51.3; nPVI-V: from M = 50.3 to M = 48.0). Although these changes were not statistically significant, they suggest a possible trend toward reduced variability in vocalic intervals, potentially reflecting a shift away from stress-timed rhythmic features among the control group in the post-test phase.

As discussed earlier in Section 5.1, the decline in ΔV values may be attributed to the increasing familiarity with testing materials. In other words, participants read more

fluently, leading to a faster speech rate during the post-test. Because ΔV is sensitive to speech rate, such speed changes might result in smaller variability scores. However, for the nPVI-V metric which has been normalized, changes in speech tempo would not lead to similar fluctuation. One possible interpretation is that the control group, by repeatedly practicing the same texts, may have reinforced their habitual rhythmic patterns from pre-test to post-test. As there was no training between the two phases, the control group might have treated the two tests as purely identical. When reading the same material the second time, they might have produced the sentences more naturally and employed their familiar rhythmic patterns. For most L2 English speakers in the current study, these patterns would be more syllable-timed due to the influence of their L1, Mandarin. Therefore, without explicit instruction, the control group's unguided repetition may have inadvertently deepened their L1-influenced rhythmic habits, rather than promoting their targeted-like stress-timed L2 patterns.

However, this interpretation of the ΔV values is speculative, as it was based solely on descriptive trends. Further empirical research directly comparing rote practice and guided instruction is necessary to strengthen this inference.

5.1.2. Strategy Use (from follow-up questionnaire)

Participants in the experiment group showed a general increasing trend in nPVI-V values following the training intervention. Based on this finding, a new question emerged about how learners engaged with the training task. Previous studies have suggested that rhythm-related behaviors, such as hand-clapping, beat gestures, can facilitate L2 performance (Baills et al., 2018; Llanes-Coromina et al., 2018). In order to capture any strategy that participants applied during the intervention, a follow-up questionnaire was presented immediately after the training session to all participants in the experiment group (n = 19). Table 15 summarizes the main strategy type reported, along with examples and their frequency of mention.

Strategy type	Examples	Frequency
Repetition	“Repeat silently/loudly”	12
Physical gesture	“Tapping, clapping, stomping”	8
Following the recording	“Following the provided audio” “Mimic audio”	6
Rhythm awareness	“Guess the beat” “Feel the rules”	3
No explicit strategy	“No strategy”	2

Table 15 Strategies used during training

Most participants reported repetition as the primary strategy. Although often viewed as basic, repetition is a powerful but underestimated tool in language learning (Mattys & Baddeley, 2019). Without enough repetition and practice, learners are unlikely to consolidate phonological patterns or progress toward more advanced forms of speech production. The second most frequently mentioned strategy was using physical gestures, such as tapping the table, clapping hands, or stomping feet. These movements serve to externalize and reinforce auditory structures. This resonates with prior studies, which have stated that the mapping of speech timing and physical movement is effective in language learning (Baills et al., 2018; Llanes-Coromina et al., 2018; Tierney & Kraus, 2014). Surprisingly, a small number of participants employed more advanced strategies such as guessing the rhythmic beats or trying to capture rhythmic rules. This rhythmic awareness reflects their metacognitive engagement with the training intervention. Moreover, two participants reported using ‘no strategy’. But this does not necessarily imply passive learning or a lack of metacognitive awareness during the training. These participants may have engaged in an unconscious learning process but need more explicit guidance to realize the strategies they used.

Though not the focus of the research questions, this questionnaire may help explain why the experiment group showed greater gains in stress-timed rhythmic patterns, particularly among those using movement-based and metacognitive strategies. Moreover, the findings of this questionnaire suggest potential pedagogical implications regarding the integration of strategy instruction into pronunciation training. These implications will be further discussed in the implications section.

5.2. Role of Rhythmic Perception Skills in Speech Rhythm Acquisition (RQ2)

The current findings provide meaningful insights into how individual abilities influence L2 speech rhythm acquisition, with particular emphasis on learners' rhythmic perception skills (measured by MET). In Model 2, the significant effect of MET on nPVI-V outcomes was found, offering partial support for RQ2, suggesting that greater rhythm sensibility may enhance learners' acquisition of speech rhythm. Therefore, among the three variables (working memory, English proficiency level, and rhythmic perception), only the rhythmic perception ability emerged as a consistent factor in shaping learners' production of stress-timed patterns in English. In other words, regardless of group differences and textual environments, participants with higher MET scores demonstrated greater durational variability in their vocalic intervals, which is the defining feature of stress-timed rhythmic patterns.

This finding resonates with previous studies that have established the link between musical aptitude and better L2 phonological learning outcomes (Posedel et al., 2012; Milovanov et al., 2008, 2010). For example, Posedel et al. (2012) found pitch perception to be a significant predictor of L2 Spanish pronunciation quality. Milovanov et al. (2008, 2010) reported that higher musical aptitude was closely associated with better English pronunciation for both children and adults. The present study extends prior research by isolating rhythmic perception as part of music aptitude and demonstrating that this skill greatly facilitates the L2 speech rhythm learning process.

In contrast, working memory and English proficiency did not significantly influence rhythmic acquisition in the current study. Regarding the language proficiency, the nonsignificant effect might be attributed to the characteristics of the sample group. All participants were advanced English learners with more than 10 years of learning experience, leading to minimal variability in proficiency level and possible ceiling effects in LexTALE results. As for working memory, though its association with various L2 abilities has been widely explored (Darcy et al., 2015; Mattys & Baddeley, 2019), its specific role in music-based L2 pronunciation training remains underexplored. Findings in existing literature are mixed: some suggested that auditory working memory mediates the relationship between musical aptitude and L2 outcomes (e.g., Herrera et al., 2011), while others did not find supportive evidence (e.g., Posedel et al., 2012). And studies that reported significant associations between working memory and L2 performance (Darcy et al., 2015; Mattys & Baddeley, 2019) often employed complex span tasks (such as L2

sentence recall tasks), which assess both storage and processing components of working memory. In contrast, this present study only used a simple digit span task, which may not be sensitive enough to capture the potential relationships between working memory and L2 pronunciation learning.

In sum, the positive role of rhythmic perception skills in this study strengthens the pedagogical importance of rhythm-related activities and instruction. Even without explicit training, learners with stronger rhythm sensitivity reacted more effectively to rhythmic cues and generally conveyed better L2 English pronunciation performance (stress-timed rhythmic patterns).

5.3. Influence of Text Type on Rhythmic Production (RQ3)

Models 3 and 4 addressed RQ3 by comparing participants' speech production across two text types: conversational and rhythmic. The results support the prediction that text type significantly influences L2 learners' English speech rhythm. Specifically, participants produced more stress-timed speech, as reflected in higher nPVI-V and ΔV values, when reading rhythmic texts compared to conversational ones. This finding aligns with earlier pedagogical-based proposals that poetic or rhythmically structured texts can enhance learners' rhythmic awareness (e.g., Finch, 2003; Ariningsih, 2023). It provides empirical support for the theoretical argument that poetic forms may benefit developing rhythmic sensitivity in L2 learners. The significant effect of text type, together with its alignment to the prior literature, provides a direction for further research on how the textual environment influences L2 pronunciation development.

Based on the present findings and existing research, two main factors may explain why rhythmic texts facilitate more stress-timed pronunciation. The most direct one is the presence of explicit rhythmic cues embedded within the utterances, such as trochaic and iambic structures. This explanation is supported by Ariningsih's (2023) systematic review, which emphasized the inherent suprasegmental features like stress and timing in poetic texts.

The second explanation relates to the design of the rhythmic materials. The six rhythmic utterances were sequenced in a progressive format, with the first two sharing a trochaic pattern and the remaining four following an iambic-like structure. This repetition may have helped learners detect recurring rhythmic patterns during the task. Even if the rhythmic cues were not immediately noticed in the initial utterances, the repetition of

similar items might have promoted implicit pattern recognition (Mattys & Baddeley, 2019).

Interestingly, though the main effect of `text type` was robust across Model 3 and 4, the interaction between `text type` and `group` (experiment vs. control) was not significant. Moreover, the negative estimate values indicated that the control group showed a greater 'boost' in stress-timed patterns than the experiment group from conversational to rhythmic texts. One plausible explanation for this trend lies in the baseline difference between groups. Specifically, the experiment group had already shown higher rhythmic variability (both in ΔV and nPVI-V) in the conversational condition (represented by post-test data), leaving less room for further improvement when transitioning to the rhythmic condition (i.e., ceiling effect). However, it is important to emphasize that these differences between groups cannot be attributed to the training intervention. This is because the statistical analyses in Models 3 and 4 were cross-sectional and did not account for repeated measures across time points, limiting any causal interpretation related to the training effect.

In sum, even without explicit instruction, learners' rhythmic awareness can be triggered through the specific textual environment, leading to the pedagogical potential of incorporating rhythmically rich texts into L2 teaching contexts.

5.4. Limitations

Three methodological limitations in the current study should be noted, as they may influence the interpretation and generalizability of the findings.

First, the relatively small sample size (n = 38) limits both statistical power and the generalizability of the results. With fewer participants, random variability can have a greater influence, meaning that patterns observed may be shaped by some participants rather than reflecting broader trends. In the present study, although recruitment targeted adults, most participants were aged 20–35, so the results may not represent the wider population of L2 English learners. The small sample also constrains the ability to detect smaller but potentially meaningful effects. Nonetheless, as an exploratory study designed to trial original rhythmic training materials, the primary goal of this study was not to draw definitive conclusions, but to explore promising directions and lay the groundwork for future, replicable research in this field.

Second, no delayed post-test was included in the experimental design, meaning the persistence of L2 rhythm learning outcomes could not be assessed. Although efforts were

made to maximize the time interval between test phases, the pre-test and post-test were still conducted within a single 40-minute session. Furthermore, the use of fully identical materials across both tests, together with the short time gap, may have introduced familiarity effects, potentially leading to improvements driven by repetition rather than by the training intervention. Despite these limitations, this approach was the most feasible option given the practical constraints of the study, such as time restrictions and the lack of funding for participant recruitment or extended testing sessions.

Third, the study lacked a direct comparison across conversational and rhythmic texts over time. Conversational texts were used in both the pre- and post-tests, whereas rhythmic texts were introduced only after the post-test. This design was intended to avoid pre-exposure to explicit rhythmic cues that might interfere with training effects. However, the design choice limits the possibility of comparing L2 performance in the rhythmic-text context at different time points. Future research could adopt a longitudinal design with a longer interval between test phases to minimize familiarity effects. Within the extended time frame, researchers can then combine rhythmic texts with conversational ones into both pre-test and post-test materials, allowing a more balanced within-subject comparison of how text type interacts with training outcomes.

5.5. Pedagogical Implications

Despite the listed methodological limitations, this study provides empirical evidence that rhythmic training can enhance L2 English stress-timed rhythmic patterns. These findings suggest several practical applications for teachers, curriculum developers, and material designers.

Integrating rhythmic training into pronunciation instruction. For L2 language teachers, especially those working with learners whose L1 rhythm differs substantially from the target L2 (e.g., syllable-timed Mandarin vs. stress-timed English), it is important to address rhythm as a key component of pronunciation. The current study indicates that even short-term rhythmic training can help learners shift from more syllable-timed toward more stress-timed patterns. Teachers could incorporate similar training into classroom activities or supplementary sessions to raise students' awareness of speech rhythm.

Employing rhythm-related strategies to enhance engagement. In addition to structured rhythmic training, other rhythm-related strategies, such as rhythmic perception

activities and the use of physical gestures, can be integrated into lessons to promote both pronunciation performance and learner engagement. This study found that rhythmic perception skills were associated with better stress-timed production, suggesting that musical activities, particularly those with strong rhythmic elements, could be embedded in curricula or serve as extracurricular workshops. Strategies, such as using body movements (e.g., clapping, tapping) to embody speech rhythm, can be introduced in classroom settings. While teachers can also prompt students to create their own movements to accompany spoken tasks, material designers can include guidance on gesture integration in textbooks or teaching resources.

Using rhythmic and poetic texts in pronunciation teaching. Although poetry is commonly used in writing instruction (e.g., Finch, 2003), its potential in pronunciation teaching remains underexplored. The present study found that learners produced more target-like rhythmic patterns when reading rhythmic texts than conversational ones, regardless of the training condition. This suggests that rhythmic and poetic materials may inherently support the learning of stress-timed patterns. Curriculum designers could therefore incorporate a great variety of rhythmic texts, such as poetry and lyrics, into pronunciation teaching materials. Getting exposed to such texts, learners' rhythmic awareness may be raised unconsciously.

6. Conclusion

This study investigated the effectiveness of short-term rhythmic training in enhancing L2 English speech rhythm, as well as the role of individual differences and the influence of text type in this language learning process. Participants were assigned to experiment and control groups, with only the experiment group receiving short-term rhythmic training. Speech rhythm was assessed through pre- and post-tests using conversational texts. As a follow-up, a separate rhythmic text test was conducted after the post-test.

Findings showed that rhythmic training significantly improved stress-timed rhythmic patterns in L2 English when measured by the nPVI-V metric, but not by ΔV. Rhythmic perception skills emerged as a crucial predictor of rhythmic training outcomes, whereas working memory and English proficiency level did not show significant effects. Text type was also found to influence L2 learners' speech performance, with rhythmic texts facilitating more stress-timed patterns than conversational ones.

Overall, the results of this study provide empirical support for the integration of short-term rhythmic training and rhythm-rich materials in L2 pronunciation teaching. Future longitudinal research with replicable designs is needed to further explore the role of rhythm and its interaction with individual learner variables in L2 English pronunciation learning and teaching.

References

Abercrombie, D. (1965). *Studies in phonetics and linguistics*. London.

Abercrombie, D. (1967). *Elements of General Phonetics*. Edinburgh University Press. https://www.jstor.org/stable/10.3366/j.ctvxcrw9t

Adams, C. (1979). *English speech rhythm and the foreign learner*. Mouton.

Allen, G. D., & Hawkins, S. (1980). Chapter 12 - PHONOLOGICAL RHYTHM: DEFINITION AND DEVELOPMENT. In G. H. Yeni-komshian, J. F. Kavanagh, & C. A. Ferguson (Eds), *Child Phonology* (pp. 227–256). Academic Press. https://doi.org/10.1016/B978-0-12-770601-6.50017-6

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Aoyama, K., & Guion, S. G. (2007). Acoustic analyses of duration and F0 range: Prosody in second language acquisition. In O.-S. Bohn & M. J. Munro (Eds), *Language Experience in Second Language Speech Learning: In honor of James Emil Flege* (pp. 281–297). John Benjamins Publishing Company. https://doi.org/10.1075/lllt.17.24aoy

Aoyama, K., & Reid, L. (2006). Cross-linguistic tendencies and durational contrasts in geminate consonants: An examination of Guinaang Bontok geminates. *Journal of The International Phonetic Association - J INT PHON ASSOC*, *36*. https://doi.org/10.1017/S0025100306002520

Ariningsih, T. W. (2023). Impact Of Poetic Style For Pronunciation: A Systematic Literature Review. *Interdiciplinary Journal and Hummanity (INJURITY)*, *2*(5), 390–400. https://doi.org/10.58631/injurity.v2i5.69

Avery, P. (with Ehrlich, S.). (1992). *Teaching American English pronunciation*. University Press.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Baills, F., Zhang, Y., & Prieto, P. (2018). Hand-clapping to the rhythm of newly learned words improves L2 pronunciation: Evidence from Catalan and Chinese learners of French. *Speech Prosody 2018*, 853–857. https://doi.org/10.21437/SpeechProsody.2018-172

Barry, W. J., Andreeva, B., Russo, M., Dimitrova, S., & Kostadinova, T. (2003). *Do Rhythm Measures Tell Us Anything about Language Type?*

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using**lme4**. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Blackburn, H. L., & Benton, A. L. (1957). Revised administration and scoring of the Digit Span Test. *Journal of Consulting Psychology*, *21*(2), 139–143. https://doi.org/10.1037/h0047235

Boersma, Paul & Weenink, David (2025). Praat: doing phonetics by computer [Computer program]. Version 6.4.39, retrieved 13 2025 from https://praat.org

Brown, G. (2017). *Listening to Spoken English* (2nd edn). Routledge. https://doi.org/10.4324/9781315538518

Celce-Murcia, M., Brinton, D., & Goodwin, J. M. (1996). *Teaching Pronunciation: A Reference for Teachers of English to Speakers of Other Languages*. Cambridge University Press.

Correia, A. I., Vincenzi, M., Vanzella, P., Pinheiro, A. P., Lima, C. F., & Schellenberg, E. G. (2022). Can musical ability be tested online? *Behavior Research Methods, 54*(2), 955–969. https://doi.org/10.3758/s13428-021-01641-2

Crystal, T. H., & House, A. S. (1988). The duration of American-English vowels: An overview. *Journal of Phonetics, 16*(3), 263–284. https://doi.org/10.1016/S0095-4470(19)30500-5

Darcy, I., Park, H., & Yang, C.-L. (2015). Individual differences in L2 acquisition of English phonology: The relation between cognitive abilities and phonological processing. *Learning and Individual Differences, 40*, 63–72. https://doi.org/10.1016/j.lindif.2015.04.005

Dauer, R. M. (1983). Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics, 11*(1), 51–62. https://doi.org/10.1016/S0095-4470(19)30776-4

Dellwo, V., & Wagner, P. (2003). *Relations between language rhythm and speech rate*. https://doi.org/10.5167/UZH-111779

Deterding, D. (1994). The rhythm of Singapore English. In *Paper presented at Fifth Australian International Conference on Speech Science and Technology* (Vol. 6, p. 8).

Finch, A. (2003). Using poems to teach English. *English Language Teaching, 15*(2), 29–45.

Fischler, J. (2005). The rap on stress: Instruction of word and sentence stress through rap music. *School of Education and Leadership Student Capstone Theses and Dissertations*. https://digitalcommons.hamline.edu/hse_all/315

Gay, T. (2009). Mechanisms in the Control of Speech Rate. *Phonetica, 38*(1–3), 148–158. https://doi.org/10.1159/000260020

Gilbert, J. B. (Judy B. (with Internet Archive). (2008). *Teaching pronunciation: Using the Prosody Pyramid*. Cambridge, UK ; New York : Cambridge University Press. http://archive.org/details/teachingpronunci0000gilb

Gordon, R. L., Fehd, H. M., & McCandliss, B. D. (2015). Does Music Training Enhance Literacy Skills? A Meta-Analysis. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01777

Gower, R. (with Walters, S.). (1983). *Teaching practice handbook*. Heinemann Educational Books.

Grabe, E., & Low, E. L. (2008). Durational variability in speech and the Rhythm Class Hypothesis. In C. Gussenhoven & N. Warner (Eds), *Laboratory Phonology 7* (pp. 515–546). De Gruyter Mouton. https://doi.org/10.1515/9783110197105.2.515

Graham, C. (with Internet Archive). (1993). *Grammarchants: More jazz chants*. New York, NY, USA : Oxford University Press. http://archive.org/details/grammarchantsmor0000grah

Gut, U. (2003). *Prosody in second language speech production: The role of the native language: Mündliche Produktion in der Fremdsprache.*

Herrera, L., Lorenzo, O., Defior, S., Fernandez-Smith, G., & Costa-Giomi, E. (2011). Effects of phonological and musical training on the reading readiness of native- and foreign-Spanish-speaking children. *Psychology of Music*, *39*(1), 68–81. https://doi.org/10.1177/0305735610361995

Kenworthy, J. (1987). *Teaching English pronunciation*. Longman.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*, 1–26. https://doi.org/10.18637/jss.v082.i13

Ladefoged, P. (with Johnson, K.). (2015). *A course in phonetics* (Seventh edition / Peter Ladefoged, Keith Johnson.). Cengage Learning.

Lehiste, I. (1977). Isochrony reconsidered. *Journal of Phonetics*, *5*(3), 253–263. https://doi.org/10.1016/S0095-4470(19)31139-8

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods*, *2*, 325–343. https://doi.org/10.3758/s13428-011-0146-0

Lezak, M. D., Howieson, D. B., Bigler, E. D., Tranel, D., Lezak, M. D., Howieson, D. B., Bigler, E. D., & Tranel, D. (2012). *Neuropsychological Assessment*. Oxford University Press.

Lin, H., & Wang, Q. (2005). *Vowel Quantity and Consonant Variance: A Comparison Between Chinese and English*.

Ling, L. E., Grabe, E., & Nolan, F. (2000). Q uantitative Characterizations of Speech Rhythm: Syllable-Timing in Singapore English. *Language and Speech*, *43*(4), 377–401. https://doi.org/10.1177/00238309000430040301

Llanes-Coromina, J., Prieto, P., & Rohrer, P. (2018). Brief training with rhythmic beat gestures helps L2 pronunciation in a reading aloud task. *Speech Prosody 2018*, 498–502. https://doi.org/10.21437/SpeechProsody.2018-101

Low, E. L. (1994). *Intonation patterns in Singapore English* (M.Phil. dissertation). Department of Linguistics, University of Cambridge.

Low, E. L. (1998). *Prosodic prominence in Singapore English.* https://www.repository.cam.ac.uk/handle/1810/251470

Mattys, S., & Baddeley, A. D. (2019). Working Memory and Second-Language Accent Acquisition. *Applied Cognitive Psychology*. https://doi.org/10.1002/acp.3554

Milovanov, R., Pietilä, P., Tervaniemi, M., & Esquef, P. A. A. (2010). Foreign language pronunciation skills and musical aptitude: A study of Finnish adults with higher education. *Learning and Individual Differences*, *20*(1), 56–60. https://doi.org/10.1016/j.lindif.2009.11.003

Mok, P. P. K., & Dellwo, V. (2008). Comparing native and non-native speech rhythm using acoustic rhythmic measures: Cantonese, Beijing Mandarin and English. *Speech Prosody 2008*, 423–426. https://doi.org/10.21437/SpeechProsody.2008-93

Munro, M. J., & Derwing, T. M. (1995). Foreign Accent, Comprehensibility, and Intelligibility in the Speech of Second Language Learners. *Language Learning*, *45*(1), 73–97. https://doi.org/10.1111/j.1467-1770.1995.tb00963.x

Nolan, F., & Asu, E. L. (2009). The Pairwise Variability Index and Coexisting Rhythms in Language. *Phonetica*, *66*(1–2), 64–77. https://doi.org/10.1159/000208931

Oxford University Press. (2024). *Oxford English Dictionary* (Online ed.). https://www.oed.com/

O'Connor, J. D. (1980). *Better English pronunciation* (2nd ed). University Press.

Peterson, G. E., & Lehiste, I. (1960). Duration of Syllable Nuclei in English. *The Journal of the Acoustical Society of America*, *32*(6), 693–703. https://doi.org/10.1121/1.1908183

Picavet, F., Aubergé, V., & Rossato, S. (2012). *CAN A GUIDED RHYTHMIC APPROACH CONTRIBUTE TO THE ORAL PERFORMANCE OF LEARNERS OF L2 ENGLISH? A CASE STUDY*.

Picciotti, P. M., Bussu, F., Calò, L., Gallus, R., Scarano, E., Cintio, G. D., Cassarà, F., & D'alatri, L. (2018). Correlation between musical aptitude and learning foreign languages: An epidemiological study in secondary school Italian students. *ACTA*

Otorhinolaryngologica Italica, *38*(1), Article 1. https://doi.org/10.14639/0392-100X-1103

Pike, K. L. (1946). *The intonation of American English*. University of Michigan Press.

Posedel, J., Emery, L., Souza, B., & Fountain, C. (2012). Pitch perception, working memory, and second-language phonological production. *Psychology of Music*, *40*(4), 508–517. https://doi.org/10.1177/0305735611415145

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

Ramus, F., Nespor, M., & Mehler, J. (1999). *Correlates of linguistic rhythm in the speech signalq*.

Roach, P. (1982). On the distinction between “stress-timed” and “syllable-timed” languages. In D. Crystal (Ed.), *Linguistic controversies: Essays in linguistic theory and practice in honour of F. R. Palmer* (pp. 73–79). Edward Arnold.

Sun, B., Yang, J., & Liang, Z. (2024). How music facilitates second language (L2) learning: A systematic review from 2002 to 2022. *Innovation in Language Learning and Teaching*, *0*(0), 1–19. https://doi.org/10.1080/17501229.2024.2350490

Suzukida, Y. (2021). The Contribution of Individual Differences to L2 Pronunciation Learning: Insights from Research and Pedagogical Implications. *RELC Journal*, *52*(1), 48–61. https://doi.org/10.1177/0033688220987655

Thomson, R. I., & Derwing, T. M. (2015). The Effectiveness of L2 Pronunciation Instruction: A Narrative Review. *Applied Linguistics*, *36*(3), 326–344. https://doi.org/10.1093/applin/amu076

Tierney, A., & Kraus, N. (2014). Auditory-motor entrainment and phonological skills: Precise auditory timing hypothesis (PATH). *Frontiers in Human Neuroscience, 8*. https://doi.org/10.3389/fnhum.2014.00949

Tuan, L. T., & An, P. T. V. (2010). *Teaching English Rhythm by Using Songs*.

Ur, P. (1996). *A course in language teaching: Practice and theory*. University Press.

Wallentin, M., Nielsen, A. H., Friis-Olivarius, M., Vuust, C., & Vuust, P. (2010). The Musical Ear Test, a new reliable test for measuring musical competence. *Learning and Individual Differences, 20*(3), 188–196. https://doi.org/10.1016/j.lindif.2010.02.004

Wang, H., Mok, P., & Meng, H. (2010). *MusicSpeak: Capitalizing on Musical Rhythm for Prosodic Training in Computer-Aided Language Learning*.

Wang, H., Mok, P., & Meng, H. (2016). Capitalizing on musical rhythm for prosodic training in computer-aided language learning. *Computer Speech & Language, 37*, 67–81. https://doi.org/10.1016/j.csl.2015.10.002

Wechsler, D. (1981). Manual for the Wechsler Adult Intelligence Scale-III. *Psychol Corp, 70*.

Wechsler, D. (2012). *Wechsler Adult Intelligence Scale—Fourth Edition* [Data set]. American Psychological Association (APA). https://doi.org/10.1037/t15169-000

Wei, M. (2006). A Literature Review on Strategies for Teaching Pronunciation. In *Online Submission*. https://eric.ed.gov/?id=ED491566

White, L., & Mattys, S. L. (2007). Calibrating rhythm: First language and second language studies. *Journal of Phonetics, 35*(4), 501–522. https://doi.org/10.1016/j.wocn.2007.02.003

Whitworth, N. (2002). Speech rhythm production in three German–English bilingual families. *Leeds Working Papers in Linguistics and Phonetics, 9*, 175–205.

Wiener, S., & Bradley, E. (2020). Harnessing the musician advantage: Short-term musical training affects non-native cue weighting of linguistic pitch. *Language Teaching Research*. https://doi.org/10.1177/1362168820971791

Wong, R. (1987). *Teaching Pronunciation: Focus on English Rhythm and Intonation. Language in Education: Theory and Practice, No. 68*. Prentice-Hall, Inc. https://eric.ed.gov/?id=ED283386

Woods, D. L., Kishiyama, M. M., Yund, E. W., Herron, T. J., Edwards, B., Poliva, O., Hink, R. F., & Reed, B. (2011). Improving digit span assessment of short-term verbal memory. *Journal of Clinical and Experimental Neuropsychology*, *33*(1), 101–111. https://doi.org/10.1080/13803395.2010.493149

Yalçın, Ş., Çeçen, S., & Erçetin, G. (2016). The relationship between aptitude and working memory: An instructed SLA context. *Language Awareness*, *25*(1–2), 144–158. https://doi.org/10.1080/09658416.2015.1122026

Zhang, R., & Yuan, Z. (2020). Examining the effects of explicit pronunciation instruction on the development of L2 pronunciation. *Studies in Second Language Acquisition*, *42*(4), 905–918. https://doi.org/10.1017/S0272263120000121

Appendice

Appendix 1—The full list of pre-test and post-test materials (16 utterances in randomized order)

1. I’ll have sushi.
2. I’ll have gulash.
3. I’ll have spaghetti.
4. I’ll have pepperoni.
5. How much is this sweater?
6. How much is this baseball?
7. How much is this pair of pants?
8. How much is this bag of socks?
9. To the library from here...
10. To the bus station from here...
11. How to get to the library from here?
12. How to get to the bus station from here?
13. Tell me how to get to the library from here?
14. Tell me how to get to the bus station from here?
15. Can you tell me how to get to the library from here?
16. Can you tell me how to get to the bus station from here?

Appendix 2—The full list of training materials (32 utterances)

*Utterances with (T) were for the training section only

Section 1:

four beats

1. I'll have sushi.
2. I'll have gulash.
3. I'll have tacos. (T)
4. I'll have curry. (T)

Five beats

5. I'll have spaghetti.
6. I'll have pepperoni.

Six beats

7. I'll have a hamburger. (T)
8. I'll have a cranberry. (T)

Section 2:

Six beats

1. How much is this sweater?
2. How much is this skateboard? (T)
3. How much is this baseball?
4. How much is this jacket? (T)

Seven beats

5. How much is this pair of pants?
6. How much is this wall of hats? (T)
7. How much is this bag of socks?
8. How much is this rack of shirts? (T)

Section 3:

Seven beats

1. To the library from here…
2. To the bus station from here…
3. To the shopping mall from here… (T)
4. To the post office from here…(T)

Ten beats

5. How to get to the library from here…

6. How to get to the bus station from here…
7. How to get to the shopping mall from here…(T)
8. How to get to the post office from here…(T)

Twelve beats

9. Tell me how to get to the library from here…
10. Tell me how to get to the bus station from here…
11. Tell me how to get to the shopping mall from here… (T)
12. Tell me how to get to the post office from here…(T)

Fourteen beats

13. Can you tell me how to get to the library from here?
14. Can you tell me how to get to the bus station from here?
15. Can you tell me how to get to the post office from here? (T)
16. Can you tell me how to get to the shopping mall from here? (T)

Appendix 3—The full list of rhy-test materials (six utterances)

1. Sticks and stones are never gonna shake me.
2. Winds and waves will never try to break me.
3. The moon is bright, the stars are high.
4. They shine so soft across the sky.
5. They watch the world so big and wide.
6. They follow dreams where wishes hide.

Appendix 4—Questionnaires (before pre-test/after training intervention)

Language Learning Experience Questionnaire (before pre-test)
Thank you for participating in this experiment. Your responses will help us better understand your background and English language use. Please answer the following questions as accurately as possible.

1. **Age:**
   _____ years old

2. **Gender:**
   ☐ Male
   ☐ Female
   ☐ Non-binary
   ☐ Prefer not to say

3. **How long have you been learning English?**
   _____ years

4. **At what age did you first start learning English?**
   Age: _____

5. **Where are you currently based?**
   Country: _____

6. **How long have you lived in an English-speaking country (in total)?**
   ☐ I have never lived in an English-speaking country
   ☐ 0–1 year
   ☐ 1–3 years
   ☐ 4–6 years
   ☐ More than 7 years

For the questions below, please consider your habits in recent months.

7. **How often do you listen to English (e.g., conversations, media, audio)?**
   ☐ Never
   ☐ Not very often
   ☐ Sometimes
   ☐ Quite often
   ☐ All the time

8. **How often do you read in English (e.g., texts, websites, books)?**
   ☐ Never
   ☐ Not very often
   ☐ Sometimes
   ☐ Quite often
   ☐ All the time

9. **How often do you speak English (e.g., in conversations, presentations, class)?**
   - ☐ Never
   - ☐ Not very often
   - ☐ Sometimes
   - ☐ Quite often
   - ☐ All the time

10. **How often do you write in English (e.g., messages, emails, assignments)?**
   - ☐ Never
   - ☐ Not very often
   - ☐ Sometimes
   - ☐ Quite often
   - ☐ All the time

---

follow-up question (after training intervention for the experiment group)

For the activity just now (the one you heard beats and sentences), what strategies do you use during the process?
(For example: clapping hands or tapping the table to create your own beats, reading or repeating the sentences out loud while following the recordings, etc.)
*If you did not use any specific strategy, please write "no strategy."

Answer:_____

Appendix 5—CUREC Approval

Education (Educ) DREC
15 Norham Gardens, Oxford, OX2 6PY

Applicant: Yanhao Li
Principal Investigator: Faidra Faitaki
Department: Education

Study title: Short-term Rhythmic Training for L2 Pronunciation Enhancement
(Version: 2.0)

Ethics reference: Education (Educ) DREC - 1351018

Dear Faidra Faitaki,

On behalf of the Committee, I confirm that the above research study described in the application and other supporting documentation submitted to the committee has been carefully considered on behalf of the Education (Educ) DREC in accordance with the University's regulations and policy for ethics approval of research involving human participants, human tissue and/or personal data. The opinion is as follows:

Opinion of Research Ethics Committee: Favourable Opinion

Subject to the following conditions:

Decision Date: 5 Aug 2025, 09:40

Opinion End Date: 14 Oct 2026

If favourable, insurance-provided indemnity arrangements will be in place between the decision date and opinion end date and you may now commence your study activities. Should you plan to continue the research beyond the end date above, it is your responsibility to ensure that you request, and receive, an extension (via amendment) from the committee for indemnity to remain in place. You may be required to provide a justification.

Please note the following:

Amendments: Should there be any subsequent changes to the reviewed study, applications for amendments can be made via the Oxford Ethics Application System (Worktribe Ethics).

Reports: Studies considered by OxTREC are expected to submit an *annual progress report* on each anniversary of study approval, until the study is completed. An end of study report is also required.

Audit: This study may be selected for audit at the discretion of the Research Governance, Ethics and Assurance Team.

Data safety: It is the responsibility of the PI to ensure that all data collected during the course of the study is stored and transferred safely and securely in accordance with University requirements. Further guidance and advice are available from the Research Data Team. Additional information is available at https://researchsupport.web.ox.ac.uk/governance/ethics

Yours Sincerely

Education Ethics Officer