

Towards Automating the Rhythmic Analysis of Speech: A Comparative Study of English Spoken by American and Thai Speakers

Sirinda Palahan¹ & Yuwaree Tongvivat² 

Abstract

Efforts to classify language rhythm have significantly contributed to methodologies for identifying acoustic features that define speech “beats”. This study advances the field through three key contributions: (1) validation of algorithmic rhythm detection, (2) characterization of distinctive rhythmic features in native versus non-native speech, and (3) development of an automated, scalable approach for rhythmic analysis. The effectiveness of the *maxD* parameter—representing the moment of fastest energy increase—was evaluated as an alternative to manual annotation for identifying syllabic beats. Speech samples from 34 speakers (17 American and 17 Thai) were analyzed using statistical and machine learning models, including Support Vector Machine, Random Forest, Gradient Boosting, and Logistic Regression, to classify rhythmic patterns. Results indicate that the *maxD* parameter demonstrates strong alignment with manually annotated beat locations, achieving high predictive accuracy for native speakers (RMSE: 0.1182, MAPE: 0.3290%). Additionally, SHapley Additive exPlanations (SHAP) analysis revealed key rhythmic features distinguishing the two groups: *maxD value*, *intensity ratio*, and *max intensity* were identified as key characteristics of native speech, while shorter vowel and syllable durations were established as distinctive native features. Systematic differences in *maxD*, pitch, and intensity alignment patterns between native and non-native speech were documented, particularly in content words and multisyllabic structures. These findings contribute to the ultimate goal of improving non-native speech by identifying critical rhythmic features that differentiate native and non-native speakers. This research study lays the groundwork for targeted pronunciation training tools and language-learning applications, offering a foundation for enhancing naturalness and fluency in second language acquisition.

Keyword: *maxD*, language beat, L2 pronunciation, rhythm detection, speech technology

Introduction

Rhythm is referred to the sequential pattern of durations, which is relevant to, but not necessarily dependent with, phrase structure in the sense that timing and tempo interplay (Gabrielsson, 1999; Honing, 2002; London, 2001). Rhythm can be identified based on the components of each music note within an interval, which is related to shorter, equal or longer duration of timing between one note and another, while perceived tempo involves beats which are focused on listeners’ sensation of pulses. Timing is another element which involves the duration of music notes but can be deviated by “playing after the beat,” when lengthening accentuating notes for expressive purposes (Honing, 2002, pp. 1-5). In the application for rhythm in language based on these concepts, recurrent patterns of the elements of individual syllable would contribute to a rhythm. The recurrent (more) prominent sounds of the syllables within a phrase or sentence would correspond to the tempo, which is associated with the notion of “beats.” Timing would somewhat coincide with the component of sounds that contribute to rhythm (perceived as e.g. relaxing, laidback, elegant), but timing can be lengthened or shortened upon speakers and situations. Whether language exhibits a similar rhythmic structure, particularly regarding the emphasis of voices within consistently equal patterns in speech, remains a subject of debate. Linguists have identified recurring patterns of emphasis in language, where sounds are relatively evenly distributed across syllables. These patterns, by extension, manifest as predefined rhythmic structures based on these characteristics (Abercrombie, 1967). Languages in which speech adheres to such rhythmic patterns are classified as syllable-timed languages, where stress patterns result in a relatively equal distribution of weight across syllables. Conversely, another rhythmic classification is characterized by stress-timed languages, where timing is segmented according to stress patterns



Yuwaree Tongvivat
yuwaree.tongvivat@sunderland.ac.uk

Sirinda Palahan
sarinda_pal@utcc.ac.th

¹ University of the Thai Chamber of Commerce, Bangkok, Thailand

² The University of Sunderland, Sunderland, United Kingdom

(Abercrombie, 1967). In these languages, unequal timing occurs between intervals—such as in the phrasal stress of the English language—where the most prominent syllabic sounds are concentrated.

Thai language is sometimes classified as exhibiting syllable-timed rhythmic patterns (Suntornsawet, 2022). This classification yet remains a subject of debate. For instance, findings by Luangthongkum (1977) suggested that Thai is a stress-timed language. It is noteworthy that this particular study is in part based on appropriate standard Thai pronunciation and old poem, in which the former can at times be more nobly, and the latter, is likely be either more musical or elegant than everyday spoken language. Studies by Romano et al. (2011) and Tingsabadh & Abramson (1999) have also suggested that Thai aligns more closely with stress-timed languages, based on measurements obtained using the Pairwise Variability Index (PVI). Similarly, Arvaniti (2012) also found that Thai is part of the stress-timed category based on the nPVI measure. In contrast, the language is found to be part of the syllable-time language based on the Normalized Pairwise Variability (nPVI), percent vowel (%V) and delta consonant (ΔC), measure by Grabe and Low (2002). These divergent findings raise critical empirical questions regarding the applicability of rhythmic classification across languages (Arvanity, 2012). As Arvaniti (2012) observes, various factors—including elicitation methods, sentence types, effect sizes, and the metrics employed—may contribute to these discrepancies. Consequently, the choice of metrics for measuring acoustic features is essential, explored further in Section 2.

Considering the phonological properties intrinsic to Thai pronunciation, the rule-based system governing Thai sounds may indeed align with the categorical features associated with syllable-timed rhythmic classes. Prominent sounds in Thai occur in almost every monosyllabic word and in the final syllable of disyllabic and polysyllabic words (Warotamasikkkhadit, 2000). Final syllables of Thai are also evidenced to be longer than other syllables in the time measurement by Luangthongkum's (1997) study, suggesting that Thai pronunciation is characterized by predefined stressed syllables. Given this consistent assignment of stress, the Thai sound system can be, at least partially, categorized as syllable-timed, especially when considering the “tempo” when prominent sounds occur.

The rhythmic characteristics of the Thai language have significant implications for the pronunciation of English by Thai speakers, particularly in the context of second language acquisition. When a second language is acquired after the critical period, the likelihood of retaining a non-native accent increase, especially when the rhythmic classes of the two languages differ markedly (Brown & Chang, 2023, pp. 4-5). Consequently, English speech produced by Thai speakers, who inherently transfer Thai rhythmic features into their English pronunciation, is less likely to resemble natural English.

Considering the following Thai sentence ทานกาแฟกันไหม (tān kāfāe kan mǎi), where the speech rhythm can be represented as follows: |ทา|น|กา|แฟ|กัน|ไ|ม|. In this notation, the symbol “|” indicates a word boundary, and the texts in an italic font inside word boundaries mark the consonant(s) and associated vowel(s) where the syllable stress is located. A comparable English sentence with a similar meaning is, “Would you like to have a cup of coffee?” The word boundaries and locations of stress in this sentence can be identified as |*Would*|*you*|*like*|to have|a *cup*|of *coffee*|. The word *coffee* (noun) is pronounced /kɒːfɪ/ in English, with the primary stress on the first syllable. In contrast, Thai speakers of English often pronounce it as /kɒːfɪː/, placing stress on the second syllable. This discrepancy arises because, in Thai, stress is typically placed on the final syllable of disyllabic words, especially in nouns (Warotamasikkkhadit, 2000), which contrasts with English stress patterns. When examining stress syllables at the sentence level, the placement of prominence varies depending on the word class of each lexical unit. For instance, in the sentence |naa0 li3 kaa0 [the alarm clock] (noun) | pluk1 (verb) [wake] | chan4 [me] (object pronoun) |, the stress syllables differ from those in the same words when they function as a compound noun: |naa0 li3 kaa0 pluk1 (noun) [alarm clock] | chan4 (possessive pronoun) [my] |.

The rule-based rhythmic pattern of Thai suggests characteristics akin to those of syllable-timed languages, with subtle variations in the prominence of syllables in function words. This bears some resemblance to the treatment of unstressed grammatical words in English. The malleable nature of Thai rhythmic patterns, particularly when transferred into English speech, may influence the interplay among duration, loudness, intensity, and pitch—key variables in the measurement of rhythmic patterns, as assessed by various metrics (Grabe & Low, 2002; Keating et al., 2004; White, 2018), further elaborated in Section 2.

This study seeks to examine these rhythmic features as manifested in a sampled English sentence. The primary objective is to contribute to the analysis of metrics that effectively measure the differences in the rhythm of English produced by native speakers and those acquiring English as a second language. Speech samples from American and Thai speakers were analyzed. To identify the contingent features, we measured the maxD and manually

annotated timing of beats as acoustic landmarks, as well as automatically extracted features including syllable durations, vowel durations, consonant durations, intensity levels, speech rates, pitch characteristics, and normalized pairwise variability indices.

This research offers three primary contributions:

1. Validation of maxD as an efficient tool for algorithmic rhythm detection.
2. Characterization of distinctive rhythmic features in native versus non-native speech.
3. Development of an automated approach for rhythmic analysis of large datasets.

This study addresses two primary research questions:

1. Can the algorithmically determined maxD parameter serve as a viable alternative to manually annotated beat locations, particularly when comparing Thai and American speakers of English?
2. What differences in acoustic features, as evaluated by statistical measures and manual annotations, can be observed when comparing the sampled speech sounds of American speakers with those of Thai speakers of English? We refer to the differences in acoustic features, especially duration, intensity, loudness/intensity of volume, within each syllable that contributes to the rhythm and tempo of speech sounds produced by the two groups of speakers.

The ultimate goal of this research study is to identify rhythmic features that distinguish native and non-native English speakers. The first research question supports this goal by exploring whether the maxD parameter can be used as a reliable and scalable tool to identify rhythmic features, providing a practical foundation for further analysis. The second research question builds upon this by identifying specific acoustic differences in rhythm between native and non-native speakers, which are critical for understanding and addressing speech rhythm disparities. Together, these questions provide the basis for comparing non-native speech with native speech and developing targeted recommendations to improve speech rhythms based on these differences. The automated nature of the analysis makes it particularly suitable for implementation in real-world applications and large-scale studies.

1. Rhythm of Thai and English

Both English and Thai languages exhibit distinct rhythmic patterns, with the primary differences arising from how stressed syllables and lexical units are articulated. English is often classified as a stress-timed language, based on the concept of the (relative) “uniformity” in the time lapse between the onsets of prominent syllables (Pike, 1945, as cited in Arvaniti, 2012, p. 351). Some key features of stress-timed rhythm include a higher degree of vowel reduction and greater consonantal variability compared to syllable-timed languages (Grabe & Low, 2002; Patel & Daniele, 2002). The classification of Thai language as a syllable-timed language is not universally accepted (cf. Romano et al., 2011), the prominence typically assigned to nearly all monosyllabic words and to the final syllables of polysyllabic words, regardless of word class (cf. Warotamasikhhadit, 2000), suggests a greater emphasis on length, duration, and intensity in specific syllables—potentially attributes commonly associated with syllable-timed languages.

In English, on the other hand, stressed syllables are assigned based on word class (Roach, 2009), with the notion of “strong form” and “weak form.” Strong form refers to the content words that convey more meaningful elements for communications (ibid.). Weak form, in general occurrences, are unstressed, which refers to words that serve as grammatical functions (e.g. articles, prepositions, pronouns) (ibid.). These words, however, can be stressed when communicating discourse functions contribute to the intended meanings. For example, the differences between “She is *my* wife.” and “She *is* my wife.” is the focused intention. The former puts an emphasis on a possession, while the latter is focused on the status as the wife. On a general note, without specific discourse or pragmatic functions, the primary and secondary syllable stress serves as tempo contributing to beats and, on a large scale, the rhythmic structure of a sentence. In specifically interpreting attitudinal and intentional meanings in discourse, three established schools of thought can be taken into consideration (Wennerstorm (2001). Firstly, tones and prominence in intonation can suggest meanings and attitudes of speakers (Brazil, 1997). Secondly, interlocutors’ tune of voice comprising pitch accents, phrase accents and boundary tones can suggest attitudinal and intentional meanings (Pierrehumbert & Hirschberg (1990). Lastly, discourse intonation and systemic functional phonology incorporates the interaction among the choices of lexico-grammatical, discourse intonation and phonological structures in the interpretation of meaning-making (Halliday, 1967). Recent studies by Kang

(2013a) and Kang & Johnson (2021) found that tones, pitch and word stress contribute to the identification of English proficiency of non-native speakers and native speakers. Specifically, low and falling tones are frequently found in utterances of non-native speakers, while rising and level tones more likely appear in the utterances of native speakers. Although not necessarily causal among the choices of lexicons, grammatical structures and tones, more complex word and grammatical structures are more commonly found in speech produced by speakers with higher English proficiency (*ibid.*). These differentiations in part align with Brazil's (1997) intonational discourse, Pierrehumbert & Hirschberg's (1990) intonational contour in interpreting discourse and Halliday's (1967) interrelationship among the contexts of lexicons, grammatical complexity and the phonological system in an interpretation of meanings.

While the rhythmic patterns of Thai and English have been observed, the differences in the levels of intensity and duration of sounds within stressed and unstressed lexical units that contribute to a perceived tempo or beats remain underexplored, especially when it comes to algorithmic evaluations. This study aims to identify the rhythmic patterns of English sentences as spoken by Thai speakers of English compared to native English speakers. To achieve this, we will examine the intensity and duration of sounds within lexical units to discern differences in the patterns of utterances between the two groups. The findings from this analysis could inform the development of teaching tools designed to train Thai speakers of English in stressed and unstressed patterns and sentence rhythm.

2. Identifying Rhythm

Various approaches have been employed to identify English rhythm, encapsulated within both sensorimotor and algorithmic frameworks (Rathcke et al., 2021; Rathcke & Lin, 2023; Šturm & Volín, 2016; Brown & Chang, 2023). Traditional methods rely on human perception to identify the tapping locations of prominent sub-syllabic sounds and the interval durations between these sounds. These methods include techniques for understanding rhythm in speech through foot-tapping and time measurement of utterances (Abercrombie, 1967; Luangthongkum, 1977), finger-tapping, and more recently, identifying the p-center (Šturm & Volín, 2016; Rathcke et al., 2021; Rathcke & Lin, 2023). With the ongoing advancement of statistical measures and specialized tools for acoustic analysis, several metrics have gained popularity for rhythm identification. These metrics include %V (percent vowel), ΔV (delta vowel), ΔC (delta consonant), VarcoV (variation coefficient for vowels), VarcoC (variation coefficient for consonants), nPVI-V (normalized pairwise variability index for vowels), and nPVI-C (normalized pairwise variability index for consonants) (Brown & Chang, 2023). These metrics have been utilized to quantify rhythm class distinctions, such as those between syllable-timed and stress-timed rhythms. Despite the adoption of these metrics for classifying language rhythms, discrepancies in findings derived from different measures persist (White & Malisz, 2020), prompting ongoing debates regarding the most appropriate metrics for identifying specific rhythmic patterns.

The choice of metrics for our analysis is guided by the specific research objectives, the relevance and applicability to the analysis of English rhythm, particularly in comparison between Thai and American speakers of English. To achieve this, the metrics must effectively capture the temporal and prosodic features that characterize the rhythm of each language group. The selected metrics—%V, ΔV , ΔC , VarcoV, VarcoC, nPVI-V, and nPVI-C—are particularly suited for this purpose. These metrics are widely recognized in the field for their ability to quantify rhythmic distinctions, especially between stress-timed and syllable-timed languages (Grabe & Low, 2002; Patel & Daniele, 2003).

The selected metrics are grounded in robust theoretical and empirical foundations. For instance, %V and $\Delta V/\Delta C$ provide insights into the relative proportion of vocalic and consonantal intervals, which are crucial for distinguishing between stress-timed and syllable-timed languages. The normalized pairwise variability index (nPVI) further refines this distinction by quantifying the degree of temporal variability between successive vowel or consonant intervals (Brown & Chang, 2023; Grabe & Low, 2002; Low et al., 2000). These metrics have been extensively validated in previous studies, making them reliable tools for analyzing the rhythmic properties of different languages.

Given that this study involves speakers from different linguistic backgrounds, it is essential that the selected metrics can accommodate the variability inherent in cross-linguistic comparisons. Metrics such as VarcoV and VarcoC are valuable, as they account for variations in speech rate and allow for a more nuanced analysis of rhythmic patterns (White et al., 2007). The nPVI-V and nPVI-C metrics are also well-suited for cross-linguistic

studies (Grabe & Low, 2002), as they can effectively capture the rhythmic differences between native English speakers and Thai speakers, whose languages exhibit different rhythmic typologies.

While the metrics offer significant advantages, it is important to acknowledge their limitations. For example, metrics like ΔV and ΔC may be influenced by speech rate (Barry et al., 2003), which could affect their reliability in certain contexts. Additionally, the nPVI metric, while useful for comparing rhythmic variability, may not fully capture the complexity of rhythmic patterns in languages with less rigid rhythmic structures (White & Malisz, 2020). These limitations have been considered in the design of this study, and steps have been taken to mitigate their impact, such as controlling speech rate in the analysis.

The selection of nPVI-V, and nPVI-C among others (see Section 3.2.1.) as the primary metrics for this study is justified based on their relevance to the research objectives, their theoretical and empirical foundations, and their applicability to cross-linguistic rhythm analysis (Low et al. (2000)). While these metrics have certain limitations, they are among the most reliable and widely used tools for quantifying rhythmic patterns in spoken language. Their application in this study is expected to yield valuable information into the rhythmic differences between native English speakers and Thai speakers of English, contributing to the broader understanding of language rhythm and its implications for language learning and pedagogy.

Recently, the algorithm-based (“maxD”) metric has demonstrated promising results in identifying the locations of beats in speech. Šturm and Volín (2016) explored this metric in their study on the synchronization of spoken Czech words with metronome beats, particularly focusing on the identification of perceptual centers in syllables, known as p-centers. In their experiment, participants synchronized their tapping with the rhythm of 37 disyllabic Czech words at two different speeds. The study found that p-centers tended to align more closely with the onset of the first vowel and initial consonants. Additionally, the findings highlighted that maxD—representing the moment of the fastest energy increase—provided a reliable cue for synchronization. This insight underscores the potential of algorithm-based approaches in enhancing our understanding of speech rhythm synchronization, suggesting a significant role for such methods in acoustic studies and speech technology.

Expanding on the research involving maxD and synchronization, Rathcke et al. (2021) employed two distinct finger-tapping tasks—Synchronization and Non-Synchronized Motor Reproduction (NMR)—to assess sensorimotor synchronization (SMS) with English sentences. Their study demonstrated effective rhythmic entrainment during synchronization tasks, where participants tapped in time with vowel onsets. Conversely, the NMR task revealed that participants reverted to their intrinsic tapping rates when reproducing rhythms after listening, highlighting notable differences in rhythmic perception in the absence of auditory feedback. This comparative analysis emphasizes the value of SMS paradigms in examining the dynamics of rhythmic synchronization in linguistic contexts, providing a broader perspective on how rhythmic perception varies across different tasks.

Building on this earlier work, which suggested the advantages of movement-based paradigms for understanding speech rhythm, Rathcke and Lin (2023) further explored the influence of linguistic and acoustic structures on rhythmic movement. Their study involved a finger-tapping task synchronized with English sentences, focusing on the role of various acoustic landmarks, such as syllable onsets and stress markers, in rhythm synchronization. The results confirmed that both manually annotated beats and algorithmically determined moments of rapid energy changes (maxD) effectively predict synchronization peaks, validating their function as robust perceptual anchors in speech rhythm. Unlike other traditional metrics including %V, ΔV , ΔC , VarcoV, VarcoC, nPVI-V, and nPVI-C that are algorithmic-based, maxD effectively predict synchronization peaks that have been tested between the manually defined beat locations and algorithmically determined moments of rapid energy changes or maxD, this parameter serves as a good potential for identifying beats in this present study.

Drawing on the findings from these studies, the first research question emerges: Can the algorithmically determined maxD parameter serve as a viable alternative to manually annotated beat locations, particularly when comparing Thai and American speakers of English? Along with this research question, this study additionally seeks to explore what statistical measures can be used to identify prominent features of sounds at the syllabic level and what these measures reveal about the representation of rhythmic features in the sampled speech sounds produced by American and Thai speakers of English. If maxD effectively predicts beats in language, the metric can be used for algorithmically identifying the length of time and timing for utterances at the syllable, word and sentence levels (both segmental and suprasegmental levels). By extension, this metric, together with assistance of a large language model, could potentially be used to create a tool that provides guide directions and automatic

feedback to train language learners for the way in which specific rhythm of speech could be pronounced. In doing so, the metric would help contextualize the rhythm theory, in the sense that the time and timing of beats in speech in which the first language and the target language differ (e.g., time and timing of syllable-timed versus stress-timed languages) could be identified and the duration of time and timing when beats as prominent sounds should occur in [supposed] pronunciation.

3. Native and Non-native Speaker Classification using Speech Rhythm

The classification of native and non-native speakers based on speech rhythm has been an area of active research, aiming to enhance the understanding and automation of language proficiency assessments. By examining various prosodic features such as duration, pitch, and intensity, researchers have developed models to distinguish between native and non-native speech patterns. This section reviews relevant literature, focusing on methodologies, features used, and their implications for improving natural-sounding speech in non-native speakers.

Chen et al. (2018) developed the SpeechRaterSM v. 5.0 engine to evaluate non-native English speech using Automatic Speech Recognition (ASR) and Natural Language Processing (NLP) technologies. The system employs prosodic features like duration-based metrics and rhythm metrics derived from segmental and syllabic durations to capture speech rhythm, along with intensity and pitch features. These features help distinguish native from non-native speech, highlighting temporal variability and prosodic patterns.

Research by Lai et al. (2013) and White et al. (2014) further demonstrates the effectiveness of rhythm metrics, such as nPVI, VarcoV, and rPVI_{syl} in capturing temporal patterns and variability in speech, essential for accurate classification. They focused on models for foreign accent recognition using both native and non-native speech samples, showing that these rhythm metrics could effectively distinguish between native and non-native speakers.

Lesnichaia et al. (2022) explored the use of convolutional neural networks (CNNs) to classify accented English speech from Germanic, Romance, and Slavic language families. The study utilized amplitude mel-spectrograms on a linear scale, which were found to be more effective for accent classification compared to traditional features like Mel-Frequency Cepstral Coefficients (MFCCs). The CNNs model, which included convolutional layers with ReLU activation, batch normalization, and pooling layers, achieved high accuracy (0.964 to 0.987) on data from the Speech Accent Archive.

Polyanskaya, Ordin, and Busa (2017) investigated how speech rhythm and speech rate independently contribute to perceived foreign accents in second language (L2) learners. The study used a resynthesis technique to create stimuli that isolated differences in speech rhythm and speech rate while controlling segmental and tonal idiosyncrasies. Native English speakers rated these stimuli for foreign accents, showing that both rhythm and rate influence perceived accent, with rhythm having a more significant impact.

Taken together, the reviewed studies have highlighted the critical role of rhythmic features, such as duration, pitch, and intensity, in distinguishing between native and non-native speakers. This leads to the second research question: what are the rhythmic differences between Thai and American English speakers? To answer this research question, we examined the differences, at the acoustic level, evaluated by statistical measures and manual annotations, when comparing sampled speech sounds produced by American speakers with those of Thai speakers of English? Identifying these specific rhythmic characteristics can help non-native speakers improve their speech patterns to achieve more natural-sounding speech.

3. Methodology

3.1. Predicting Beat Locations using maxD

The initial step in the methodology involves predicting beat locations using the maxD parameter, which identifies moments of rapid energy change in speech, typically corresponding to stress peaks. To evaluate the alignment of maxD locations with theoretical rhythmic beats (hereafter referred to as “beats” or “beat location(s)”), a comparative analysis will be conducted between these algorithmically determined points and the onset of vowels, or nuclei, within the metrical structure of sentences. These theoretical beats, which represent the natural rhythm of speech, will be used as a benchmark to assess the accuracy of maxD in predicting manually annotated beats. For this analysis, we employed the Montreal Forced Aligner (MFA) by McAuliffe et al. (2017).

3.1.1. Data Collection

Speech samples were collected from the Speech Accent Archive (Weinberger, 2015), which has been established to systematically document a wide range of speech accents from various language backgrounds. In this archive, native and non-native speakers of English read the same standardized English paragraph, with recordings made under controlled conditions. For the purposes of this study, only the final sentence of the elicitation paragraph was analyzed: “We will go meet her Wednesday at the train station.” The analysis was restricted to recordings in which speakers produced the sentence with exactly 12 syllables. A total of 17 recordings from Thai speakers of English met this criterion, and an equal number of recordings from American native speakers were selected for comparison. Thus, the study analyzed 34 audio files in total, comprising 17 native English speakers from the United States and 17 non-native English speakers from Thailand. It is important to note that this analysis was not exclusively focused on the American accent; future research will aim to include a broader range of English accents. We also acknowledge that the analysis is based on acoustic features derived from one sentence. Future explorations will include a more variety of sentences.

3.1.2. Identifying Beat Locations by maxD

Before proceeding with the analysis, it was essential to extract the necessary features from the collected data. This process involved identifying the maxD locations, which correspond to moments of rapid energy change, and comparing them to the manually annotated beats that represent the rhythmic structure of the speech. The feature extraction process was carried out in several detailed steps:

The procedure for calculating the local time point of maxD was adapted from Šturm and Volín (2016). Initially, a raw energy contour for each sentence was generated by squaring the raw amplitudes, using a 40-ms window and a 1-ms shift. This raw energy contour was then smoothed by applying a 6th-order moving average filter. Following this, a raw energy-difference contour was created by calculating the difference between two adjacent samples of the smoothed energy contour ($E_T - E_{T-1}$) which was further smoothed using a 10th-order moving average filter Rathcke & Lin (2023, p. 6). For each syllable, the maxD location was identified as the highest local value within the smoothed energy-difference function. One maxD location was extracted per syllable, representing the point of maximal energy increase (see Fig. 1).

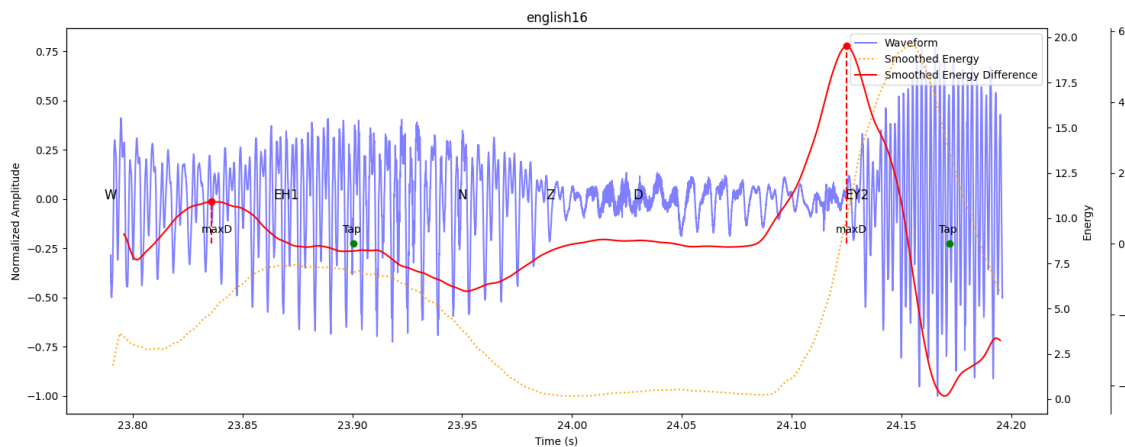


Fig. 1 maxD locations and taps within the word *Wednesday* spoken by an American speaker

3.1.3. Extraction of Manually Annotated Beat Locations

The manual annotation of sounds involved pinpointing the precise timing at which intensity (of volume) occurred within the duration of each syllable. Given that the elements contributing to rhythm in phrases or sentences—such as intensity of volume, duration, and pitch—are integral to stressed and unstressed syllables (White, 2018), the mathematical duration of time when the high intensity of volume occurs within each syllable was identified and designated as “beat(s)” or “beat location(s).”

The beat was annotated using ELAN (ELAN (version 6.8), 2024), where the location of beats within syllables was identified and annotated as a time range. These annotations were then exported as JSON files and aligned with beat predictions generated by maxD and other metrics using the Montreal Forced Aligner (MFA) (McAuliffe et al., 2017) and Python version 3.7 (Python Software Foundation, 2021). The annotation process was conducted by a linguist with 15 years of music training and practice (the corresponding author). The annotation approach of

identifying beats differs from conventional sensory-motor finger-tapping techniques. Our method allowed the annotator to listen repeatedly to the points where the intensity of volume and its duration occur, ensuring as precise a recording of the beats as possible. We could record the beats as a duration of time (in millisecond).

This approach to annotation aligns with the concept of the p-center, which relates to human perception of the moment a syllable occurs (Šturm & Volín, 2016). In our annotations, the timing began from the onset of the syllable, marked by the high intensity of volume and duration, and continued until the final part of the sound with high intensity of volume occurs within each syllable. In other words, the loudest sound duration within each syllable was annotated. The annotated duration of time was then used both to compare with the maxD locations and to analyze the duration of each syllable produced by the American and Thai speakers. The annotated time range is defined as a “tap” in our analysis (see Fig. 1). The taps, approximated by the vowel onsets or nuclei within the speech's metrical structure and extracted manually. The average duration of time in taps within each syllable was calculated, resulting in a single number representing the average mathematical time of each tap, served as a reference point for evaluating the predictive capability of maxD.

3.1.4. Regression Analysis

A linear regression model was employed to investigate the relationship between the maxD locations (independent variable) and the manually annotated beat locations (dependent variable). The primary objective of the model was to predict the timing of beats based on the identified maxD locations. The regression analysis involved fitting a linear model to the data and calculating the coefficients to assess the strength and nature of the relationship between maxD and beats. To evaluate the model accuracy and generalizability, cross-validation techniques were implemented, where the data was divided into training and testing sets to ensure the model's performance was assessed on unseen data.

Metrics such as Mean Absolute Percentage Error (MAPE) and Root Mean Squared Error (RMSE) were computed to quantify the prediction errors. The model's predictions were then compared to the actual manually annotated beat locations, and the degree of alignment was analyzed. Additionally, statistical significance tests, including t-tests for the regression coefficients, were conducted to determine the reliability and robustness of the predictors.

3.2. Classification Models using Rhythmic Features

The second phase in the methodology involves the development of a classification model to address the research question: “What differences in acoustic features, as evaluated by statistical measures and manual annotations, can be observed when comparing the sampled speech sounds of American speakers with those of Thai speakers of English?” This phase focuses on using rhythmic features extracted from speech data to differentiate between the two groups. Rhythmic features, which capture the temporal and stress patterns in speech, are critical in distinguishing between different language speakers and their unique prosodic characteristics.

Previous research in machine learning identified pitch, intensity and duration as features classifying prominent syllables in English speech (Johnson & Kang (2015)). This present study includes other features to classify and differentiate between Thai and American English speakers based on their rhythmic patterns, various rhythmic features, including maxD locations, duration variability, and other prosodic markers, will be utilized. The analysis of these features aims to uncover distinctive rhythmic signatures that characterize Thai and American English speech.

In the following sections, the process of data preprocessing, the selection of rhythmic features, the classification models used in the experiments, and the techniques used to evaluate its performance will be described. This approach will identify the key rhythmic features that differentiate Thai and American English speakers and assess the effectiveness of these features in a classification context.

3.2.1. Feature Extractions of Sounds within and between Syllables

The following rhythmic features were extracted for the classification model:

- **Max and Min of Syllable Durations:** The maximum and minimum durations of syllables within each sentence.
- **Max and Min of Vowel Durations:** The maximum and minimum durations of vowels within each sentence.

- **Max and Min of Consonant Durations:** The maximum and minimum durations of consonants within each sentence.
- **Max and Min of Syllable Intensity:** The maximum and minimum intensity levels of syllables within each sentence.
- **Syllable Speech Rate:** The rate of syllable production within each sentence.
- **Word Speech Rate:** The rate of word production within each sentence.
- **Pitch Range:** The difference between the highest and lowest pitch values within each sentence.
- **Pitch Mean:** The average pitch value within each sentence.
- **Pitch Variation:** The variation in pitch values within each sentence.
- **Pitch Max:** The maximum pitch value within each sentence.
- **maxD Location:** The location of the maxD within the sentence. This automated measure is intended to serve as an alternative to manual annotations for identifying beat locations.
- **maxD Value:** The value of the maxD within the sentence.
- **Syllable Duration Ratio:** The ratio of the maximum to minimum syllable durations.
- **Vowel Duration Ratio:** The ratio of the maximum to minimum vowel durations.
- **Consonant Duration Ratio:** The ratio of the maximum to minimum consonant durations.
- **Intensity Ratio:** The ratio of the maximum to minimum intensity levels.
- **Pitch Ratio:** The ratio of the maximum to minimum pitch values.
- **nPVI-V for Vowels and Consonants:** The normalized Pairwise Variability Index for vowels and consonants, capturing the variability in duration.

All features were automatically extracted using computational methods. The measurements were processed through a combination of tools:

- Basic acoustic features (duration, intensity, pitch) were extracted using Praat scripts
- maxD parameters were calculated using the algorithm described in Section 3.1.2
- Speech rate features were computed automatically from the forced alignment output
- Ratio calculations were derived programmatically from the primary acoustic measurements
- nPVI measurements were generated through automated scripts

The selection of duration, intensity, and pitch as features for distinguishing rhythmic patterns in speech is well-supported in the literature. As White (2018) highlighted, prosody—particularly the modulation of pitch and timing—plays a crucial role in conveying linguistic meaning, turn-taking, and expressing physical and emotional states. These elements are vital for understanding and segmenting speech. Furthermore, the inclusion of speech rate as a distinguishing feature is underscored by research that highlights its significance. Polyanskaya and Ordin (2019) emphasize that speech rate has a substantial impact on the perception and assessment of pronunciation in a second language, noting that variations in speaking rate can alter rhythmic patterns, making it an essential factor in analyzing and differentiating languages.

The use of nPVI-V for vowels and consonants, as proposed by Low et al. (2000), captures the variability of interval durations, aiding in the differentiation of rhythmic properties across languages. Additionally, maxD parameters (location, relative location, and value per sentence) were included as they represent moments of rapid energy change, indicative of stress peaks, and capture the dynamic aspects of speech rhythm and stress patterns. By incorporating these well-established features, our study aims to provide a comprehensive analysis of the rhythmic differences between Thai and American English speakers, contributing to a deeper understanding of language-specific rhythmic patterns.

3.2.2. Classification Models

To address the research question concerning the rhythmic differences between Thai and American English speakers, classification models were developed using the rhythmic features discussed in the previous section. These models were designed to differentiate between the two speaker groups based on the extracted features. The following classifiers were employed in the construction of these models:

- **Support Vector Machine (SVM):** SVM provides outstanding performance in optimizing speech features (Hansen et al., 2014; Louni et al., 2024; Nainan & Kulkarni, 2020). Both linear and Radial Basis Function (RBF) kernels were utilized. The linear kernel was chosen to identify the optimal hyperplane

that separates the classes with maximum margin, while the RBF kernel was applied to capture non-linear relationships within the data.

- **Random Forest (RF):** An ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes from individual trees. This approach is effective in capturing non-linear relationships and feature interactions.
- **Gradient Boosting (GB):** An ensemble technique that sequentially builds models, each iteration improving performance by focusing on the errors of previous models.
- **Logistic Regression (LR):** Both original and polynomial features were employed to capture linear and non-linear relationships. This linear model offers probabilities for class membership, facilitating the interpretation of feature coefficients.
- **XGBoost:** An optimized gradient boosting algorithm that constructs models sequentially, with each new model correcting the errors of its predecessors. XGBoost is renowned for its speed and performance. Adopting this model is evidenced to yield excellent results in differentiating speech characteristics (Ashwini & Bharathi, 2024).

These models were selected for their distinct strengths in managing complex classification tasks. SVM is particularly effective for maximizing the margin between classes, making it suitable for binary classification. Random Forest excels at capturing non-linear relationships and interactions among features, thereby enhancing robustness and generalization. Gradient Boosting and XGBoost were chosen for their high predictive power and iterative improvement of model accuracy. Logistic Regression, with its straightforward and interpretable nature, allows for an understanding of the impact of individual features on classification outcomes.

While Lesnichaia et al. (2022) explored the use of convolutional neural networks (CNNs) to classify accented English speech from Germanic, Romance, and Slavic language families, this study opts for simpler models due to their explainability. The focus on simpler models enables a clearer interpretation of the rhythmic features contributing to the differentiation between Thai and American English speakers, thereby aiding non-native speakers in understanding and improving their speech patterns.

3.2.3. Model Validation and Performance Comparison

All models underwent a rigorous training and validation process to ensure robustness and generalizability. The dataset was first divided into training and testing sets using an 80/20 split. The training set (80% of data) was then subjected to 3-fold cross-validation, where it was divided into three equal subsets. Each model was trained on two subsets while being validated on the third, with this process repeated three times so that each subset served as validation data once. The held-out test set (20% of data) was used only for final model evaluation. By employing these classifiers and validation techniques, the models were designed to accurately classify and differentiate between Thai and American English speakers based on their rhythmic features.

To evaluate the effectiveness of the models in classifying speaker rhythms, several performance metrics were utilized. This included accuracy, which measures the overall correctness of the model; precision, which evaluates the model's ability to correctly identify positive instances; recall, which assesses the model's capacity to capture all relevant positive instances; and F1-score, which provides a harmonic mean of precision and recall, offering a balanced metric to evaluate both aspects. These metrics provided a comprehensive evaluation of each model's performance, ensuring that the chosen features and classification approaches effectively distinguished between the rhythmic patterns of Thai and American English speakers.

3.3. Analysis of the Location Patterns of maxD, Pitch and Intensity

The analysis of the location patterns of alignments among maxD, pitch, and intensity was conducted, with a focus on identifying and interpreting these patterns within the framework of acoustic phonology. The investigation aimed to uncover the spatial and temporal relationships between moments of rapid energy change maxD, pitch peaks, and intensity levels within syllables. By meticulously examining these elements, we sought to determine how their occurrences align with established theories, which posit that such features play a critical role in shaping the rhythmic and prosodic structure of speech. The manual analysis allowed for a nuanced understanding of the interplay between these acoustic features, revealing patterns that may reflect the underlying or to some extent

explaining the similarities and differences between the two groups of speakers in our samples. In doing so, the findings could contribute to a deeper understanding of how maxD, pitch, and intensity interact to create perceptually salient rhythmic and prosodic patterns in language.

3.4. Computational Requirements and Scalability

The proposed approach demonstrates strong potential for scalability due to its automated feature extraction pipeline. All rhythmic features, including the maxD parameter, intensity measurements, and duration calculations, are extracted automatically without manual intervention, making the method suitable for processing large speech datasets. The selected classification models (SVM, Random Forest, Gradient Boosting, and Logistic Regression) are also well-established for handling high-dimensional feature sets efficiently.

While the current study utilized a focused dataset of 34 speakers to establish methodological validity, the automated nature of the analysis suggests strong potential for scaling to larger datasets. However, systematic evaluation of processing times and computational resource usage with larger datasets was not conducted in this study. Future work will include comprehensive benchmarking of computational performance and scalability testing with expanded datasets, measuring processing times for feature extraction and model execution across different data volumes and hardware configurations. These measurements will provide valuable insights for implementing the proposed methods in large-scale speech analysis systems and interactive language learning tools.

4. Experimental Results and Discussions

This section presents the findings from the experimental analyses, followed by a comprehensive discussion of the results in relation to existing theories. The experimental outcomes are analyzed to evaluate the effectiveness of the methodologies employed, and the implications of these findings are explored in the context of language rhythm studies. By critically examining the data, we aim to elucidate the key patterns and relationships that emerged, offering insights into the rhythmic differences between Thai and American English speakers and their broader linguistic significance.

4.1. Predicting Beat Locations

In this section, the outcomes of the regression model predicting manually annotated beats using the maxD parameter are presented. The regression model, designed to predict manually annotated beats based on maxD locations, provides the key findings of the analysis.

4.1.1. Regression Model Results

Three linear regression models were employed to predict beats using the maxD locations. The dataset was split into training and testing sets with an 80/20 ratio to ensure robust evaluation of the model's performance. The models were created as follows:

Model 1: maxD and onsets of Thai speakers

Model 2: maxD and onsets of English speakers

Model 3: maxD and onsets of both Thai and English speakers combined

The linear regression models fitted to our data are represented by the following equations:

Model 1: maxD and onsets of Thai speakers

Coefficients: Intercept = -2.4039, Slope = 1.1010

Model 2: maxD and onsets of English speakers

Coefficients: Intercept = -0.1125, Slope = 1.0082

Model 3: maxD and onsets of both Thai and English speakers combined

Coefficients: Intercept = -1.5254, Slope = 1.0734

The coefficients of each model indicate a strong predictive relationship between maxD and manually annotated beats. For Thai speakers, Model 1 shows a significant positive relationship with a slope of 1.1010, suggesting that maxD is a reliable predictor of manually annotated beats, albeit with an adjustment factor reflected in the negative intercept of -2.4039. Model 2, for English speakers, reveals a near one-to-one relationship with a slope of 1.0082 and an intercept close to zero (-0.1125), indicating minimal adjustment needed for this group. Model 3, combining both Thai and English speakers, demonstrates a general positive relationship with a slope of 1.0734 and an intercept of -1.5254, indicating an average adjustment across both groups.

The slopes of the regression models also provide insight into the correlation between maxD and manually annotated beats. All models show slopes close to 1, indicating a strong positive relationship. Model 2 (English speakers) demonstrates a nearly one-to-one relationship with a slope of 1.0082, suggesting maxD locations align very closely with manually annotated beats for native speakers. Model 1 (Thai speakers) shows a slightly steeper slope of 1.1010, indicating a tendency for maxD to predict slightly later beat locations compared to manual annotations. Model 3's slope of 1.0734 reflects the combined pattern across both speaker groups. These slope values, all being close to 1, indicate strong positive correlations between maxD and manually annotated beats, supporting maxD's effectiveness as a predictor of beat locations.

The performance of these models was further assessed using Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE) on the test data, along with the p -values of the coefficients to determine their statistical significance.

	RMSE	MAPE	p -values
Model 1	1.6348	2.7213%	0.0000*
Model 2	0.1182	0.3290%	0.0000*
Model 3	0.8879	1.4937%	0.0000*

Table 1 Performance of model 1, 2 and 3 when accessed by RMSE and MAPE

As illustrated in Table 1, The RMSE and MAPE values indicate the model's prediction accuracy, with lower values suggesting better performance. Model 2, which pertains to English speakers, demonstrated the lowest RMSE (0.1182) and MAPE (0.3290%), indicating highly accurate predictions. Model 3 also performed well, with an RMSE of 0.8879 and a MAPE of 1.4937%. Model 1 showed higher error metrics (RMSE: 1.6348, MAPE: 2.7213%) compared to the other models, reflecting greater variability in predicting beat locations for Thai speakers. The p -values for the coefficients in all three models were extremely low ($p < 0.001$), indicating a statistically significant relationship between maxD and manually annotated beats, (exact p -values below 0.001 are reported as 0.000* in the table). These results validate the effectiveness of the maxD parameter in predicting manually annotated beats and underscore its potential application in distinguishing the rhythmic patterns of native and non-native speakers.

4.1.2. Conclusion for Research Question 1

The experiments addressed the research question: can the maxD parameter, determined algorithmically, serve as an effective alternative to manually annotated beat locations, particularly when comparing Thai and American speakers? Through the analysis, strong evidence for the effectiveness of maxD in predicting beat locations was demonstrated. Model 2 (English speakers) showed the highest accuracy (RMSE: 0.1182, MAPE: 0.3290%), indicating exceptional precision in predicting native speaker beat patterns. Model 3, which combined both Thai and English speakers, also demonstrated robust performance (RMSE: 0.8879, MAPE: 1.4937%), validating the general applicability of maxD across both speaker groups. Although Model 1 for Thai speakers showed higher error metrics, significant predictive capability was maintained.

The low p -values across all models confirmed the statistical significance of maxD in predicting manually annotated beats. Through these findings, maxD has been validated as an efficient tool for phonetic research, capable of distinguishing rhythmic patterns between Thai and American English speakers. By identifying beat locations with a degree of reliability, a practical foundation for analyzing rhythmic differences has been established, supporting the goal of identifying features that can guide non-native speakers in improving their speech rhythms.

The validation of maxD as an effective tool for automated rhythm analysis contributes to both theoretical understanding of speech rhythm and practical applications in language teaching. These findings support the development of targeted pronunciation training tools and automated language-learning applications focused on enhancing naturalness and fluency in second language acquisition.

4.2. Classification Models using Rhythmic Features

Classifying native and non-native speakers based on their rhythmic features is essential for improving the naturalness of speech in second language learners. This section presents the results of various classification models used to distinguish between Thai and American English speakers, with a focus on the rhythmic characteristics of their speech. The following subsections provide a detailed analysis of the classification model outcomes, including an examination of misclassifications to better understand the challenges in distinguishing near-native speakers. Additionally, the importance of various rhythmic features is explored using SHAP (SHapley Additive exPlanations) values, which shed light on the contributions of specific features to the model's decisions. This approach underscores the capabilities of machine learning models in speech classification and identifies potential areas for improvement, ultimately supporting non-native speakers in achieving more natural-sounding speech.

4.2.1. Classification Results

This section presents the results of the classification models used to distinguish between Thai and American English speakers based on their rhythmic features. The performance of each model was evaluated using multiple metrics: accuracy to measure overall correctness, precision to assess true positive rate, recall evaluating the model's ability to identify all relevant cases, and F1-score to provide a balanced measure between precision and recall. These metrics collectively provide a comprehensive assessment of model performance across different aspects of classification. The results are summarized in Table 2 below.

Model	Accuracy	F1	Precision	Recall
Linear SVM	0.9091	0.9231	0.8571	1.0000
Non-linear SVM	1.0000	1.0000	1.0000	1.0000
XGBoost	0.8182	0.8571	0.7500	1.0000
Random Forest	1.0000	1.0000	1.0000	1.0000
Gradient Boosting	0.9091	0.9231	0.8571	1.0000
Linear Logistic Regression	0.9091	0.9231	0.8571	1.0000
Non-linear Logistic Regression	1.0000	1.0000	1.0000	1.0000

Table 2 Classification models for rhythmic feature differentiation in speech produced by Thai and American speakers

The classification results reveal that several models performed exceptionally well in distinguishing between Thai and American English speakers based on their rhythmic features. The Non-linear Logistic Regression, Random Forest, and Non-linear SVM models achieved perfect scores across all metrics (accuracy: 1.0000, F1-score: 1.0000, precision: 1.0000, recall: 1.0000). This balanced performance across metrics suggests these models are robust in both identifying native speakers correctly (precision) and capturing all native speakers in the dataset (recall). However, these perfect scores should be interpreted with caution due to the relatively small sample size (34 speakers total), which may have led to overfitting despite cross-validation efforts.

The Linear SVM, Gradient Boosting, and Linear Logistic Regression models demonstrated strong but slightly lower performance (accuracy: 0.9091, F1-score: 0.9231, precision: 0.8571, recall: 1.0000). The perfect recall but lower precision indicates these models successfully identified all native speakers but occasionally misclassified non-native speakers as native speakers. This pattern suggests these models might be more generous in classifying speech as native-like, which could be advantageous in identifying highly proficient non-native speakers but may require additional validation in practical applications.

The XGBoost model showed relatively lower but still strong performance (accuracy: 0.8182, F1 -score: 0.8571, precision: 0.7500, recall: 1.0000). The lower precision with perfect recall again indicates a tendency to classify some non-native speakers as native speakers, suggesting this model might be most suitable for applications where identifying all potential native-like speech patterns is prioritized over strict classification.

4.2.2. Error Analysis

The classification models demonstrated robust performance in distinguishing between native and non-native speakers, with all models achieving a recall of 1.0000 for native speaker identification. However, examination of false positives revealed important insights about model limitations. Most models, except XGBoost, produced one consistent false positive: a Thai speaker who began learning English at age 6. The XGBoost model identified an additional false positive: a Thai speaker with 41 years of English residency. These misclassifications highlight the models' difficulty in cases of early language acquisition and extended English exposure, suggesting that both factors can lead to native-like rhythmic patterns.

The pattern of misclassifications points to specific limitations in the current approach. The models particularly struggled with speakers who either acquired English during the critical period or had extensive exposure to the language environment. This suggests that while rhythmic features effectively distinguish most native and non-native speakers, they may not capture the subtle differences present in highly proficient non-native speakers who have achieved native-like rhythm through early acquisition or long-term immersion. These findings underscore both the power of rhythmic features as classification indicators and their limitations in exceptional cases.

The high performance of these models supports the existence of specific rhythmic patterns in English speech that non-native speakers can successfully adopt. However, the misclassifications of highly proficient speakers suggest that these patterns can be acquired through various paths, including early age of onset and long-term language exposure. While the current models demonstrate strong overall performance, future refinements could focus on developing more nuanced approaches for handling the cases of advanced language proficiency and incorporating features that better capture subtle rhythmic variations.

4.2.3. Feature Importance Analysis

In this study, the best-performing models for distinguishing between Thai and American English speakers were identified as Random Forest (RF), Non-linear Support Vector Machine (SVM), and Non-linear Logistic Regression (LR). Due to the non-linear nature of these models, SHAP (SHapley Additive exPlanations) values were utilized to measure feature importance. SHAP values offer a unified approach to interpreting the contributions of each feature, providing insights into how individual features influence the model's predictions.

The SHAP values analysis revealed complex interactions between features that contribute to model decisions. The combined SHAP values across all three models demonstrated that the decision-making process relies on multiple interacting features rather than single features in isolation. These results are shown in Table 3.

In the Random Forest model, decisions were made through a hierarchical process where *maxD_value* served as the primary splitting feature at the top levels of the trees, followed by intensity-related features at deeper levels. This hierarchical structure suggests that the model first evaluates the energy peaks in speech before considering finer acoustic details. The Non-linear SVM's decision boundaries were primarily shaped by the interaction between *maxD_value* and *intensity_ratio*. High values in both features typically resulted in classification as native speech, while lower values in either feature led to classification as non-native speech. This interaction reveals that natural-sounding English speech requires both appropriate energy peaks and intensity variation working in concert. The Non-linear Logistic Regression model demonstrated how features work together multiplicatively. For instance, when high *maxD_value* coincided with shorter vowel durations, the probability of native classification increased more than what would be expected from either feature alone. This suggests that the ability to produce strong energy peaks while maintaining precise timing is a key characteristic of native speech.

Feature	Mean SHAP RF	Mean SHAP SVM	Mean SHAP LR
maxD_value	0.014008	0.011795	0.020492

max_vowel_duration	-0.01439	-0.00863	-0.012788
min_syllable_duration	-0.003162	0.001062	-0.029615
max_syllable_duration	-0.001799	-0.002947	-0.01899
intensity_ratio	0.002993	0.007444	0.0101
max_consonant_duration	-0.003373	0.003784	0.01296
max_intensity	0.001774	0.007989	0.009359
consonant_duration_ratio	0.000236	0.001311	0.016142
nPVI-V	-0.004493	0.000853	0.011009
vowel_duration_ratio	0.001621	-0.007161	-0.003881

Table 3 SHAP evaluation of rhythmic characteristics that differ between Thai and American speakers

The combined SHAP values provide a comprehensive view of the most influential features in distinguishing native speakers. Features with consistently positive SHAP values across all models include *maxD_value*, *intensity_ratio*, and *max_intensity*. Higher values for these features increase the likelihood of being classified as a native speaker, indicating that more pronounced energy peaks and greater variability in speech intensity are characteristic of native speech patterns. These features highlight the role of intensity and energy peaks in achieving natural-sounding speech. This finding aligns with research on English prosody, where clear intensity contrasts between stressed and unstressed syllables are fundamental to rhythm. The importance of *maxD_value* particularly suggests that native speakers maintain more distinct energy peaks, which contributes to the stress-timed nature of English.

Conversely, features with consistently negative SHAP values across all models are *max_vowel_duration*, *min_syllable_duration*, and *max_syllable_duration*. Lower values for these features are associated with native speakers, suggesting that shorter maximum vowel and syllable durations are typical of native speech. These features emphasize the importance of shorter vowel and syllable durations in distinguishing native speech patterns. This characteristic reflects the English tendency toward vowel reduction in unstressed syllables, a key feature that differentiates it from syllable-timed languages like Thai. The consistent negative SHAP values for syllable duration metrics indicate that native speakers maintain more compact syllable lengths, particularly in unstressed positions.

Some features show mixed SHAP values across different models, reflecting varying influences. The *max_consonant_duration* demonstrates an interesting pattern - positive SHAP values in the SVM and LR models suggest that longer maximum consonant durations are associated with native speakers, while the negative value in the RF model indicates some variability. This could reflect the complex role of consonant clusters in English, where native speakers may produce longer consonant durations in certain positions while maintaining overall rhythmic timing. Similarly, *consonant_duration_ratio* shows positive SHAP values in the LR model, indicating higher ratios are linked to native speakers, though the SVM model shows a slight negative influence. This variability might represent different strategies in consonant timing between stressed and unstressed syllables.

The nPVI-V (normalized Pairwise Variability Index for vowels) feature shows positive SHAP values in the LR model, suggesting higher variability in vowel durations is associated with native speakers. This aligns with theoretical expectations about English rhythm, where systematic alternation between long and short vowels creates the characteristic stress-timed pattern. The mixed values across models indicate that while this variation is important, it operates in concert with other rhythmic features rather than in isolation. *Vowel_duration_ratio* also exhibits mixed SHAP values, with negative values in the SVM and LR models suggesting lower ratios may be characteristic of native speech, while the positive value in the RF model indicates some variability. This complexity might reflect the interaction between vowel length contrasts and stress patterns in English.

These insights help in understanding the rhythmic differences between native and non-native speakers, emphasizing the significance of energy changes, intensity, and duration in achieving natural speech patterns. The analysis reveals that native English rhythm is characterized by a complex interplay of features: sharp energy contrasts (positive SHAP for *maxD_value* and intensity measures), efficient syllable timing (negative SHAP for duration measures), and systematic variation in both consonant and vowel durations (mixed SHAP values for duration ratios). This multifaceted pattern suggests that achieving native-like rhythm requires coordinating multiple acoustic dimensions rather than focusing on any single aspect.

4.2.4. Conclusion for Research Question 2

The experiment aimed to answer the question: What differences in acoustic features, as evaluated by statistical measures and manual annotations, can be observed when comparing the sampled speech sounds of American speakers with those of Thai speakers of English? Through comprehensive analysis, several critical features were identified that distinguish native from non-native speech patterns.

High classification accuracy was achieved through non-linear Support Vector Machine, Random Forest, and non-linear Logistic Regression models in distinguishing between native and non-native speakers. Through SHAP analysis, *maxD_value*, *intensity_ratio*, and *max_intensity* were consistently identified as key characteristics of native speech, with consistently positive SHAP values across all models. Higher values for these features were associated with native speakers, indicating that more pronounced energy peaks and greater variability in speech intensity are characteristic of native speech patterns.

Conversely, shorter vowel and syllable durations were established as distinctive features of native speech, as evidenced by the consistently negative SHAP values for *max_vowel_duration*, *min_syllable_duration*, and *max_syllable_duration*. Systematic differences in *maxD*, *pitch*, and *intensity* alignment patterns between native and non-native speech were documented, particularly in content words and multisyllabic structures.

These findings highlight how energy changes, as reflected in the timing and intensity of speech peaks, are crucial in capturing the prosodic patterns that differentiate native and non-native speakers. The identified features provide specific acoustic parameters for enhancing naturalness in second language pronunciation. These results offer a practical foundation for developing targeted pronunciation training tools and language-learning applications, further supporting the ultimate goal of guiding non-native speakers toward achieving more natural speech rhythms.

While this study provides valuable insights into the rhythmic features that distinguish Thai and American English speakers, it is noted that the features examined are not exhaustive. Future research should consider a broader range of features and larger, more diverse datasets to further validate and expand upon these findings.

4.3. Comparative Rhythm Analysis: Alignment Location Patterns of maxD, Pitch and Intensity

This section presents a comparative analysis of features that contribute to temporal rhythms between Thai and American English speakers. By examining the alignment of rhythmic features such as *maxD* positions in relation to vowel onsets, the analysis aims to uncover distinctive prosodic patterns that differentiate the two language groups. This evaluation provides insights into the temporal structuring of speech and its implications for understanding language-specific rhythmic characteristics. In doing so, we examined the alignment patterns of *maxD*, *pitch*, and *intensity* within speech, focusing on their roles in shaping the prosodic contours of Thai and American English. By analyzing the alignment (see Fig. 2) and the distribution of acoustic features (Tables 4, 5, 6 and 7), the study seeks to elucidate the underlying rhythmic structures that contribute to the naturalness and fluency of speech in different language contexts.

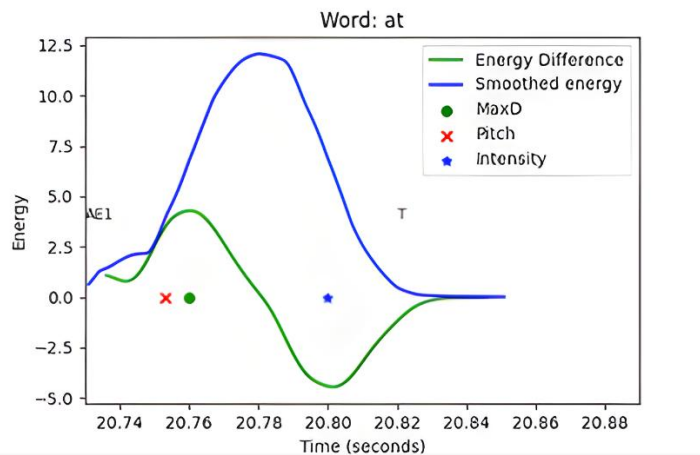


Fig. 2 An alignment pattern for the word *at* spoken by an American speaker

To begin, Table 4 exhibits significant patterns in the location of pitch, intensity, and maxD. Noticeably, *the* is the only word that the location of maxD is not shown. This occurrence could be due to the fact that the word is uttered very softly which clearly does not function as prominence in sound of English speech. The most common pattern is the alignment of pitch followed by maxD and intensity, found in *train*, *go*, *her* and *at*. The word *train* and *go* contain mid vowels /eɪ/ and /əʊ/. The alignments of maxD followed by intensity and pitch are frequently found in *meet*, *will* and *we*. These words share a commonality in that they contain high vowels /i:/, /ɪ:/ and /i:/ respectively. These observations suggest that there could be a relationship between the types of vowels and the locations of maxD, pitch and intensity in the series of alignments. The higher number of occurrences in Pattern 1 is depicted in *train*, *go* and *meet*. All of which are content words, thus it would be possible to hypothesize that the specific patterns containing maxD could be associated with words that convey meanings in communication.

Most words exhibit a consistent “full” (sound) with a mid-tone, meaning that the utterances are produced in (relatively) full and clear words. While distinct characteristics are shown in *the* which is pronounced with weaker sounds. All words are pronounced with mid tone, with the exception of *meet* pronounced with high tone.

Word	Pattern 1	Pattern 2	Intensity of volume	Tone
the	pitch-intensity (15)	pitch (9)	weak (30)	mid (34)
train	pitch-maxD-intensity (30)	maxD-intensity-pitch (2)	full (34)	mid (33)
go	pitch-maxD-intensity (26)	maxD-pitch-intensity (5)	full (34)	mid (34)
her	pitch-maxD-intensity (15)	maxD-pitch-intensity (6)	full (31)	mid (32)
at	pitch-maxD-intensity (9)	maxD-intensity-pitch (8)	full (15)	mid (20)
meet	maxD-intensity-pitch (27)	maxD-pitch-intensity (2)	full (34)	high (22)
will	maxD-intensity-pitch (14)	maxD-pitch-intensity (7)	full (24)	mid (34)
we	maxD-pitch-intensity (9)	maxD-intensity-pitch (8)	full (28)	mid (34)

Table 4 Patterns of the locations of maxD, pitch and intensity occurring in the sounds of monosyllabic words: all speakers (34)

Table 5 details the differences between American speakers (“en”) and Thai speakers (“th”). There are distinct patterns in the sequence of maxD, intensity, and pitch, as well as variations in loudness and tone across different words. While *meet* and *train* show consistent characteristics between American and Thai speakers, suggesting a similarity of the utterance of sounds. Other words, on the other hand, for example, *at* and *her* exhibit more variability between American and Thai speakers. This variability suggests that certain words are articulated differently depending on the dialect or individuals, potentially influencing their stress, emphasis, and overall pronunciation.

Looking closely into the alignments of maxD, intensity and pitch, American and Thai speakers share similarities in the patterns when pronouncing the monosyllabic word, *her* [pitch-maxD-intensity], *will*, *meet* [maxD-intensity-pitch], *go*, *train* [pitch-maxD-intensity] and *we* [maxD-pitch-intensity]. The different patterns are shown in the sounds of *the* and *at* produced by American and Thai speakers. Only pitch appears in the alignment pattern of *the*, while the alignment of *at* is maxD-intensity-pitch by American speakers and pitch-maxD-intensity by Thai

speakers. These occurrences could be attributable to the fact that *the* and *at* are clearly unstressed words in the sentence, and it is noticeable the absence of maxD in Pattern 1 and 2 of *the* of both American and Thai speakers and pattern 2 of *at* pronounced by American speakers. Thus the absence of maxD can potentially suggest an unstressed nature of the word utterances.

Word	Pattern 1	Pattern 2	Intensity of Volume	Tone
her (en)	pitch-maxD-intensity (11)	maxD-pitch-intensity (3)	full (15)	mid (17)
her (th)	pitch-maxD-intensity (4)	maxD-pitch-intensity (3) intensity-pitch-maxD (3) pitch-intensity-maxD (3)	full (16)	mid (15)
will (en)	maxD-intensity-pitch (8)	pitch-maxD-intensity (3) maxD-pitch-intensity (3)	full (17)	mid (17)
will (th)	maxD-intensity-pitch (6)	pitch-maxD-intensity (4) maxD-pitch-intensity (4)	full (14)	mid (17)
meet (en)	maxD-intensity pitch (14)	pitch-maxD-intensity (1) intensity-maxD-intensity (1) maxD-pitch-intensity (1)	full (17)	high (12)
meet (th)	maxD-intensity-pitch (13)	maxD-pitch-intensity (1) intensity-pitch-maxD (1) intensity-maxD-pitch (1) pitch-maxD-intensity (1)	full (17)	high (10)
go (en)	pitch-max-D-intensity (14)	maxD-pitch-intensity (2)	full (17)	mid (17)
go (th)	pitch-maxD-intensity (12)	maxD-pitch-intensity (3)	full (17)	mid (17)
train (en)	pitch-maxD-intensity (15)	maxD-pitch-intensity (2)	full (17)	mid (16)
train (th)	pitch-maxD-intensity (15)	maxD-intensity-pitch (2)	full (17)	mid (17)
we (en)	maxD-pitch-intensity (4)	maxD-intensity-pitch (3)	full (14)	mid (17)
we (th)	maxD-pitch-intensity (5)	maxD-intensity-pitch (5)	full (15)	mid (17)
the (en)	pitch (8)	pitch-intensity (7)	weak (17)	mid (17)
the (th)	pitch-intensity (8)	pitch-maxD-intensity (5)	weak (13)	mid (17)
at (en)	maxD-intensity-pitch (5)	pitch (3) pitch-maxD-intensity (3)	weak (17)	mid (16)
at (th)	pitch-maxD-intensity (6)	maxD-pitch-intensity (4)	full (15)	falling (8)

Table 5 Patterns of the locations of maxD, pitch and intensity occurring in the sounds of monosyllabic words: American (17) and Thai speakers (17)

Table 6 shows the alignment patterns of two disyllabic words *Wednesday* and *station*, depicting distinct patterns in the sequence of maxD, intensity, and pitch, along with variations in the intensity of volume, stress location, and

tone. As shown in Pattern 1, there is a similarity in part of the alignments of maxD followed by intensity, pitch and pitch. *Wednesday* exhibits a complex structure with variations in stress across syllables and a dynamic tonal pattern, while *station* shows more consistency with a strong initial emphasis and a mid-falling tone. These patterns highlight the importance of syllable structure and stress placement in determining the overall pronunciation and emphasis of these words in speech. The only difference in tonal pattern in the mid tone produced by American speakers and falling tone produced by Thai speakers. Because falling tone produced by Thai speakers only appears in *at*, it is unclear whether or not the findings contrast to Kang (2013a) and Kang & Johnson (2021) suggesting that falling tone occurs more commonly in speakers with higher proficiency of English. Further research may evaluate a more variety of sentences in order to identify the similarities and differences in the tones produced by Thai speakers with higher and lower English proficiency.

The composition of phonetic features for *Wednesday* and *station* is different. *Wednesday* contains the following features: approximant /w/ – low vowel /ɛ/ – coda /nz/ – alveolar stop /d/ – high vowel (diphthong) /eɪ/, while *station* comprises post-alveolar fricative /s/ – alveolar stop /t/ – high vowel /eɪ/ – post-alveolar /ʃ/ – nasal stop /n/. It could be noteworthy that further exploration may consider comparing the alignments patterns of disyllabic words sharing the same composition of phonetic features. In doing so, it could be possible to identify whether there is a relationship between commonalities in the alignment patterns and the phonetic composition of words.

Word	Pattern 1	Pattern 2	Intensity of Volume	Stress Location	Tone
Wednesday	maxD-intensity-pitch-pitch-maxD-intensity (11)	maxD-intensity-maxD-intensity-pitch-pitch (5)	full (34)	first (18) first and second (16)	high-mid (34)
station	maxD-maxD-intensity-pitch-pitch-intensity (5)	maxD-intensity-maxD-intensity-pitch-pitch (4)	full (26) full-weak (8)	first (34)	mid-falling (34)

Table 6 Patterns of the locations of maxD, pitch and intensity occurring in the sounds of disyllabic words: all speakers (34)

Table 7 shows that there are some consistent patterns across American and Thai speakers, with notable differences, particularly in stress location and tone. *Wednesday* exhibits more consistency in stress and tone across both speaker groups, while *station* shows greater variability, especially in how syllables are emphasized and how the pitch contour is applied. These differences highlight the differences on the articulation of disyllabic words, reflecting the unique phonetic characteristics of speech sounds produced by American English and Thai speakers.

Looking deeply into the alignment patterns, both American and Thai speakers share a commonality in patterns of *Wednesday* [maxD-intensity-pitch-pitch-maxD-intensity], but *Station* shows different alignment patterns. It is clear that the stress locations are different in the pronunciation of both words. While American speakers emphasize the first syllable, many Thai speakers emphasize both first and second syllables. This allude to the fact that the last syllables of disyllabic words must be stressed in Thai language (Warotamasikkhadit, 2000), and it is possible that the same sound system is transferred to the way in which English is spoken by Thais. This means that instead of stressed sound located at the first syllables of a disyllabic word functioning as a noun, the second syllable is also stressed as exemplified in the stress locations of *Wednesday* and *station*.

Word	Pattern 1	Pattern 2	Intensity of Volume	Stress Location	Tone
Wednesday (en)	maxD-intensity-pitch-pitch-maxD-intensity (5)	varied patterns appearing once or twice each	full (17)	first (17)	high-mid (12)
Wednesday (th)	maxD-intensity-pitch-pitch-maxD-intensity (6)	varied patterns appearing once or twice each	full (17)	first and second (16)	high-mid (13)
station (en)	pitch-pitch-maxD-maxD-intensity-intensity (3)	varied patterns appearing once or twice each	full (9) full and weak (8)	first (17)	mid-falling (9) mid-low (8)
station (th)	maxD-maxD-intensity-pitch-pitch (3)	varied patterns appearing once or twice each	full (17)	first (9) first and second (8)	mid-low (9) mid-falling (4)

maxD-intensity-
pitch-pitch-maxD-
intensity (3)

Table 7 Patterns of the locations of maxD, pitch and intensity occurring in the sounds of disyllabic words: American (17) and Thai speakers (17)

To bring this project to close, the findings suggest that maxD is an effective algorithmic-based parameter that can help identify beats in syllables that contribute to characteristics of distinctive rhythm, while an absence of maxD can indicate a weak form of words that should not be stressed in simply normal speech following the idea of locutionary act (Searle, 1969). This complements to the findings in Section 4.2 that intensity and energy peak together with shorter vowels and syllable duration contribute to the speech-like rhythmic patterns, longer maximum consonant duration, which can be differentiated from those of Thai speakers. Further research could consider experimenting with a more variety of sentences and delve deep into the phonetical structure of words and the alignments of maxD, pitch and intensity.

5. Limitations and Future Research

Several limitations of this study should be acknowledged. First, our analysis did not systematically investigate the method's robustness across diverse speaker attributes. While demographic factors like gender distribution, age variations, and sociolinguistic backgrounds were present in our data, we did not explicitly evaluate how these variables impact the model's performance or analyze potential biases in rhythm detection across different speaker groups. Variations in vocal quality, including pitch range, voice resonance, and speaking rate, which could significantly influence rhythmic patterns, were not controlled for in our experimental design. These omissions may limit the model's reliability when applied to heterogeneous speaker populations.

Regional dialect variations within both Thai and American English speakers were not accounted for, and speakers' educational backgrounds, English proficiency levels, and prior exposure to English instruction were not standardized. Cultural biases may also exist in the elicitation paragraph's Western context. In support of diversity and inclusion, it is important to note that the issues of language regard and the equality of value in language dialects and accents should be taken into consideration when it comes to pronunciation training. We support that pronunciation training technology and the practice of automatic speech analysis should be based on naturalness of sounds aimed at achieving comprehensibility in communication rather than superiority of one language/dialect/accents over another.

The sampling approach presents additional limitations. The small sample size of 34 speakers and reliance on a single sentence limit the generalizability of our findings. The recordings from the Speech Accent Archive were collected under optimal conditions with minimal background noise, which may not reflect real-world speech environments. The effectiveness of maxD in detecting beat locations and the accuracy of classification models under various noise conditions, recording devices, and audio qualities remain untested.

Methodological limitations include potential systematic biases in our classification models through feature selection that could favor certain speaking styles or accents. The binary classification approach may oversimplify the spectrum of English pronunciation, and the maxD parameter's effectiveness may vary across different phonological systems. Additionally, the lack of direct comparison with recent deep learning approaches to rhythm analysis limits our understanding of our method's comparative advantages.

Future research should address these limitations through:

- 1) Larger, more demographically diverse speaker samples with detailed background documentation,
- 2) Multiple sentence types and speaking contexts,
- 3) Testing under various noise conditions and recording qualities,
- 4) Benchmarking against deep learning approaches,
- 5) Comprehensive analysis of model performance across speaker attributes including age groups, gender, and vocal characteristics,
- 6) Comparing the effects of capturing the rhythmic features of speech using the maxD parameter and other traditional metrics, such as %V, ΔC , or PVI,
- 7) Incorporating a multi-parameter approach to accommodate more acoustic and temporal features for a more nuanced representation of rhythm,

- 8) Comparing and expanding the analyses to other languages with different rhythmic properties, such as syllable-timed and stress-timed languages,
- 9) Evaluating the generalizability by adopting the model parameters to analyze the rhythm of speech in other languages to broaden the model's applicability.

The code and model parameters will be made publicly available to support reproducibility and enable broader application of these findings in the research community.

6. Conclusion

This study analyzed rhythmic differences between Thai and American English speakers, focusing on algorithmic beat prediction and key acoustic features. Our findings validate maxD as an efficient alternative to manual beat annotation and identify specific acoustic features distinguishing native speech patterns. These results have important implications for language teaching: the maxD parameter could be implemented in automated pronunciation feedback systems, while the identified acoustic features could guide the development of targeted rhythm training exercises.

The study makes several key contributions: 1) demonstrating the effectiveness of automated rhythm analysis through maxD, 2) identifying specific acoustic markers of native-like rhythm, and 3) developing a scalable feature extraction pipeline. These advances are particularly significant for large-scale pronunciation research and the development of data-driven language teaching tools.

While this study demonstrates promising results, future work should include comparative analysis with recent deep learning approaches to rhythm analysis to establish the relative advantages of our method. Future research will also focus on validating these findings across broader speaker populations and linguistic contexts, particularly examining the effectiveness of maxD-based feedback in actual language learning settings. While current limitations in sample size and scope necessitate careful application, this work provides a foundation for developing evidence-based pronunciation training tools that could enhance second language rhythm acquisition.

References

- Abercrombie, D. (1967). *Elements of general phonetics*. Aldine.
- Arvaniti, A. (2009). Rhythm timing and the timing of rhythm. *Phonetica*, 66, 46-63.
<http://doi.org/10.1159/000208930>
- Arvaniti, A. (2012). The Usefulness of Metrics in the Quantification of Speech Rhythm. *Journal of Phonetics*, 40, 351–373. <https://doi.org/10.1016/j.wocn.2012.02.003>
- Ashwini, P. & Bharathi, S.H. (2024). Continuous feature learning representation to XGBoost classifier on the aggregation of discriminative Features using DenseNet-121 architecture and ResNet 18 architectures towards Apraxia Recognition in the Child Speech Therapy. *International Journal of Speech Technology*, 27, 1-13.
10.1007/s10772-024-10089-6
- Barry, W.J., Andreeva, B., Russo, M. Dimitrova, S. & Kostadinova, T. (2003). Do Rhythm Measures Tell us Anything about Language Type? In *Proceedings of the 15th ICPHS Barcelona*, 2693-2696.
- Brazil, D. (1997). *The Communicative Value of Intonation in English*. Cambridge University Press.
- Brown-Bousfield, M. M., & Chang, C. B. (2023). Regressive cross-linguistic influence in multilingual speech rhythm: The role of language similarity. In Brown-Bousfield, M. M., Flynn, S., & Fernández-Berkes, É. (Eds.), *L3 Development After the Initial State* (pp. 49-71). John Benjamins Publishing.
<https://doi.org/10.1075/sibil.65.03bro>
- Chen, L., Zechner, K., Yoon, S.-Y., Evanini, K., Wang, X., Loukina, A., Tao, J., Davis, L., Lee, C. M., Ma, M., Mundkowsky, R., Lu, C., Leong, C. W., & Gyawali, B. (2018). *Automated scoring of nonnative speech using the speechrater sm v. 5.0 engine*. ETS Research Report Series, 2018(1), 1-31.
- ELAN (Version 6.8) [Computer software]. (2024). Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. Retrieved from <https://archive.mpi.nl/tla/elan>

- Gabrielsson, A. (1999). The Performance of Music. In Deutsch, D. (Eds.), *Psychology of Music* (2nd ed, pp. 501-602). Academic Press. <https://doi.org/10.1016/B978-02213464-4/50015-9>
- Grabe, E., & Low, E. L. (2002). Durational Variability in Speech and the Rhythm Class Hypothesis. In Gussenhover, C., Warner, N. (Ed.), *Papers in Laboratory Phonology 7* (pp. 515-546), Berlin: Mouton de Gruyter. <https://doi.org/10.1515/9783110197105.2.51>
- Halliday, M. A. K. (1967). *Intonation and Grammar in British English*, 48. Jaunua Linguarum. Series Practica: De Gruyter Mouton.
- Hansen, J., Suh, J., Angkititrakul, P., & Lei, Y. (2014). Effective background data selection for SVM-based speaker recognition with unseen test environments: More is not always better. *International Journal of Speech Technology*, 17, 211-221. <https://doi.org/10.1007/s10772-013-9219-z>
- Honing, H. (2002). Structure and interpretation of rhythm and timing. *Dutch J. Music Theory* 7.
- Johnson, D. O., & Kang, O. (2015). Automatic prominent syllable detection with machine learning classifiers. *International Journal of Speech Technology*, 18, 583–592. <https://doi.org/10.1007/s10772-015-9299-z>
- Kang, O., & Johnson, D. O. (2021). Linguistic features and automatic scoring of Aptis speaking performances. AR- G/ 2021/ 3, ARAGs Research Reports Online, Series Editors: Richard Spiby and Carolyn Westbrook, ISSN 2057- 5203. London: British Council.
- Kang, O. (2013). Linguistic analysis of speaking features distinguishing general English exams at CEFR levels B1 to C2 and examinee L1 background. *Research Notes*, 52, 40-48.
- Lai, C., Evanini, K., & Zechner, K. (2013). Applying rhythm metrics to non-native spontaneous speech. In *Speech and Language Technology in Education*.
- Lesnichaia, M. Khaustova, V. Bogach, N. Lezhenin, I., Blake, J., & Pyshkin, E. (2022). Classification of accented English using CNN model trained on amplitude Mel-spectrograms. *Interspeech*, 3669-3673. <http://doi.org/10.21437/Interspeech.2022-46>
- London, J. (2001). Rhythm. In S. Sadie (Ed.), *The New Grove Dictionary of Music and Musicians* (2nd edn., pp. 277-308). Oxford University Press.
- Louni, A., Rizoug, L. & Belmadani, A. (2024). A quantal model for Algerian vowel features identification using formants and subglottal resonances. *International Journal of Speech Technology*, 27, 437-445. <https://doi.org/10.1007/s10772-024-10114-8>
- Low, E. L, G, E., & Nolan, F. (2000). Quantitative Characterization of Speech Rhythm: Syllable-Timing in Singapore English. *Language and Speech*, 4394, 377-401. <https://doi.org/10.1177/00238309000430040301>
- Luangthongkum, T. (1977). *Rhythm in Standard Thai*. Doctoral Thesis. University of Edinburgh.
- McAuliffe, M. Socolof, M. Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: trainable text-speech alignment using Kaldi. In *Proceedings of the 18th Conference of the International Speech Communication Association*, (pp. 498-502). <https://doi.org/10.21437/Interspeech.2017-1386>
- Nainan, S., & Kulkarni, V. (2021). Enhancement in speaker recognition for optimized speech features using GMM, SVM and 1-D CNN. *International Journal of Speech Technology*, 24, 809–822. <https://doi.org/10.1007/s10772-020-09771-2>
- Patel, A. D., & Daniele, J. R. (2003). An Empirical Comparison of Rhythm in Language and Music. *Cognition*, (87), B35-B45. [https://doi.org/10.1016/S0010-0277\(02\)00187-7](https://doi.org/10.1016/S0010-0277(02)00187-7)
- Pierrehumbert, J. B., & Hirschberg, J. (1990). The meaning of intonational contours in the interpretation of discourse. In P. R. Cohen, J. Morgan & M. E. Pollack (Eds). *Intentions in communication* System Development Foundation Benchmark Series. The MIT Press.
- Pike, K. (1945). *The Intonation of American English*. Ann-Arbor: University of Michigan Press.

- Polyanskaya, L., & Ordin, M. (2019). The effect of speech rhythm and speaking rate on assessment of pronunciation in a second language. *Applied Psycholinguistics*, 40(3), 795–819. <https://doi.org/10.1017/S0142716419000067>
- Polyanskaya, L., Ordin, M., & Busa, M.G. (2017). Relative Salience of Speech Rhythm and Speech Rate on Perceived Foreign Accent in a Second Language. *Language and Speech*, 60(3), 333–355. <https://doi.org/10.1177/0023830916648720>
- Python Software Foundation (2021). *Python*. Retrieved from <https://www.python.org/>
- Rathcke, T., & Lin, C. (2023). An acoustic study of rhythmic synchronization with natural English speech. *Journal of Phonetics*. <http://doi.org/10.1016/j.wocn.2023.1012630095-4470>
- Rathcke, T., Lin, C., Falk, S., & Bella, S. D. (2021). Tapping into linguistic rhythm. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, 12(1), pp. 1–32. <https://doi.org/10.5334/labphon.248>
- Roach, P. (2009). *English Phonetics and Phonology: A Practical Course*, 2nd ed. Cambridge University Press.
- Romano, A., Mairano, P., & Calabro, L. (2011). Measures of Speech Rhythm in East-Asian Tonal Languages. In *Proceedings of the 17th International Congress of Phonetic Sciences*, Hong Kong.
- Searle, J. (1969). *Speech Act*. Cambridge University Press.
- Šturm, P., & Volín, J. (2016). P-centres in natural disyllabic Czech words in a large-scale speech-metronome synchronization experiment. *Journal of Phonetics*, 55(2016), 38–52. <http://dx.doi.org/10.1016/j.wocn.2015.11.003>
- Suntornsawet, J. (2022). A Systemic Review of Thai-Accented English Phonology. *PASAA* (63), 348–360.
- Tingsabadh, M.R.K., & Abramson, A.S. (1999). Thai. *Handbook of the IPA*, 147–150.
- Warotamasikkhadit, U. (2000). The accentual system of Thai polysyllabic words. *Manusya: Journal of Humanities*, 3(2), 32–44. <https://doi.org/10.1163/26659077-00302004>
- Weinberger, S. (2015). *Speech Accent Archive*. George Mason University. Retrieved 14 Jan 2024, from <http://accent.gmu.edu>
- Wennerstorm, A. (2001). Intonation and evaluation in oral narratives. *Journal of Pragmatics*, 33(8), 1183–1206. [https://doi.org/10.1016/S0378-2166\(00\)00061-8](https://doi.org/10.1016/S0378-2166(00)00061-8)
- White L, Mattys SL, Series L, & Gage S. (2007). Rhythm metrics predict rhythmic discrimination. In *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS XVI)*. Universität des Saarlandes.
- White L. (2018). Segmentation of speech. In S.-A. Rueschemeyer & G. Gaskell (Eds.), *Oxford handbook of psycholinguistics* (pp. 5–30). Oxford University Press.
- White, L., & Malisz, Z. (2020). Speech rhythm and timing. In C. Gussenhoven & A. Chen (Eds.). *Oxford Handbook of Language Prosody* (pp. 167–182). Oxford: Oxford University Press.
- White, L., Mattys, S. L., & Wiget, L. (2014). Automatic classification of L2 rhythm in syllable-timed languages. *Journal of Phonetics*, 43, 52–68. <https://doi.org/10.1016/j.wocn.2014.02.002>