languages

MDPI

Article

An Open CAPT System for Prosody Practice: Practical Steps towards Multilingual Setup

John Blake [1,*,†], Natalia Bogach [2,*,†], Akemi Kusakari [3], Iurii Lezhenin [2,4], Veronica Khaustova [1], Son Luu Xuan [5], Van Nhi Nguyen [6], Nam Ba Pham [2], Roman Svechnikov [2], Andrey Ostapchuk [2], Dmitrei Efimov [2] and Evgeny Pyshkin [1,*,†]

[1] School of Computer Science and Engineering, The University of Aizu, Aizu-Wakamatsu 965-8580, Japan
[2] Institute of Computer Science and Technology, Peter the Great St. Petersburg Polytechnic University, St. Petersburg 195251, Russia
[3] Hachinohe College, National Institute of Technology, Hachinohe 039-1104, Japan
[4] Speech Technology Center, St. Petersburg 194044, Russia
[5] Suntory System Technology Ltd., Osaka 530-8204, Japan
[6] FPT Japan Holdings Co., Ltd., Tokyo 105-0011, Japan
[*] Correspondence: jblake@u-aizu.ac.jp (J.B.); bogach_nv@spbstu.ru (N.B.); pyshe@u-aizu.ac.jp (E.P.)
[†] These authors contributed equally to this work.

Abstract: This paper discusses the challenges posed in creating a Computer-Assisted Pronunciation Training (CAPT) environment for multiple languages. By selecting one language from each of three different language families, we show that a single environment may be tailored to cater for different target languages. We detail the challenges faced during the development of a multimodal CAPT environment comprising a toolkit that manages mobile applications using speech signal processing, visualization, and estimation algorithms. Since the applied underlying mathematical and phonological models, as well as the feedback production algorithms, are based on sound signal processing and modeling rather than on particular languages, the system is language-agnostic and serves as an open toolkit for developing phrasal intonation training exercises for an open selection of languages. However, it was necessary to tailor the CAPT environment to the language-specific particularities in the multilingual setups, especially the additional requirements for adequate and consistent speech evaluation and feedback production. In our work, we describe our response to the challenges in visualizing and segmenting recorded pitch signals and modeling the language melody and rhythm necessary for such a multilingual adaptation, particularly for tonal syllable-timed and mora-timed languages.

Keywords: CAPT; iCALL; multimodality; signal processing; L2; speech prosody; stress-timed language; syllable-timed language; mora-timed language; speech melody; intonation pattern; rhythm

check for updates

Citation: Blake, John, Natalia Bogach, Akemi Kusakari, Iurii Lezhenin, Veronica Khaustova, Son Luu Xuan, Van Nhi Nguyen, Nam Ba Pham, Roman Svechnikov, Andrey Ostapchuk, and et al. 2024. An Open CAPT System for Prosody Practice: Practical Steps towards Multilingual Setup. *Languages* 9: 27. https://doi.org/10.3390/languages9010027

Academic Editors: Paolo Mairano and Sandra Schwab

Received: 20 July 2023
Revised: 19 December 2023
Accepted: 23 December 2023
Published: 12 January 2024

1. Introduction

Language learners need to learn the meaning, form, and pronunciation of words and phrases to be able to hold a conversation. Pronunciation teaching trends vary over time, but a current trend is the continuing development and improvement of web and mobile applications to assist with pronunciation (Pennington 2021). Multimedia-based educational environments for language learning largely draw on the recent findings from linguistic, pedagogical, and psychological research on language perception, representation, and production. This theoretical base acts as a platform to bring together language-related theories with human–machine interaction paradigms, creating completely new learning possibilities. The rapidly growing field of intelligent Computer-Assisted Language Learning (iCALL) combines the power of artificial intelligence to process language with standard Computer-Assisted Language Learning. Text-based iCALL systems tend to utilize artificial intelligence to classify and categorize language features, while iCALL systems for speech draw on artificial intelligence to process and visualize speech signals.

Computer-assisted pronunciation training (CAPT) systems comprise an important component of iCALL (Pennington and Rogerson-Revell 2019). CAPT systems provide learners with opportunities to practice and test their ability to pronounce the target language items intelligibly. Functionalities among CAPT systems vary, but one of the most common tasks is the listen-and-repeat activity (Couper 2021), which was popularized by behavioristic approaches to language learning (Carey et al. 2015) in which learners repeat model utterances and receive feedback on their performance. CAPT systems rely on signal processing to transform sound waves using algorithms to enable the sounds to be visualized. The transformation process of sound capture and visualization is the same regardless of language. The primary components of CAPT pipelines used to process speech aspects are the same across almost all languages. Despite the variation in articulation among different languages, all human spoken languages use the vocal tract to create meaningful sounds to communicate ideas, intentions, facts, and so forth. In a similar way, despite the wide variation in amplitude, time, pitch, frequency, and phase of sound waves, existing CAPT pipelines capture and transform physical sound waves into two-dimensional visualizations, which could be easier to display and more comprehensible, for example, compared to spectrograms (Sztahó et al. 2014).

The underlying architecture of CAPT environments is, therefore, the same regardless of target language or languages. Present-day innovative CAPT solutions incorporate artificial intelligence, speech processing, and natural language processing in a single CAPT pipeline, thus contributing to the growing domain of iCALL.

Modern CAPT systems may be categorized into three classes according to their functionality, namely, tools designed to help learners improve their pronunciation (dedicated tools), tools that can be adopted for pronunciation teaching (repurposed tools), and tools designed for the analysis of speech sounds (analysis tools) (Velázquez-López and Lord 2021), as shown in Table 1. Dedicated CAPT tools may address discrete units of sound, such as individual vowel or consonant sounds. CAPT tools may deal with larger units, such as the patterns of intonation that spread over multiple sounds, a short string of words, or whole phrases. Despite their numerous reported achievements, extant CAPT tools still lack explicit instructive feedback for the acquisition and assessment of foreign language suprasegmentals (Hardison 2021; Pennington 2021) and lack the ability to support different learning styles by allowing learners to select their preferred feedback modality (Mikhailava et al. 2022).

Table 1. Major classes of CAPT tools.

No.	Class	Examples of Tools
1	Analysis tools	BetterAccent (Hamlaoui and Bengrait 2016), KaSPAR (Tallevi 2017), WinPitch (Martin 2010)
2	Repurposed tools	Forvo (Pierson 2015), Lingua Libre (https://lingualibre.org/) accessed on 22 December 2023, YouGlish (McCarthy 2018)
3	Dedicated tools	CAPT system for hearing impaired (Sztahó et al. 2014), Inton@Trainer (Lobanov et al. 2018), ELSA Speak (Samad and Ismail 2020), StudyIntonation (Boitsova et al. 2018; Bogach et al. 2021)

Thus, in our work we strive to rise to the challenges of the personalization and customization of CAPT feedback to both the target language and the individual learners. Although most CAPT systems address a single language, our aim is to develop an environment suitable for all languages, allowing the visualizations to be tailored to the prosody of the target language. We also aim to cater to the different needs of learners, particularly their cognition channels, by advancing the system features to adapt prosody training for different learning styles and preferences.

This article contributes to the body of research by showing that despite the language-agnostic nature of sound signal processing (Datta et al. 2020), the idiosyncrasies of the pronunciation of each target language present learners with different challenges and require tailored solutions to optimize learning potential. Axiomatically, learners of tonal languages need to focus on tones, yet as Tsukada (2019) points out, the number, range, duration, quality, and importance of tones varies by language, necessitating bespoke solutions even for languages within the same family.

The remainder of the paper is structured as follows. Section 2 provides a brief introduction to pronunciation and prosody. Key pronunciation terminology is defined, and different language taxonomies based on pronunciation are explained. Section 3 outlines the four core challenges we faced in creating a multilingual CAPT environment. These challenges are elaborated in the following four sections. Section 4 introduces the challenges overcome in developing the *StudyIntonation* proof-of-concept prototype, a multimodal CAPT environment. Section 5 details the steps taken to verify the suitability of *StudyIntonation* as a learning tool. Section 6 describes and explains the challenges presented by different target languages. In this section, which is the primary focus of this paper, we discuss the steps taken to tailor the CAPT environment to the different prosodic requirements of learners for three target languages: English, Vietnamese, and Japanese. Section 7 provides an overview of how a CAPT environment can be tailored to meet the needs of individuals based on their mother tongue, learning styles, and learning preferences. A summary of the challenges overcome and potential approaches to adopt to address the outstanding challenges is presented in Section 8.

2. Prosody

One area that merits considerable attention in the field of CAPT is the integration of prosody, the rhythmic and melodic aspects of speech, especially within a multilingual context. Prosody encompasses an array of linguistic features such as pitch, tempo, and stress, which serve as critical constituents of meaning and communicative efficacy (Fox 2000). In a multilingual setting, prosodic patterns often vary widely across languages, thereby constituting a notable challenge for non-native speakers who strive to achieve high levels of proficiency. Understanding and mastering the prosodic elements of a language can impact the intelligibility and naturalness of spoken discourse. As such, incorporating prosody in CAPT applications substantially enriches the learning experience by offering language-specific guidance on these often overlooked but crucial elements of speech.

Segmental features refer to individual phonemes that serve as the building blocks of words. These features often constitute the primary focus of language instruction as they are readily identified and isolated for practice. However, suprasegmental features, which include intonation, stress, and tone, operate over stretches of speech units such as syllables, words, or phrases (Roach 2009). These features are critical for conveying meaning and emotion, and their mastery can significantly influence the comprehensibility and fluency of a non-native speaker. In a multilingual context, the intricacies of segmental and suprasegmental features can vary considerably across languages, making it particularly challenging for learners to adapt to the phonetic and prosodic norms of the target language. Therefore, an understanding of both feature types is essential for non-native learners aiming to achieve a natural and intelligible pronunciation in any given language.

Stress, intonation, and tone are linguistic features that operate on the suprasegmental level (Lehiste 1970), shaping the prosodic landscape of speech. Stress relates to the emphasis placed on specific syllables within words or phrases, often realized through increased loudness, higher pitch, and longer duration. Intonation refers to the variation in pitch across stretches of speech, influencing the communicative functions of declarative, interrogative, and imperative statements (Collier and Hart 1975; Ladd 2008). Tone, particularly relevant in tonal languages, involves the use of pitch variations at the syllable level to distinguish lexical or grammatical meaning. In contrast, pitch is a perceptual parameter related to the frequency of the vocal fold vibrations, which listeners interpret as either high or low. Stress, intonation, and tone are linguistic features that utilize variations in pitch to convey meaning. As McDermott and Oxenham (2008) point out, pitch is a perceptual correlate to frequency. Each language, dialect, or even individual speaker exhibits a specific pitch range and contour. The intricacy of pitch goes beyond its role as a building block for stress, intonation, and tone; its manipulation can convey emotional state, social context, or rhetorical emphasis, adding an additional layer of complexity to spoken communication.

2.1. Pitch

Pitch is often conflated with fundamental frequency (f_0) (Yu 2014), though the two are distinct (Hirst and de Looze 2021). Fundamental frequency reflects the rate of vocal fold vibrations, whereas pitch is the perceptual correlate of this frequency and is subjectively interpreted by listeners as being high or low. As pitch is a perceptual property, it cannot be directly extracted from speech recordings; what can be measured is the fundamental frequency or f_0 contour (Yu 2014), which serves as an acoustic indicator of what listeners perceive as pitch. This distinction is crucial for non-native speakers seeking mastery in the pronunciation of a target language as it underlines the difference between the physical and perceptual aspects of speech sounds.

Thus, understanding the relationship between pitch and fundamental frequency can provide valuable insights for mastering linguistic features such as stress, intonation, and tone. In stress-based languages like English, for example, syllabic emphasis is often marked by a higher fundamental frequency, which listeners perceive as stress due to a higher pitch. In the domain of intonation, the modulation of fundamental frequency over a stretch of speech, often colloquially referred to as the "pitch contour", can significantly impact the meaning of an utterance. For example, the rising fundamental frequency at the end of the English question "You're coming?" signals a question intonation.

For tonal languages like Vietnamese, the manipulation of fundamental frequency at the syllabic level results in different lexical or grammatical meanings, which are then perceived as different tones or pitch contours by the listener. Meanwhile, pitch-accent languages like Japanese employ variations in fundamental frequency to distinguish words that are otherwise phonemically identical. For example, the word *hashi* can signify *bridge* or *chopsticks* depending on its pitch pattern, which in turn is dictated by its fundamental frequency.

Hence, an understanding of pitch and fundamental frequency is indispensable for non-native speakers striving to perfect their pronunciation and to interpret the spoken language more accurately. Whether the target language is stress-timed like English, pitch-accented like Japanese, or tonal like Vietnamese, mastering these acoustic and perceptual aspects of speech will significantly enhance communicative efficacy.

Pitch accent languages are those that assign a pitch to one syllable within a word. This syllable may be pronounced with a higher pitch so that it stands out from the other syllables. Pitch accent languages include Japanese and Swedish. English may use pitch to emphasize a syllable, but the stressed syllable may be emphasized by saying it louder or elongating the sound. Syllables that stand out are known as stressed syllables, while the other syllables are unstressed. English, therefore, does not clearly match all the criteria needed to classify it as a true pitch accent since the method of stressing the syllable differs. In English, up to two syllables may be stressed in multi-syllable words. The main stress is called primary, while secondary stress may also occur on the initial syllable of a word. Mistakes in both the placement of accent or stress may impact intelligibility.

Intonation, as a form of prosody, exists in both tonal and pitch-accent languages. However, the function and prominence of intonation can differ between these types of languages. In tonal languages, where pitch contours at the syllabic level are crucial for distinguishing lexical or grammatical meaning, intonation must be integrated carefully so as not to conflict with these tonal distinctions. As a result, the methods for incorporating intonation in tonal languages may be more constrained, leading to different conventions for conveying prosodic meaning compared to pitch-accent languages. For example, a study by Ploquin (2013) found that Chinese speakers relied on native lexical tones to produce English prosody. The importance of intonation varies by language. As a rule of thumb, when learning tonal languages, in the early stages mastering the individual tones has a greater impact on intelligibility and is therefore more of a priority than working on intonation. However, intonation is much more important when learning English since using intonation inappropriately may create a negative impression. For example, when making a polite request, such as *Could you help me?*, using falling intonation rather than the conventional rising intonation will most likely cause the listener to assume that the speaker is disappointed or angry. This could impact their willingness to provide assistance. Conversely, mastering pitch accent is more important when learning Japanese since misunderstandings may arise from mistakes in the placement of the accent. In fact, misunderstandings may happen between speakers of different varieties of Japanese. A case in point is the Japanese word あめ [ame] when the accent is placed on the second mora; it means *candy* in Tokyo but *rain* in Osaka.

2.2. Isochrony

Isochrony (Abercrombie 1967; Pike 1945) serves as a useful paradigm for categorizing languages based on their rhythmic structures. By focusing on specific pronunciation features, namely, stressed syllables, all syllables, or mora, this method allows linguists to broadly classify languages into stress-timed, syllable-timed, and mora-timed categories. These categories are not without criticism, though (Arvaniti 2009; Kim and Cole 2005). While this method offers several benefits in comparative linguistics, language education, and speech technology, it also carries certain limitations that must be acknowledged. The perception of rhythm in speech is highly subjective and may be influenced by factors including language background and the context of utterance. The perceived rhythm may not necessarily align with the physical properties of the speech signal. Despite this, one of the key advantages of employing isochrony as a classification system is its pedagogical utility. The framework can aid language learners in understanding the intricacies of pronunciation, fluency, and listening comprehension, especially when contrasting languages with different rhythmic bases. However, isochrony is not without its disadvantages. It can sometimes oversimplify the complex nature of language prosody, forcing languages into predefined categories that may not capture their full rhythmic richness. The framework also tends to overlook dialectal variation as different dialects of a single language may exhibit distinct rhythmic properties that challenge the basic classification.

2.2.1. Stress-Timed Languages

In stress-timed languages, such as English, the temporal intervals between stressed syllables are approximately consistent. This rhythmic pattern often results in the shortening of unstressed syllables, contributing to phenomena such as contractions and elisions. For language learners, this can pose a challenge, particularly in hearing and producing unstressed syllables, especially those occurring in indefinite articles and prepositions. We have used this stress-timed English for all the models in our project, but it should be noted that while the English spoken in Anglophone countries is generally categorized as a stress-timed language, some regional varieties of English may not be characterized as such. For example, the English spoken in parts of the Caribbean, India, or Singapore might exhibit syllable-timed characteristics or a hybrid rhythmic structure. Therefore, the landscape of spoken language classification is more complex than often assumed.

2.2.2. Syllable-Timed Languages

Vietnamese exemplifies syllable-timed languages, in which each syllable occupies roughly the same amount of time. This consistent timing can make individual syllables more distinct, simplifying pronunciation for language learners and yielding more regular rhythmic patterns. Syllable-timed languages like Italian and Vietnamese have rather stable patterns in which the time allocated to pronouncing each syllable is approximately the same. Vietnamese is both syllable-timed and tonal, while Italian is syllable-timed and non-tonal.

2.2.3. Mora-Timed Languages

A mora is a phonological unit that contributes to the syllable weight, where each mora occupies approximately the same amount of time during speech. This temporal uniformity of morae renders mora-timed languages distinct in their rhythm and prosodic characteristics. In such languages, linguistic rhythm is dictated not by the number of syllables *per se* but by the number of morae. This moraic structure profoundly influences various linguistic aspects, including phonetics, phonology, and speech processing. Japanese is a quintessential example of a mora-timed language. For example, although the word *Tokyo* in Japanese is two syllables, it is not two morae ('To-kyo') but rather four morae ('To-u-kyo-u' or 'Tō-kyō') due to the extended vowel sounds, represented by the macrons. This moraic structure is fundamental to the rhythm and pacing of the language. This is supported by studies showing that the mora is a key unit in spoken Japanese (Kubozono 1989; Kureta et al. 2006).

3. Challenges

There are numerous challenges to conquer in the development of a multilingual multimodal CAPT environment. These challenges are presented here in four discrete categories to increase readability and provide readers with a linear story of the development of the CAPT environment. As with most research, the practical implementation was more complex, given the inseparably intertwined links between the categories. The four categories are: (1) the creation of the proof-of-concept prototype, (2) the verification of the effectiveness of the feedback loop, (3) tailoring the interface to different target languages, and (4) tailoring the interface to individual learners. An overview of each of these challenges is given below.

The initial challenge was to set up a proof-of-concept prototype to visualize pitch as a two-dimensional wave representing the interactive contours of the model and the learner's pitches. In the current version, the visual feedback is accompanied by the graph distance metrics, based on a dynamic time warping (DTW) algorithm (Rilliard et al. 2011), assuring tempo-invariant and more robust estimation compared to other metrics such as Pearson correlation coefficient, earth mover's distance, and mean square error, unless they are calculated after DTW (thus inheriting tempo invariance from DTW).

The next challenge was to establish whether the CAPT environment was able to provide sufficient feedback to enable the replication of target language models. Pilot assessment studies showed learners were able to produce fundamental frequency contours with a similar curvature to the target models, confirming a positive effect on learner pronunciation.

The third challenge, which is the main focus and makes up the lion's share of this paper, is tailoring the CAPT environment to the needs of language learners for different target languages. Each language presents learners with different sets of pronunciation problems, so a bespoke graphic user interface (GUI) was created for each target language. Three languages were selected: English, Vietnamese, and Japanese. All three are representative examples of languages from significantly distinctive language groups based on isochrony. English is stress-timed, Vietnamese is syllable-timed, and Japanese is mora-timed.

One of the important continuing challenges is to provide learner-friendly feedback (Mikhailava et al. 2022) and tailor the interface to suit the needs and preferences of language learners by enabling learners to study in their own bespoke individualized CAPT environments, which optimize learning by matching the environment to the needs, preferences, and strategies of learners. This is a challenge that we have yet to overcome.

4. Challenge 1: Proof-of-Concept Prototype

Our proof-of-concept prototype is the *StudyIntonation* project, a multimodal CAPT environment comprising a toolkit that manages mobile applications for a set of languages.

4.1. Global-Local Framework

The relationship between the CAPT environment and the target language is not dissimilar to the glocalization (Roudometof 2023), i.e., the global standardization and localization that occurs in the retail and service industry. For example, the general framework of a retail franchise is prescribed, but localization occurs at national or regional levels to maximize customer satisfaction.

In programming, this global-local nexus is analogical to the instantiation of classes. For example, generic constructions (such as C++ templates) provide a way to define a generalized class, which is invariant to the specific types of objects that this class works with. However, defining a perfectly generic template is often non-trivial. This is why, in the process of template instantiation, the developer may face an actual incompatibility of the template with a particular parameter type. Thus, unless the target parameter type can be adapted, or the template itself could be updated, additional constraints, template specialization, or the construction of a new independent class may be required. Figure 1 illustrates the major possible scenarios that may happen in the process of template instantiation.

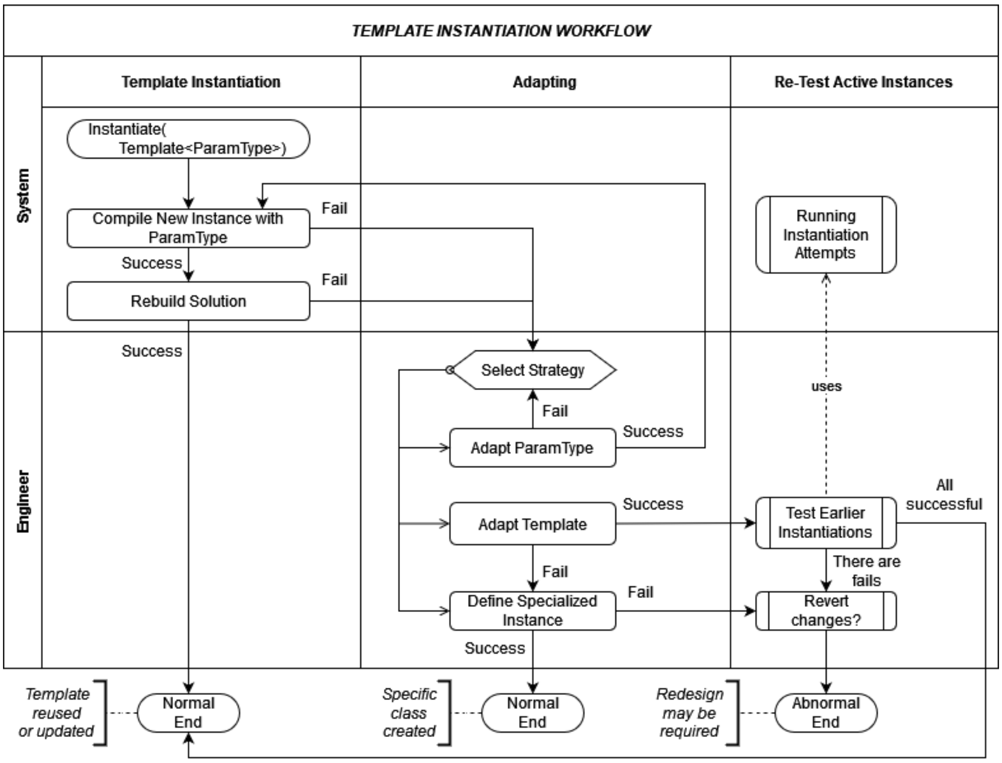


Figure 1. Template instantiation in programming as a metaphor of multilingual setup.

The programming template instantiation serves as a good metaphor for re-using the generalized models provided by signal processing and NLP algorithms in the case of setting up an ideal multilingual CAPT environment. The *StudyIntonation* CAPT environment and associated toolkit that manages mobile applications is the global environment, while the language-specific graphic user interfaces and tailormade feedback developed for the prosody of each language are local implementations.

4.2. System at a Glance

Figure 2 shows the major implemented units and proof-of-concept components developed in the scope of the project. Having developed a CAPT environment, originally aimed at helping learners of English practising spoken language using speech intonation modeling and visualization (which, in fact, are based on language-agnostic algorithms), we are now able to advance the system and describe how we adapted it for pronunciation training for languages other than English, specifically, Vietnamese and Japanese. Actually, all three are representative examples of the languages from each of the three different classes of isochrony, namely, stress-timed, syllable-timed, and mora-timed.

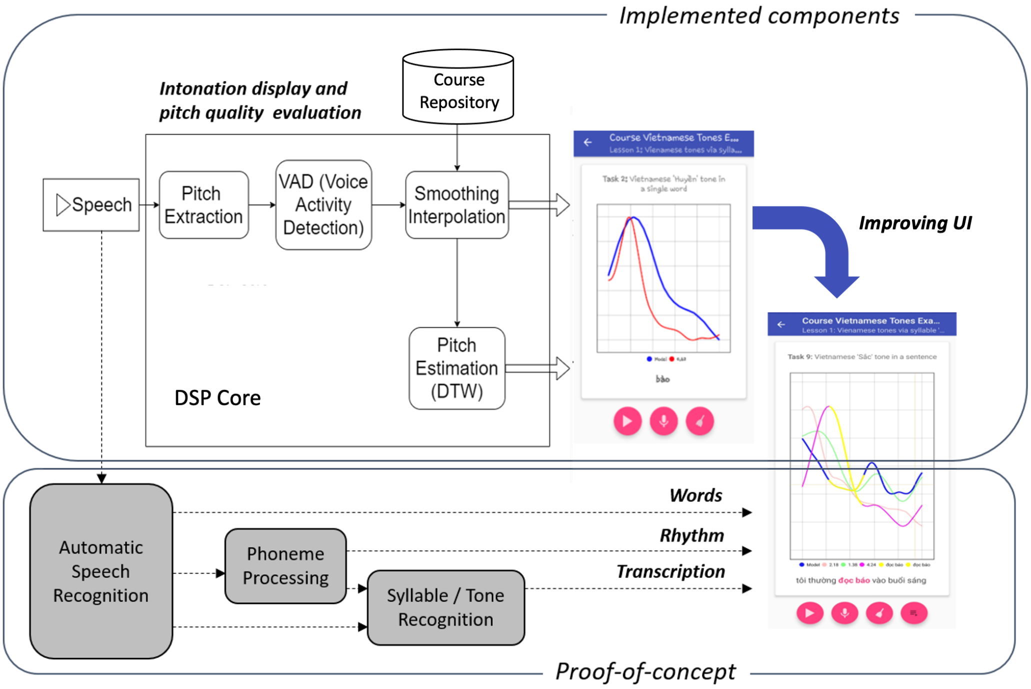
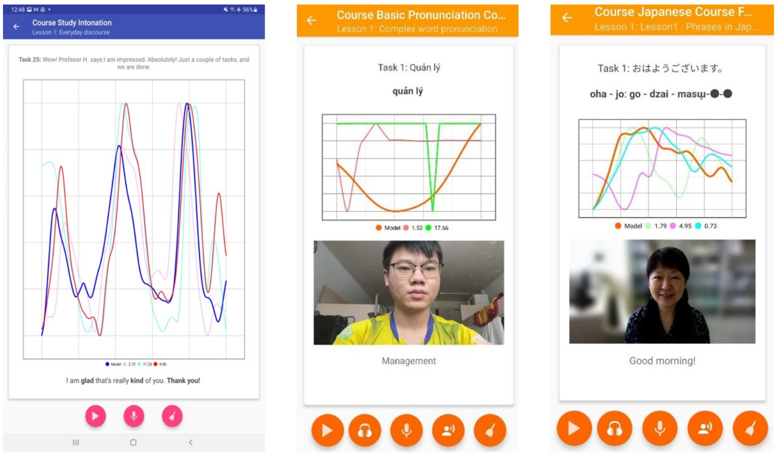

Implemented components
Intonation display and pitch quality evaluation
Course Repository
Speech
Pitch Extraction
VAD (Voice Activity Detection)
Smoothing Interpolation
Pitch Estimation (DTW)
DSP Core
Course Vietnamese Tones E...
Lesson 1: Vienamese tones via sylla...
Task 2: Vietnamese 'Huyền' tone in a single word
Model
bào
Improving UI
Course Vietnamese Tones Exa...
Lesson 1: Vienamese tones via syllable '...
Task 9: Vietnamese 'Sắc' tone in a sentence
Model 2.18 1.38 4.24 đọc báo đọc báo
tôi thường đọc báo vào buổi sáng
Automatic Speech Recognition
Phoneme Processing
Syllable / Tone Recognition
Words
Rhythm
Transcription
Proof-of-concept

Figure 2. *StudyIntonation* CAPT environment: project in progress.

From the perspective of learner experience, representing spoken language visually in the form of interactive pitch contours of the model and learner attempts is a key element of the multimodal platform. In the current version, the visual feedback is accompanied by the graph distance metrics, based on a dynamic time warping (DTW) algorithm applied for paired pitch contours as per Rilliard et al. (2011). This metric is able to provide tempo-invariant and more robust estimation than the other metrics tried, such as the Pearson correlation coefficient, the earth mover's distance, and mean square error (Hermes 1998).

Figure 3 shows examples of application screens from three pronunciation courses representing a stress-timed language (English), a syllable-timed tonal language (Vietnamese), and a mora-timed pitch accent language (Japanese).

The personalization and customization of CAPT interfaces relies on accurate automated speech recognition, which is impacted by sociocultural (e.g., geographic and demographic) and individual variations (e.g., in the shape and proportion of vocal organs). Applying speech recognition algorithms to the recorded model and user pitches can create grounds for practical steps towards providing more instructive, customized feedback of the CAPT system, thus naturally contributing to multilingual features of a CAPT system. Some of the technical challenges are elucidated below.


12:48
56%
Course Study Intonation
Lesson 1: Everyday discourse
Model 2.72 11.23 4.98
Course Basic Pronunciation Co...
Lesson 1: Complex word pronunciation
Model 1.53 17.66
Course Japanese Course F...
Lesson 1: Lesson1 : Phrases in Jap...
Model 1.79 4.95 0.73

Task 25: Wow! Profesor H. says:I am impressed. Absolutely! Just a couple of tasks, and we are done.

I am **glad** that's really **kind** of you. **Thank you!**

Task 1: Quản lý

quản lý

Management

Task 1: おはようございます。

oha - jo: go - dzai - masụ-●-●

Good morning!

(a) English (b) Vietnamese (c) Japanese

Figure 3. *StudyIntonation* interface screens from three different pronunciation courses.

4.3. Speech Signal Processing and Visualization

The construction of intonation contours of spoken phrases is enabled by using fundamental frequency extraction, which is a conventional operation in acoustic signal processing. Estimating and measuring fundamental frequency and modeling the pitch produced are non-trivial tasks (Hirst and de Looze 2021), which may require an external reference to verify all of the stages. It was shown in Bogach et al. (2021) how Praat software could be of much assistance when live speech recording for pitch extraction is conducted; voice activity detection algorithms are used on the raw pitch readings before the other signal conditioning stages. The latter stages include pitch filtering, approximation, and smoothing, which are standard digital signal processing stages to obtain a pitch curve adapted for pedagogic purposes.

4.4. Estimation Algorithms

The visual display of the fundamental frequency (which is the main acoustical correlate of stress and intonation) combined with audio feedback (as implemented in the *StudyIntonation* environment) is very helpful (Bogach et al. 2021). However, the feasibility of visual feedback increases if the learner's pitch contour is displayed along with a possible formalized interpretation of the difference between the model and the learner's pitches. There are a variety of conventional speech processing algorithms used for pitch estimation (Chun 1998; Klapuri 2009). Nevertheless, the choice of the most suitable prosody-based pitch estimation and similarity evaluation methods according to application areas is still disputable. In different works, the metrics used to compare the model and the learner pitches usually include mean square error, Pearson correlation coefficient, earth mover's distance, and dynamic time warping (DTW). DTW can be thought of as a way to "measure the similarity of a pattern with different time zones" (Permanasari et al. 2019, p. 1). Rilliard et al. (2011) demonstrated that the DTW model is effective at capturing similar intonation patterns that are tempo-invariant and, therefore, more robust than Euclidean distance.

There are two limitations to drawing on DTW scores. The first is that the score is static, giving only a glimpse of a learner's performance at one point in time. The second is its generality since the score provides a holistic overview. The lack of discrete data measures makes interpreting the score difficult, potentially leaving learners wondering what they need to alter to improve their performance. Alternatively, the effects of synchronization and coupling may be obtained by harnessing metrics constructed using cross-recurrence quantification analysis (CRQA) (Orsucci et al. 2013; Webber and Marwan 2015). This approach helps create more constructive and comprehensible CAPT feedback. When learners are expected to synchronize prosodic characteristics with the model, CRQA metrics may provide valuable insights into the learning dynamics. At the very least, the CRQA metrics should correlate with more successful and less successful attempts of learners. The effect of prosodic synchronization has been studied and reported in applications for emotion recognition (Vásquez-Correa et al. 2016), and for the analysis of conversational dynamics (Fusaroli and Tylén 2016).

5. Challenge 2: Feedback Loop

A feedback loop provides learners with information that they can use to improve their speech. The importance of feedback on pronunciation has been shown in multiple studies (see, for example, McCrocklin 2016; Neri et al. 2006). A simple feedback loop was described in the Shannon–Weaver communication model (Bell and Henderson 2022). Learners can only act on feedback that is comprehensible, so the digitalized signals need to be transformed from something that only computers can read to some kind of output that is human-readable and ideally easy to understand. Learners repeating utterances without feedback may be unable to monitor the accuracy of their speech and so are unlikely to be able to make incremental improvements in their pronunciation. A key component of a CAPT system is therefore to convey to learners what needs to be changed in order to bring their pronunciation closer to the model.

The first test of the effectiveness of the feedback loop in the *StudyIntonation* CAPT environment is to establish whether native speakers can replicate the prosody of other native-speaker utterances. Given that prosody varies not only by language but according to the individual, there is no certainty that a system can cope with such variation. If the system provides sufficient feedback for a native speaker to replicate the model, then there is the possibility that a non-native speaker may also be able to replicate the prosodic features. However, should a native speaker be unable to replicate the model, the likelihood is that a non-native speaker would also be unable to do so. Thus, the first challenge lies in establishing the sufficiency of the feedback.

The first two screenshots presented in Figure 4 illustrate that User 1, a native speaker, was able to adapt his intonation to the pattern of the model while gradually improving his scores after a series of six attempts of pronouncing the phrase *How is the conference going for you?*

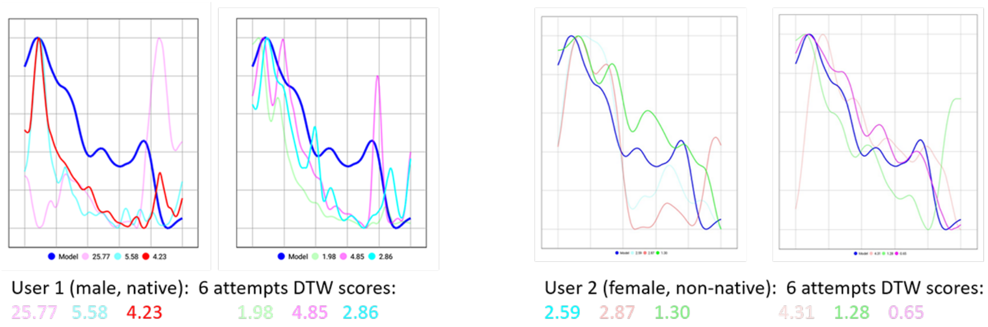


Figure 4. Example of DTW scores achieved by both native and non-native speakers. Each image shows the graph of the model in blue and three user attempts. (This figure is extracted from Mikhailava et al. (2022)). Please note that the values below the figure on the far right have been corrected.

Having established that replication is possible, the next challenge was to test whether non-native speakers could also replicate the model utterances. The following two screenshots on the right-hand side of Figure 4 illustrate an interesting phenomenon, in which User 2, a non-native speaker, was able to outperform a native speaker.

Though DTW provides an objective tempo-invariant primary estimation of the learner's ability to replicate the model, the pitch graph does not provide sufficient information to discover whether and to what extent improvements in prosody could be made. This means that the feedback combining the pitch graphs and numerical scores is constructive and consistent but not instructive enough.

6. Challenge 3: Tailoring to Languages

All spoken languages share a common mode of sound production—that of air through the vocal cords. Signal processing and visualization algorithms are applicable to all languages, so despite being initially designed to focus on processing sound signals of English language, the core of a CAPT system is language-agnostic. The key components of the *StudyIntonation* CAPT environment, such as the digital signal processing core, pitch graph visualizer, user interface, and dynamic time warping pitch quality evaluation unit, which are described in depth in Bogach et al. (2021), are not sensitive to a particular language. Therefore, they naturally fit the requirements for configuring the environment for multilingual use. However, in the process of system adaptation to languages other than English, we discovered a number of important aspects to be addressed in order to allow for the seamless incorporation of the new pronunciation training courses for different languages. The ideal CAPT prosody solution is a single environment enabling users to practice any language. The fundamental frequency (F_0) is the most important determinant of prosody for all spoken languages (De Cheveigné and Kawahara 2002). However, despite the underlying similarities in the production of spoken languages, learner needs vary by language. The problem stems from prosodic differences between languages. Prosody has semantic, structural, stylistic, and social functions. For example, using falling intonation rather than rising intonation in English may convey negativity, using a rising tone rather than a falling tone may alter the meaning in tonal languages, and using an inappropriate pitch accent in Japanese leads to confusion.

We developed CAPT environments for three languages, each of which exhibits different isochrony. In other words, the rhythm of each of the three languages differs. The controlling factor for the division is based on the (relatively) stable intervals of time between language features. English is stress-timed, so the key determiner of rhythm is the time between stressed syllables. In Vietnamese, syllable durations are more stable than other prosodic units, so Vietnamese is described as being syllable-timed. Japanese has been described as syllable-timed (and in some instances it is), but more accurately it is mora-timed. This is because the mora durations are more stable than other prosodic units. In addition to different rhythmic patterns, there are numerous other pronunciation aspects that language learners need to master to be able to pronounce words and phrases intelligibly.

6.1. Stress-Timed Language (English)

The impetus for the creation of *StudyIntonation* was to help Russian learners of English, so from the outset the system was designed to achieve this aim. The challenges faced for the English language were addressed during the development of the proof-of-concept prototype, and its fitness for this purpose was verified by ensuring that native speakers of English were able to replicate the utterances.

Both Russian and English may be classed as stress-timed languages. Typically, when the mother tongue and target language both share the same timing, learners do not usually have problems with this aspect of rhythm when first learning the language. What is problematic for foreign language learners of English regardless of mother tongue is intonation, that is, the change of pitch, over a word, phrase, or clause. Although the pitch changes in English are more gradual and may stretch over more words than those of tonal languages, selecting appropriate intonation patterns is a challenge for many learners. There is significant interaction between the word that receives the main stress in a sentence or clause, and the intonation used.

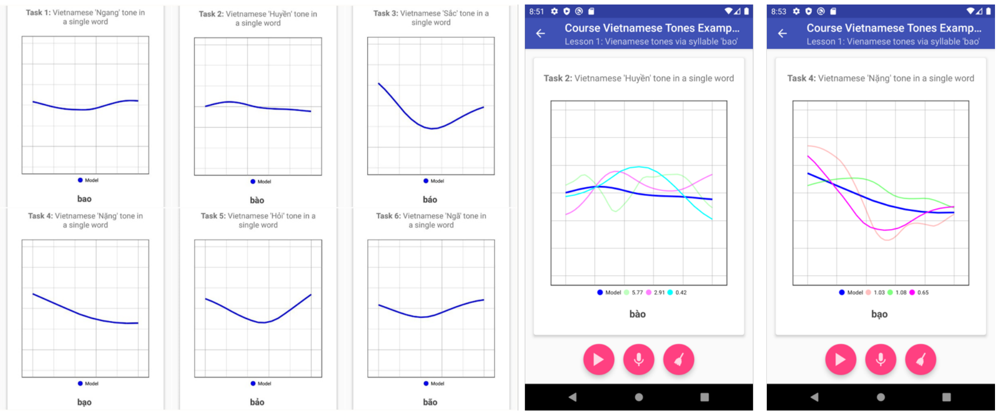
At its most simplistic level, teachers of English pronunciation tend to focus on the two common intonation patterns: falling intonation and rising intonation. Both types of intonation are both grammatical and attitudinal. This multi-functionality is one of the sources of confusion for learners of English. For example, the default intonation pattern for closed questions (e.g., Are you okay?) is rising intonation, while the default intonation pattern for open questions (e.g., How are you?) is falling intonation. The situation is exacerbated when expressing attitude using intonation since positive feelings, such as surprise and delight, are shown using rising intonation while disappointment and anger are shown using falling intonation. This may mean that a learner of English asking a closed question with falling intonation unknowingly creates a poor impression. The intonation impacts meaning, albeit it not lexical meaning.

Thus, the CAPT environment for English prosody is geared towards helping learners understand and use appropriate intonation patterns.

6.2. Syllable-Timed Language (Vietnamese)

For tonal languages, the accurate use of tone is paramount. Vietnamese is usually grouped with the so-called register languages, where tones do not rely solely on pitches but also on length, contour melody, intensity, and phonation as the constituent elements of the register complex (Pham 2004). Learners with non-tonal L1 backgrounds face difficulties recognizing and producing the tones, while appropriate tone articulation is required to convey the correct meaning. The interplay between independent tones and phrasal intonation is more complex compared to non-tonal languages. Though even in non-tonal languages one can find occasional uses of short-term pitch change, in the overwhelming majority of cases, differences in pitches do not have as much impact on meaning disambiguation or speech recognition as in tonal languages.

Van et al. (2021) discuss the required steps for adopting the CAPT interface for both tone and phrasal intonation in a pilot study using *StudyIntonation* for Vietnamese. With regard to the user interface and course organization, for mastery of independent tones (for example, for the tones shown in Figure 5), the system did not need significant adjustments: users can practice the tones using the independent exercises.



Task 1: Vietnamese 'Ngang' tone in a single word

Task 2: Vietnamese 'Huyền' tone in a single word

Task 3: Vietnamese 'Sắc' tone in a single word

Task 4: Vietnamese 'Nặng' tone in a single word

Task 5: Vietnamese 'Hỏi' tone in a single word

Task 6: Vietnamese 'Ngã' tone in a single word

Course Vietnamese Tones Examp...
Lesson 1: Vienamese tones via syllable 'bao'

Task 2: Vietnamese 'Huyền' tone in a single word

Task 4: Vietnamese 'Nặng' tone in a single word

Figure 5. Different tones in syllable *bao* along with the example of user's attempts (adapted from Van et al. 2021).

However, the experiments on a research prototype of pitch graph visualizer for a tonal language demonstrated that there are issues requiring further efforts at both the level of the digital processing core and the feedback production:

1. The parameters of smoothing interpolation that worked well for English speech may need finer adjustments for Vietnamese;
2. Better visualization and highlighting of particular tones (while used in combination with phrasal visualization) is required.

That is why, in contrast to English phrasal intonation training, for Vietnamese, we had to develop an extension for the pitch visualizer enabling pitch segmentation. Hence, DTW algorithms need to be applied to the segments, not only to the to full-scale pitch graph as Figure 6 illustrates.

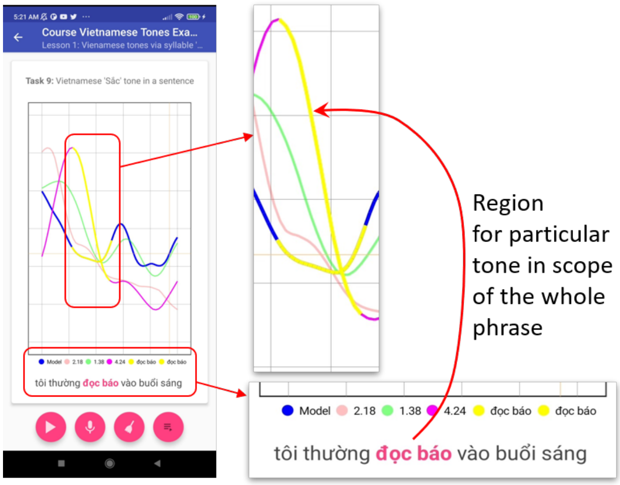

5:21 AM
Course Vietnamese Tones Exa...
Lesson 1: Vienamese tones via syllable '...
Task 9: Vietnamese 'Sắc' tone in a sentence
● Model ● 2.18 ● 1.38 ● 4.24 ● đọc báo ● đọc báo
tôi thường đọc báo vào buổi sáng
Region
for particular
tone in scope
of the whole
phrase
● Model ● 2.18 ● 1.38 ● 4.24 ● đọc báo ● đọc báo
tôi thường đọc báo vào buổi sáng

Figure 6. Segmentation and highlight of tone in the component for Vietnamese.

6.3. *Mora-Timed Language (Japanese)*

For non-tonal languages, the adequate modeling of pitch movements within an utterance helps one to achieve a better connection to the very basic cognitive mechanisms of the language in question. Specifically, though Japanese does not belong to the class of tonal languages, modeling the pitch accent of the high and low tones within words is important to achieve a naturally sounding language melody. Also, Japanese provides a good example of a language where an interface for visualizing rhythmic patterns shows great potential for the provision of instructive feedback to language learners. In order to decide on the necessary extension of visual interface features and feedback to learners, we studied a number of important particularities of the Japanese language.

Toda (2003) lists five modeling examples demonstrating major pronunciation and listening comprehension difficulties for learners of Japanese (see Table 2).

Table 2. Japanese morae, pitch, and tones in examples.

No.	Japanese	Transliteration	English Translation
(1)	来てください。	ki-te-ku-da-sa-i	Please come.
(2)	着てください。	ki-te-ku-da-sa-i	Please wear it.
(3)	切ってください。	ki-tte-ku-da-sa-i	Please cut it off.
(4)	切手ください。	ki-tte-ku-da-sa-i	Please give me stamps.
(5)	聞いてください。	ki:-te-ku-da-sa-i	Please listen.

In Table 2, the segmental content of (1) is the same as (2), and the segmental content of (3) is the same as (4), but the pitches are different. Actually, the standard transliteration does not represent these differences, thus delivering only approximate pronunciation instruction to the learner. On the contrary, the phrases (1) and (3) have the same pitch, but their morae are different: in case (1), there are six morae, while in case (3), there are seven. The model illustrated by the short sound [っ] in cases (1) and (3) is called *soku-on* (そく音), which could be modeled as a staccato-like sound in music. In turn, the pitch and morae in cases (4) and (5) are the same, but there is a tonal difference: the part where the vowel is prolonged is called *chou-on* (ちょう音); the latter, together with *hatsu-on* (はつ音), are "special morae" of the Japanese language.

Table 3 illustrates four cases of *hatsu-on*, with the latter being a nasal sound in the same place of articulation as the following sound (Yamagishi 2008). For example, when the following sound is [k] (like in cases (1) and (2)), the soft palatal nasal [ŋ] is used.

Table 3. *Hatsu-on* variations in Japanese.

No.	Kanji	Hiragana	IPA	English Translation
(1)	天下	てんか	te-ŋ-ka	The whole country
(2)	点火	てんか	te-ŋ-ka	Ignition
(3)	鉄火	てっか	te-k-ka	Lean tuna
(4)	定価	ていか	te-ː-ka	List price

To sum up, in Table 3, cases (1) and (2) are *hatsu-on* examples, the short sound *tsu* [っ] in case (3) represents *soku-on* (the staccato-like model), and case (4) is a long vowel. Hence, though in each of the cases (1) to (4), there are three morae, both the articulation and pitches are different between these examples.

In Japanese, the pitch differs at the mora level, not at the syllable level (Sukegawa 1993). This often causes difficulties for language learners. In Table 3, cases (1) and (2) are not [ten-ka] but [te-ŋ–ka]; thus, [ŋ] necessitates the production of a different pitch compared to [te]. This change can have different directions, either from higher pitch to lower as in case (1) or from lower to higher as in case (2). Figure 7 illustrates this difference.

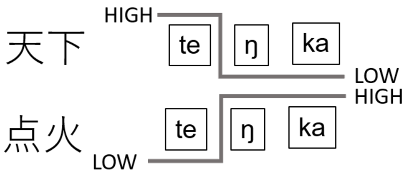


Figure 7. Pitch change in the case of *hatsu-on* sound.

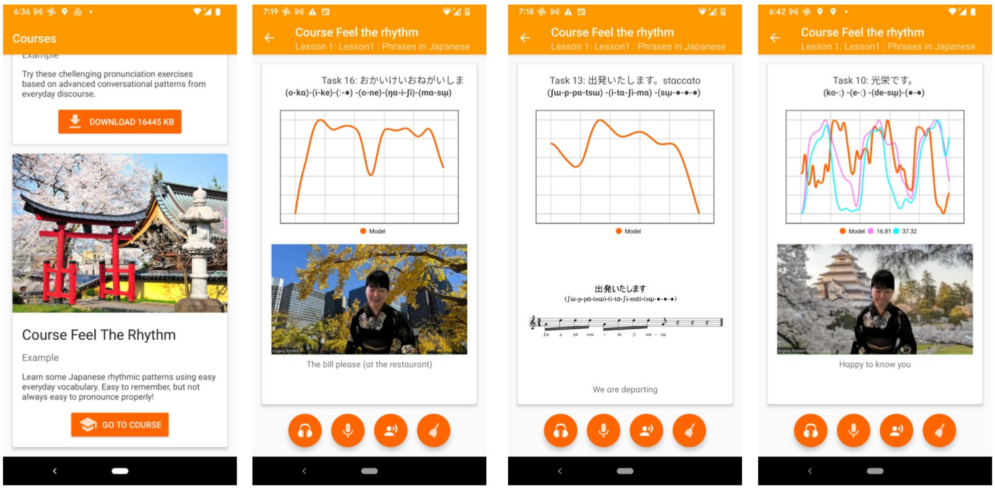
Therefore, the aforementioned cases (4) and (5) from Table 2 as well as cases (1) and (2) from Table 3 all have changes in pitches from low to high or vice versa. Thus, in order to achieve Japanese-like speech, learners themselves need to be aware of the number of morae and the difference in pitches in the particular morae.

Another feature of Japanese that makes it difficult to produce instructive multimodal feedback is the significant difference between the syllable patterns and the rhythmic patterns. Pauses and syncopations allow for the mapping of the syllable pattern to a rhythmic template, which is necessary for meaningful spoken interpretation. Similar to the pitch issues, this mapping could not be clearly represented using IPA transcription. Thus, there is a demand to find other ways to represent such rhythmic templates.

There is a theory that the rhythm of Japanese is mostly four beats (Suzuki 2014; Yamada 1997); thus, musically it could conform to time signatures such as $\frac{2}{4}$ or $\frac{4}{4}$. Japanese language has a constant rhythm assigned to each sentence, a characteristic known as *isochrony* (Yamada 1997). The tick-tock rhythm of a clock and the swinging rhythm of a pendulum are isochronous. The assigned rhythm can be expanded or contracted by a factor of 2 or $\frac{1}{2}$ around this rhythm, but the *shape* of the rhythm remains constant. Figure 8 draws on the application interface illustrating the use of music and extended IPA models to represent the rhythmic patterns of Japanese in conversational contexts.

Though in the majority of practical situations, the rhythmic patterns of Japanese language can be naturally mapped to a two- or four-beat model, the latter is not the only option, especially if we consider the possibility of "meaningful" mapping, which not only takes aspects of the pure pronunciation into the consideration but also the decomposition of the phrase with respect to its semantic elements. Figure 9 illustrates such a case using phonetic transcription along with music notation modeling the rhythm. Additionally, in this figure we can see how the *soku-on* sounds can be modeled using staccato notation.

In principle, rhythm also may help focus attention on keywords, which is important to both tonal and non-tonal languages. As shown in Figure 10, the rhythmic patterns can be retrieved using energy deviations and voice activity detection (VAD) and then displayed either jointly with the phonetic transcription or separately as energy/duration waves.



6:36 Courses

Example

Try these chellenging pronunciation exercises based on advanced conversational patterns from everyday discourse.

DOWNLOAD 16445 KB

Course Feel The Rhythm

Example

Learn some Japanese rhythmic patterns using easy everyday vocabulary. Easy to remember, but not always easy to pronounce properly!

GO TO COURSE

7:19 Course Feel the rhythm

Lesson 1: Lesson1 : Phrases in Japanese

Task 16: おかいけいおねがいしま

(o-ka)-(i-ke)-(:-•) -(o-ne)-(ŋa-i-ʃi)-(ma-sɯ̥)

● Model

The bill please (at the restaurant)

7:18 Course Feel the rhythm

Lesson 1: Lesson1 : Phrases in Japanese

Task 13: 出発いたします。staccato

(ʃɯ-p-pa-tsɯ) -(i-ta-ʃi-ma) -(sɯ̥-•-•-•)

● Model

出発いたします

(ʃɯ-p-pa-tsɯ)-(i-ta-ʃi-ma)-(sɯ̥-•-•-•)

We are departing

6:42 Course Feel the rhythm

Lesson 1: Lesson1 : Phrases in Japanese

Task 10: 光栄です。

(ko-:) -(e-:) -(de-sɯ̥)-(•-•)

● Model ● 16.81 ● 37.32

Happy to know you

(a) Course selection

(b) Model pitch graph along with teacher's video

(c) Model pitch graph along with music notation

(d) Model pitch graph for exercise with repetitions

Figure 8. UI for exercises focusing on the rhythm of Japanese language.

４日から８日までずっとバイトで、休みがないんですよ。（促音）

jokkakaɾa jokamade dzɯttobaitode jasɯmjiŋa naindesɯ̥jo (staccato)

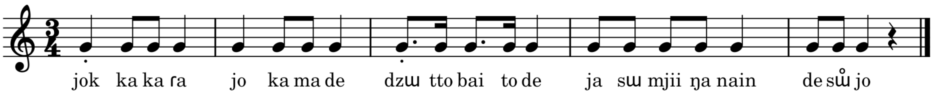


Figure 9. Three beats rhythm and example of soku-on sounds.

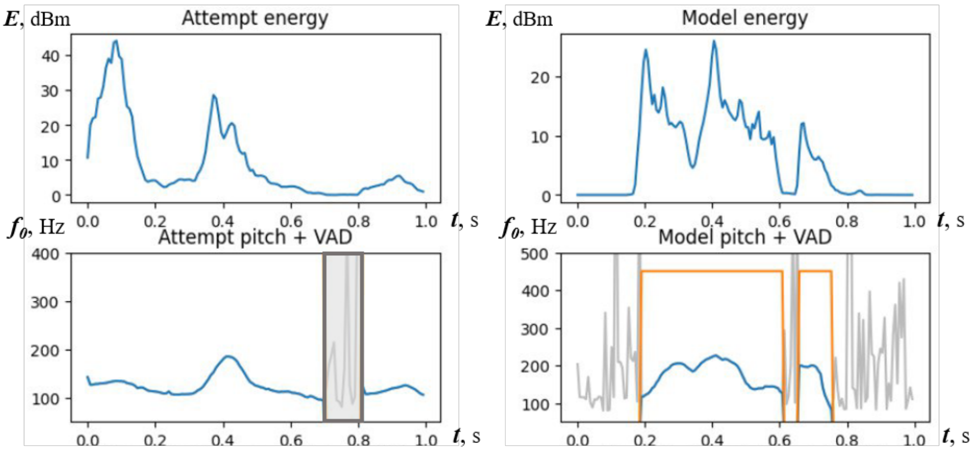


Figure 10. Rhythmic pattern recognition.

7. Challenge 4: Tailoring to Learners

Tailoring the mobile application to individual learners involves providing learners with the most appropriate feedback based on the influence of their first language on the target language, their learning style and learning preferences. Tailoring may be achieved at two levels: automatic detection and user selection. Automatic recognition has the advantage of making judgements based on objective measurements; and, thus, does not pigeonhole learners based on their declarations, but categorizes them based on the prosody produced in the target language.

7.1. First Language (L1)

A central influence on target language pronunciation is the primary language of the learner, which in most cases is the mother tongue or first language (L1) (Li et al. 2022; Ueyama and Li 2020). Given the significant impact that the prosodic features may have on the prosody of the target language, it seems prudent to take this into account in any feedback system. This could be achieved through user declaration in which users state their first language. A more sophisticated solution would be for the system to be able to pinpoint the language family that is closest to any foreign accent that is detected in the target language. For example, Spanish and Portuguese speakers who speak English with traces of their L1 may be categorized as Ibero-Romance accents. Once their language class is identified, their feedback can be tailored to typical issues faced by speakers of that language class. This tailoring is somewhat akin to the machine-learning-based personalization that online retailers use to deliver targeted advertisements to users. However, in our case, the adverts are replaced with targeted advice on improvement (unless, of course, we develop a pay-for-use model).

A system that incorporates such accent detection is therefore able to provide language-family-specific feedback to learners whose accents are more heavily influenced by their mother tongue while providing generic feedback to learners whose accent is not noticeably influenced by their first language. By drawing on accent-recognition algorithms based on the prosody of L1 on the target language, it should be feasible to automatically tailor feedback to language families. With this in mind, Lesnichaia et al. (2022) investigate state-of-the-art accent classification results for speakers who spoke English with Germanic, Romance, and Slavic accents. Their models were based on sparse data from the Speech Accent Archive, which is the same as in many other reported models (Berjon et al. 2021; Ensslin et al. 2017; Singh et al. 2020).

7.2. Learning Styles and Learner Preferences

In the same vein, Khaustova et al. (2023) make use of neural style transfer techniques to identify the L1 background to personalize CAPT feedback in *StudyIntonation*. The concepts of learning styles and learner preferences are related. There are, however, differences between them. Learning style refers to a set of factors that enable users to acquire or learn a language or, in our case, improve speech prosody using our CAPT environment. Learning preferences focus on learning conditions, that is, the factors that create conducive conditions for learning. In short, learning styles relate to how information is processed, while learning preferences relate to the learning conditions and not processing.

Much research has been published on cognitive learning styles (see, for example, Cassidy 2004; Pashler et al. 2008). Visual, auditory, kinesthetic, and tactile are four commonly referred to learning styles, which are covered in initial teacher training courses worldwide. More recently, Baker (2020) argues against using cognitive learning styles as a straitjacket. One of the reasons for this is that learners may not clearly fit into a single category. In fact, some learners may simply draw on learning styles eclectically. The debate over learner-specific learning styles and learning strategies continues unabated (Brown 2023). However, there appears to be no downside to users of CAPT systems in being able to access feedback in different modes, although this comes at a significant time cost to the developers.

Online learning, the use of mobile devices, and the ease of integrating multimodality into web or mobile applications present language learning software developers with the opportunity to tailor materials to individual learning styles, thus providing learners with the possibility to elect (or be automatically categorized as) a particular learning style or combination of styles.

To gain a better understanding of the impact of learning styles on CAPT environments, a study by Blake et al. (2019) investigated the difference between introspective and expert opinion on five learning styles: auditory, visual, kinesthetic, text, and communicative. The results showed that the self-declared learning styles differed from the observed learning style in approximately 30% of learners in the study. One implication from this is that in most cases, learners are able to select the learning style that they use.

Given the difficulty in assessing learning styles and the relative ease and objectivity of controlling learner preferences, research on learning styles has waned, and most research in the field focuses on learner preferences. When applied to *StudyIntonation*, learner preferences could be realized in the choice of media. For example, in the initial implementation of the English environment, the model expressions were presented as audio files. The Vietnamese and Japanese environments use video input. There is no need to provide an option for learners to select audio only since there is no need for them to look at the video screen, but providing the video enables learners who would prefer to watch the movement of the mouth, facial expressions, or accompanying gestures. This embedding of the utterances helps provide learners with more environmental clues about using language.

8. Conclusions

In this article, we have detailed the challenges faced in setting up a multilingual CAPT environment, some of which we were able to overcome (at least in part) and some which are outstanding. The challenge of creating the initial system architecture, pipeline, and fine-tuning algorithms to obtain a working prototype operation involved a great investment of time and resources. However, it enabled us to keep on track to complete the development of the system.

The second challenge to be addressed was to verify the sufficiency of feedback. Once we had confirmed that native speakers were able to act on the information contained in the pitch graphs in order to replicate model utterances, we confirmed that non-native speakers were able to do so. This is where we discovered that not only could non-native speakers accurately replicate the model, but in some cases were able to do so more accurately than the native speakers.

The third challenge and the primary contribution of this paper is the tailoring of the multilingual environment to English, Vietnamese, and Japanese. The personalization and customization of CAPT interfaces rely on accurate automated speech recognition, which is impacted by sociocultural (e.g., geographic and demographic) and idiosyncratic variations. This involves considering not only the differences in the speech prosody of the languages but in considering what information learners need to be given and how this information would best be presented.

The final challenge relates to providing bespoke learning environments based on learner preferences and learning styles. The degree to which personalization may be achieved is dependent on not only technological limitations but on the production of appropriate media resources. For example, the main method of feedback is provided in the form of a pitch graph. Additional video explanations or textual descriptions may address learner preferences; however, creating such resources is time-intensive, while the pay-off in terms of learning effectiveness and efficiency is unknown.

Author Contributions: Conceptualization, J.B. and E.P.; methodology, A.K., N.B. and V.K.; software, I.L., S.L.X., V.N.N., R.S., D.E., A.O. and N.B.P.; validation, V.K. and A.K.; formal analysis, I.L.; music modeling, E.P. and A.K.; phonology modeling. J.B., V.K. and A.K.; writing—original draft preparation, J.B., E.P., A.K. and N.B.; writing—review and editing, J.B. and E.P.; visualization, A.K., N.B. and E.P.; supervision and project administration, E.P.; and funding acquisition, J.B. and E.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Japan Society for the Promotion of Science (JSPS), project "Theory, Methodology and Tools for Tailoring CAPT Feedback", grant number 23K00679.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

Abercrombie, David. 1967. *Elements of General Phonetics*. Edinburgh: Edinburgh University Press. [CrossRef]

Arvaniti, Amalia. 2009. Rhythm, timing and the timing of rhythm. *Phonetica* 66: 46–63. [CrossRef]

Baker, Lewis. 2020. Releasing students from the cognitive straitjacket of visual-auditory kinaesthetic learning styles. *Impact* 3: 57–60.

Bell, Tim, and Tracy Henderson. 2022. Providing a CS unplugged experience at a distance. *ACM Inroads* 13: 26–31. [CrossRef]

Berjon, Pierre, Avishek Nag, and Soumyabrata Dev. 2021. Analysis of French phonetic idiosyncrasies for accent recognition. *Soft Computing Letters* 3: 100018. [CrossRef]

Blake, John, Natalia Bogach, Artem Zhuikov, Iurii Lezhenin, Mikhail Maltsev, and Evgeny Pyshkin. 2019. CAPT tool audio-visual feedback assessment across a variety of learning styles. Paper presented at 2019 IEEE International Conferences on Ubiquitous Computing & Communications (IUCC) and Data Science and Computational Intelligence (DSCI) and Smart Computing, Networking and Services (SmartCNS), Shenyang, China, October 21–23; Piscataway: IEEE, pp. 565–69. [CrossRef]

Bogach, Natalia, Elena Boitsova, Sergey Chernonog, Anton Lamtev, Maria Lesnichaya, Iurii Lezhenin, Andrey Novopashenny, Roman Svechnikov, Daria Tsikach, Konstantin Vasiliev, and et al. 2021. Speech processing for language learning: A practical approach to computer-assisted pronunciation teaching. *Electronics* 10: 235. [CrossRef]

Boitsova, Elena, Evgeny Pyshkin, Takako Yasuta, Natalia Bogach, Iurii Lezhenin, Anton Lamtev, and Vadim Diachkov. 2018. StudyIntonation courseware kit for EFL prosody teaching. Paper presented at 9th International Conference on Speech Prosody 2018, Poznań, Poland, June 13–16, pp. 413–17.

Brown, Steven B. 2023. The persistence of matching teaching and learning styles: A review of the ubiquity of this neuromyth, predictors of its endorsement, and recommendations to end it. *Frontiers in Education* 8: 1147498. [CrossRef]

Carey, Michael David, Arizio Sweeting, and Robert Mannell. 2015. An l1 point of reference approach to pronunciation modification: Learner-centred alternatives to 'listen and repeat'. *Journal of Academic Language and Learning* 9: A18–A30.

Cassidy, Simon. 2004. Learning styles: An overview of theories, models, and measures. *Educational Psychology* 24: 419–44. [CrossRef]

Chun, Dorothy. 1998. Signal analysis software for teaching discourse intonation. *Language Learning and Technology* 2: 74–93.

Collier, René, and J'T Hart. 1975. The role of intonation in speech perception. Paper presented at Structure and Process in Speech Perception: Proceedings of the Symposium on Dynamic Aspects of Speech Perception held at IPO, Eindhoven, The Netherlands, August 4–6; Berlin: Springer, pp. 107–23.

Couper, Graeme. 2021. Teacher cognition of pronunciation teaching: The techniques teachers use and why. *Journal of Second Language Pronunciation* 7: 212–39. [CrossRef]

Datta, Arindrima, Bhuvana Ramabhadran, Jesse Emond, Anjuli Kannan, and Brian Roark. 2020. Language-agnostic multilingual modeling. Paper presented at 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, May 4–8; Piscataway: IEEE, pp. 8239–43.

De Cheveigné, Alain, and Hideki Kawahara. 2002. YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America* 111: 1917–30. [CrossRef]

Ensslin, Astrid, Tejasvi Goorimoorthee, Shelby Carleton, Vadim Bulitko, and Sergio Poo Hernandez. 2017. Deep learning for speech accent detection in video games. Paper presented at AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, Pomona, CA, USA, October 24–28. vol. 13, pp. 69–74. [CrossRef]

Fox, Anthony. 2000. *Prosodic Features and Prosodic Structure: The Phonology of Suprasegmentals*. Oxford: Oxford University Press.

Fusaroli, Riccardo, and Kristian Tylén. 2016. Investigating conversational dynamics: Interactive alignment, interpersonal synergy, and collective task performance. *Cognitive Science* 40: 145–71. [CrossRef]

Hamlaoui, Naima, and Nawel Bengrait. 2016. Using Better Accent Tutor and Praat for learning English intonation. *Arab World English Journal* 3: 99–112.

Hardison, Debra M. 2021. Multimodal input in second-language speech processing. *Language Teaching* 54: 206–20. [CrossRef]

Hermes, Dik J. 1998. Measuring the perceptual similarity of pitch contours. *Journal of Speech, Language, and Hearing Research* 41: 73–82. [CrossRef] [PubMed]

Hirst, Daniel, and Céline de Looze. 2021. Fundamental frequency and pitch. In *The Cambridge Handbook of Phonetics*. Edited by Rachael-Anne Knight and Jane Setter. Cambridge: Cambridge University Press, pp. 336–61. [CrossRef]

Khaustova, Veronica, Evgeny Pyshkin, Victor Khaustov, John Blake, and Natalia Bogach. 2023. CAPTuring accents: An approach to personalize pronunciation training for learners with different L1 backgrounds. In *Speech and Computer*. Edited by Alexey Karpov, K. Samudravijaya, K. T. Deepak, Rajesh M. Hegde, Shyam S. Agrawal and S. R. Mahadeva Prasanna. Cham: Springer Nature Switzerland, pp. 59–70. [CrossRef]

Kim, Heejin, and Jennifer Cole. 2005. The stress foot as a unit of planned timing: Evidence from shortening in the prosodic phrase. Paper presented at Interspeech, International Speech Communication Association, Lisbon, Portugal, September 4–8. pp. 2365–68.

Klapuri, Anssi. 2009. A method for visualizing the pitch content of polyphonic music signals. Paper presented at International Society for Music Retrieval, Kobe, Japan, October 26–30. pp. 615–20.

Kubozono, Haruo. 1989. The mora and syllable structure in Japanese: Evidence from speech errors. *Language and Speech* 32: 249–78. [CrossRef]

Kureta, Yoichi, Takao Fushimi, and Itaru F. Tatsumi. 2006. The functional unit in phonological encoding: Evidence for moraic representation in native Japanese speakers. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 32: 1102. [CrossRef]

Ladd, D. Robert. 2008. *Intonational Phonology*, 2nd ed. Cambridge: Cambridge University Press. [CrossRef]

Lehiste, Ilse. 1970. *Suprasegmentals*. Cambridge: The MIT Press.

Lesnichaia, Mariia, Veranika Mikhailava, Natalia Bogach, Iurii Lezhenin, John Blake, and Evgeny Pyshkin. 2022. Classification of accented English using CNN model trained on amplitude mel-spectrograms. Paper presented at Proceeding Interspeech 2022, Incheon, Republic of Korea, September 18–22. pp. 3669–73. [CrossRef]

Li, Peng, Florence Baills, Lorraine Baqué, and Pilar Prieto. 2022. The effectiveness of embodied prosodic training in L2 accentedness and vowel accuracy. *Second Language Research* 39: 1077–105. [CrossRef]

Lobanov, Boris, Vladimir Zhitko, and Vadim Zahariev. 2018. A prototype of the software system for study, training and analysis of speech intonation. In *International Conference on Speech and Computer*. Berlin: Springer, pp. 337–46.

Martin, Philippe. 2010. Learning the prosodic structure of a foreign language with a pitch visualizer. Paper presented at Speech Prosody 2010—Fifth International Conference, Chicago, IL, USA, May 11–14.

McCarthy, Daniel. 2018. YouGlish.com: A promising tool for pronunciation dictionary lexicography. *Annual Review of Education, Communication & Language Sciences* 15: 81–96.

McCrocklin, Shannon M. 2016. Pronunciation learner autonomy: The potential of automatic speech recognition. *System* 57: 25–42. [CrossRef]

McDermott, Josh H., and Andrew J. Oxenham. 2008. Music perception, pitch, and the auditory system. *Current Opinion in Neurobiology* 18: 452–63. [CrossRef]

Mikhailava, Veranika, Evgeny Pyshkin, John Blake, Sergey Chernonog, Iurii Lezhenin, Roman Svechnikov, and Natalia Bogach. 2022. Tailoring computer-assisted pronunciation teaching: Mixing and matching the mode and manner of feedback to learners. Paper presented at Proceedings of INTED 2022 Conference, Online, March 7–8. Volume 7, pp. 767–73. [CrossRef]

Neri, Ambra, Catia Cucchiarini, and Helmer Strik. 2006. ASR-based corrective feedback on pronunciation: Does it really work? Paper presented at Interspeech 2006, Pittsburgh, PA, USA, September 17–21. https://doi.org/10.21437/Interspeech.2006-543.

Orsucci, Franco, Roberta Petrosino, Giulia Paoloni, Luca Canestri, Elio Conte, Mario A. Reda, and Mario Fulcheri. 2013. Prosody and synchronization in cognitive neuroscience. *EPJ Nonlinear Biomedical Physics* 1: 1–11. [CrossRef]

Pashler, Harold, Mark McDaniel, Doug Rohrer, and Robert Bjork. 2008. Learning styles: Concepts and evidence. *Psychological Science in the Public Interest* 9: 105–19. [CrossRef]

Pennington, Martha C. 2021. Teaching pronunciation: The state of the art 2021. *RELC Journal* 52: 3–21. [CrossRef]

Pennington, Martha C., and Pamela Rogerson-Revell. 2019. *English Pronunciation Teaching and Research: Contemporary Perspectives*. London: Palgrave Macmillan.

Permanasari, Yurika, Erwin H. Harahap, and Erwin Prayoga Ali. 2019. Speech recognition using dynamic time warping (DTW). *Journal of Physics: Conference Series* 1366: 012091. [CrossRef]

Pham, Andrea Hoa. 2004. *Vietnamese Tone: A New Analysis*. Abingdon: Routledge.

Pierson, Catherine. 2015. Forvo: All the words in the world. Pronounced. *Reference Reviews* 29: 29–30. [CrossRef]

Pike, Kenneth L. 1945. *The Intonation of American English*. Ann Arbor: University of Michigan Press, vol. 1.

Ploquin, Marie. 2013. Prosodic transfer: From Chinese lexical tone to English pitch accent. *Advances in Language and Literary Studies* 4: 68–77. [CrossRef]

Rilliard, Albert, Alexandre Allauzen, and Philippe Boula de Mareüil. 2011. Using dynamic time warping to compute prosodic similarity measures. Paper presented at Twelfth Annual Conference of the International Speech Communication Association, Florence, Italy, August 27–31.

Roach, Peter. 2009. *English Phonetics and Phonology: A Practical Course*. Cambridge: Cambridge University Press.

Roudometof, Victor. 2023. Globalization, glocalization and the ict revolution. *Global Media and Communication* 19: 29–45. [CrossRef]

Samad, Ita Sarmita, and Ismail Ismail. 2020. ELSA speak application as a supporting media in enhancing students' pronunciation skill. *Majesty Journal* 2: 1–7. [CrossRef]

Singh, Yuvika, Anban Pillay, and Edgar Jembere. 2020. Features of speech audio for accent recognition. Paper presented at 2020 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems, Durban, South Africa, August 6–7. pp. 1–6. [CrossRef]

Sukegawa, Yasuhiko. 1993. Effects of special beats on the accent of Indonesian learners of Japanese. In *D1 "Group Research Presentation Papers" Research on Priority Areas of the Ministry of Education "Japanese Pronunciations" 1992 Research Results Report*. Washington, DC: Ministry of Education, pp. 167–76.

Suzuki, Tomoyuki. 2014. An objective analysis of Japanese rhythms utilizing the "shibyoushi-ron". *Hitotsubashi Japanese Language Education Research* 2: 95–106.

Sztahó, Dávid, Gábor Kiss, László Czap, and Klára Vicsi. 2014. A computer-assisted prosody pronunciation teaching system. Paper presented at WOCCI, Singapore, September 19, pp. 45–49.

Tallevi, Francesca. 2017. Teaching English Prosody and Pronunciation to Italian Speakers: The KaSPAR Approach. Master's thesis, Politecnico di Milano, Milan, Italy.

Toda, Takako. 2003. Acquisition of japanese special beats by foreign learners (second language acquisition). *Phonetic Research* 7: 70–83.

Tsukada, Kimiko. 2019. Are Asian language speakers similar or different? the perception of Mandarin lexical tones by naïve listeners from tonal language backgrounds: A preliminary comparison of Thai and Vietnamese listeners. *Australian Journal of Linguistics* 39: 329–46. [CrossRef]

Ueyama, Motoko, and Xinyue Li. 2020. An acoustic study of emotional speech produced by Italian learners of Japanese. Paper presented at 10th International Conference on Speech Prosody 2020, Tokyo, Japan, May 25–28. pp. 36–40.

Van, Nhi Nguyen, Son Luu Xuan, Iurii Lezhenin, Natalia Bogach, and Evgeny Pyshkin. 2021. Adopting StudyIntonation CAPT tools to tonal languages through the example of Vietnamese. Paper presented at 3rd ETLTC International Conference on Educational Technology, Language and Technical Communication, Aizuwakamatsu, Japan, January 27–30. SHS Web of Conferences, vol. 102, p. 01007. https://doi.org/10.1051/shsconf/202110201007.

Vásquez-Correa, Juan Camilo, Juan Rafael Orozco-Arroyave, Julián David Arias-Londoño, Jesus Francisco Vargas-Bonilla, and Elmar Nöth. 2016. Non-linear dynamics characterization from wavelet packet transform for automatic recognition of emotional speech. In *Recent Advances in Nonlinear Speech Processing*. Cham: Springer, pp. 199–207. [CrossRef]

Velázquez-López, Diana, and Gillian Lord. 2021. 5 Things to Know about Teaching Pronunciation with Technology. CALICO Infobytes. Available online: https://calico.org/infobytes (accessed on 20 December 2023).

Webber, Charles L., and Norbert Marwan. 2015. *Recurrence Quantification Analysis: Theory and Best Practices*. Cham: Springer.

Yamada, Yoshiro. 1997. Sakano, Nobuhiko, "Unraveling the mystery of the seven-five chorus: Theory of Japanese rhythm". *Bungei Kenkyu* 143: 131–32.

Yamagishi, Tomoko. 2008. Normative awareness of the length of *matsu* in native Japanese speakers: Speakers of the metropolitan dialect and Kinki dialect. *Journal of Phonetics (Phonetic Society of Japan)* 12: 87–97.

Yu, Kristine M. 2014. The experimental state of mind in elicitation: Illustrations from tonal fieldwork. *Language Documentation & Conservation* 8: 738–77.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.