Multimodal Modeling of the Mora-timed Rhythm of Japanese and its Application to Computer-assisted Pronunciation Training

Evgeny Pyshkin*†, Akemi Kusakari‡, John Blake*, Nam Ba Pham§, Natalia Bogach§
*The University of Aizu, Aizu-Wakamatsu, Japan
‡National Institute of Technology, Hachinohe College, Hachinohe, Japan
§Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia
†Email: pyshe@u-aizu.ac.jp

***Abstract*—This paper discusses the setup of mobile components of *StudyIntonation* computer-assisted pronunciation training (CAPT) environment for pronunciation exercises for second language learners of Japanese. Grounded in signal processing, prosody-based speech graph visualization, and pitch quality estimation algorithms, the major focus of this contribution is on implementing the elements of multimodal pronunciation training for Japanese language, which serves as a model of a mora-timed language. We specifically address the approaches to CAPT feedback production with respect to representing mora-timed rhythmic patterns in a way that makes Japanese pronunciation training more enjoyable and more efficiently enables learners to master the mora-timed rhythm.**

***Index Terms*—CAPT, prosody, mora-timed language, Japanese language, mobile application**

An old pond –
A frog dives in
Water sound.

Matsuo Basho, 1686
Quoted from *1020 Haiku in Translation* [1]

I. Introduction

Present-day multimedia-based computerized educational environments for language learning bring together language learning studies and computational approaches. Specifically, automated speech recognition, applied software engineering, and human-computer interfaces are inextricably intertwined with and symbiotically connected to the knowledge scope, research methods, and practical implications from the diverse domains of linguistics, pedagogy, psychology and cognitive sciences. From this, new learning possibilities arise that do not simply engage the digitization of learning materials but create new scenarios, which are impossible to support without using intelligent technologies of knowledge representation and personalization of learning services. In language learning, these possibilities are usually considered as a vital component of intelligent computer-assisted language learning (iCALL) systems.

Specifically, intelligent computer-assisted pronunciation training (CAPT) systems form an important sub-domain of iCALL [2]. Providing second language (L2) learners with opportunities to practice and evaluate the *intelligibly* of their pronunciation is one of the keystone requirements for successful CAPT implementations. Even within the common scenario of listen-and-repeat activities [3], there is much space for improving the models used to produce a well-tailored CAPT system feedback to learners repeating model utterances [4]. Though we know a number of successful CAPT tools addressing both the mastery of phrasal intonation and specific tones, it is still a challenge to produce instructive feedback for the acquisition and assessment of foreign language suprasegmentals [5], [6], particularly with respect to the possibility to personalize feedback modality according to the user preferences regarding learning styles.

It is commonly assumed that prosodic features, such as stress, rhythm, and intonation, are produced physically in the same way regardless of language. This assumption provides the theoretical foundation for developing a multilingual CAPT system based on the same speech processing and feedback production pipeline. However, during the piloting of different languages, we realized the necessity to adapt the system interface and feedback production modes of the multilingual setup of the CAPT environment to support distinctive classes of languages including stress-timed (e. g. English), tonal (e. g. Thai or Mandarin Chinese), syllable-timed (e. g. Korean, Italian or French) and mora-timed (e. g. Japanese).

The center of attention of this work is the CAPT setup for Japanese, a prototypical example of a mora-timed language. For mora-timed languages, in particular, understanding rhythmic divisions of the utterance into portions (known as language *isochrony* [7]) is one of the critical elements of spoken language proficiency. That is why, it is crucial to discover ways to extend the CAPT feedback with features to enable learners to adhere to the rhythmic expectations of mora-timed languages. As mentioned earlier, these features need to address the inherent multimodality of a CAPT system by engaging different learning styles and differences in preferred perception channels among individual learners.

979-8-3503-2422-8/23/$31.00 ©2023 IEEE
DOI 10.1109/IIAI-AAI59060.2023.00044

II. System in Focus

This contribution builds on our ongoing project to develop a multimodal CAPT environment comprising a toolkit that manages mobile applications using speech signal processing, visualization and estimation algorithms. The current system (Figure 1) provides spoken language practice opportunities for language learners, with a particular focus on prosody-based modelling of speech intonation. Learners listen to and shadow contextualized model utterances recorded by native speakers and repeat and record their own attempts at replicating the prosody of the model utterances.

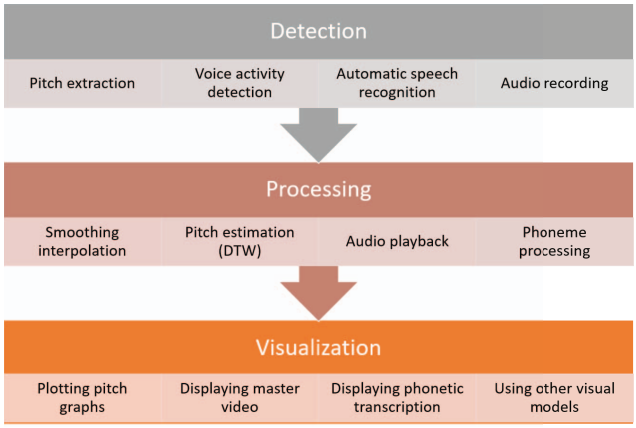


Fig. 1. *StudyIntonation* CAPT system picth processing pipeline.

The pitch curves of the model and the learner attempts are plotted enabling learners to compare and contrast their performance. Such a visualization becomes possible due to the digital signal processing core supporting the fundamental frequency detection algorithms based on methods of digital signal processing.

Feedback is generated from pitch similarity metrics currently constructed using dynamic time warping (DTW), a conventional measure of pitch curve similarity providing tempo invariant estimation; and, therefore, better robustness compared to other practical measures such as the Pearson correlation coefficient and mean square error [8], [9]. The detailed description of the *StudyIntonation* system architecture and applied signal processing algorithms can be found in our previous works [10]–[12] along with positioning of our system in line with existing state-of-the-art solutions belonging to the same class of dedicated CAPT tools.

The project design is generic and based on sound signals rather than particular languages; therefore, there are good grounds for extending the number of supported languages from English (which was the first target language) to other languages. Japanese serves as a prototypical example of a mora-timed language requiring specific efforts for implementing CAPT feedback features that would help learners to master the rhythmic language patterns. The possible learning trajectories are mapped to the courses suggested for selection as shown in Figure 2.

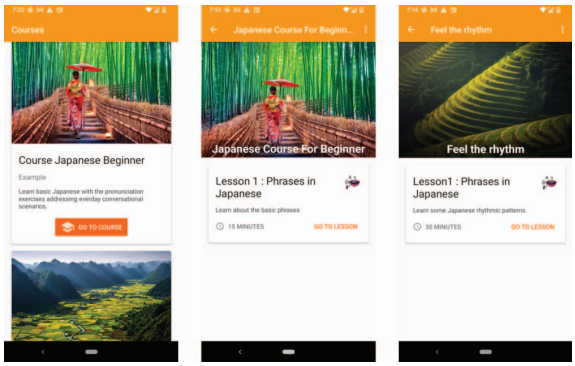


Fig. 2. *StudyIntonation* for Japanese: course selection.

III. Pronunciation Presentation, Visualization, Evaluation for Japanese

As argued in [4], assuring relevant feedback does not only assume fair evaluation of learners' pronunciation but contrasting the evaluation against a benchmark or model to increase its instructive value. Such feedback is not feasible if the conversational and pronunciation exercises (naturally implying the audio materials) are used in isolation; thus, actuating only the auditory perception channel, which is not the leading perception channel for most people. Present-day technology (including mobile solutions) enables harnessing higher levels of multimodality in language learning to address different perception channels, appeal to individual learners' manner, link the feedback mode, and, therefore, address the diversity of learning styles, an important aspect widely discussed in L2 education discourse [2], [11], [13]–[16].

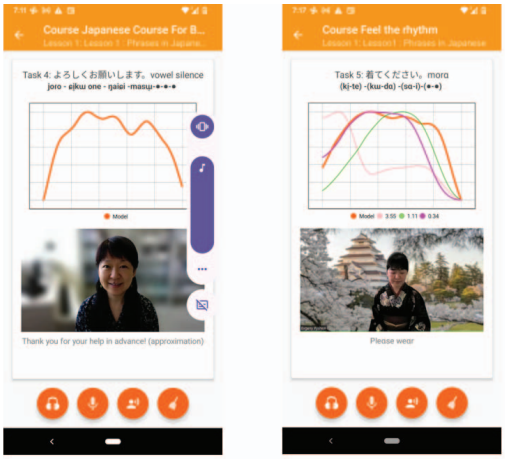
The feedback produced by a CAPT system ought to be considered in two major manifestations: the mode and the manner. First, the mode relates to the CAPT interface features that deliver multimodal pronunciation training and evaluation. Second, the manner focuses on the speech representation models that can enhance and extend the learner experience. The following two sections explore these views in more detail.

IV. Tailoring Interface: Mode

The interface is designed to provide feedback that is contrastive and actionable. This is achieved by providing data to enable learners to notice any deviance from the model pitch curve and modify their output to reduce the deviance.

A. Pitch graph

Intonation speech contour visualized in the form of a colored curve is shown along with the model and learner audio reproduction; thus, enabling the audio-visual feedback coupling. In Figure 3(a) the interface screen is presented demonstrating the pitch graph of the model speech along with the recorded model video, and exercise description, including the Japanese text, phonetic model, and translation. Figure 3(b) illustrates how the gradually improving user's attempts are exposed as pitch graphs along with DTW metrics shown for each recorded attempt.



(a) Model pitch graph is exposed along with video and exercise description

(b) Graphs of user's attempts and DTW scores are shown against the model

Fig. 3. *StudyIntonation* for Japanese: exercises.

Fig. 4. *StudyIntonation* for Japanese: contextualizing the exercises.

B. DTW pitch quality evaluation

DTW provides the primary formal feedback in the form of a numerical score expressing the distance between the model pitch (i. e., the model curve) and the learner's attempt. Smaller values indicate higher proximity to the model. Measuring the distance helps learners to see the quality of their attempts and their progress. However, such scores still lack some instructive value, which is why it is important to find additional approaches to feedback production, expanding the multimodality of the CAPT process.

C. Model speaker video

Incorporating speaker video into the exercise provides learners with the opportunity to perceive and mimic visible, facial, and articulatory movements, such as the shape of the mouth and, in some cases, the position of the tongue.

D. Contextualizing the exercise

There are several major approaches to introducing the exercise context. Adding preceding fragments of conversation (as shown in [17]) can provide additional insights, showing why the same phrase can be pronounced differently depending on the context. Target expressions can be contextualized using familiar scenarios, such as ordering a drink in a cafe. In the case of using video reproduction of model exercises, the context can be presented by the video environment (the examples in Figure 4 illustrate). It makes the pronunciation exercise more fun, and better connects the practice activity to real-life scenarios, the latter being especially important for casual conversations.

V. Tailoring Feedback Production: Manner

One of the most conspicuous challenges of a multilingual setup of a CAPT environment is to provide a corrective and instructive interpretation of the system feedback. The challenge requires language-specific ways of feedback production since learner requirements vary by language.

A. Understanding the language rhythm

As argued in [18], *"radically divergent contexts can share similar musical structures. As musicians know, feeling the sense of rhythm, and sharing rhythmic structures from beyond one's own shores creates a bridge across languages and cultures"*. Many know that traditional Japanese *haiku* comprise 17 syllables split across three lines with the first line consisting of five, the second line seven and the third line five. However, especially for the purposes of authentic spoken language interpretation and *haiku* translation, it is important to understand that the syllable patterns are not the same as rhythmic patterns.

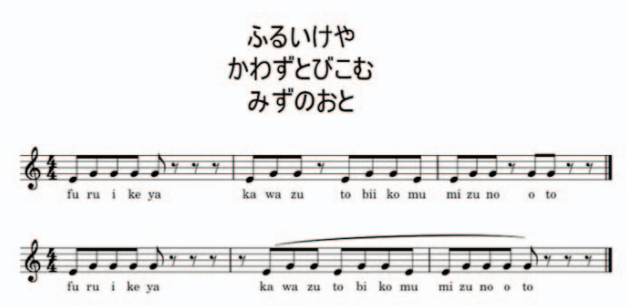


Fig. 5. Pauses and syncopations as exemplified by Basho's frog haiku.

Figure 5 uses the classic example of famous *haiku* by Matsuo Basho (that we dare to put as epigraph to this paper) to illustrate two possible spoken language interpretations of a 17-mora poem: actually, as noted in [18], the superimposition of a 17-mora verse ($5+7+5=17$) on a 24-mora ($8+8+8=24$) rhythmic template means that 7 mora are left "floating" to provide the meaningful pauses, also with a possibility of

syncopation, the latter can often be influenced by one's fashion or taste. In Figure 5, the first interpretation represents the pauses necessary to support the internal rhythmic structure of the pattern, while the second one includes syncopation (in the beginning of the second measure) serving to keep the 12-on (12-unit) phrase together. This small but nice example explains well, how the $5-7-5$ (or any other) syllable structure can be mapped to a four-beat rhythmic template used for meaningful spoken interpretation [19], [20].

B. Phrase repeated in scope of one exercise

Japanese language teachers often recommend that in order to feel natural Japanese language rhythm, the same phrase can be repeated several times consecutively following the suggested time signature and pauses. Hence, even the pauses occurring at the end of a target phrase are important as they enable the phrase to be repeated while maintaining the rhythm.

The only practical way to help learners feel the rhythm of short phrases is to construct exercises requiring repetition of the same phrase several times within one exercise attempt. The same as in music education, rhythmic patterns are easier to recognize and to follow, if the fragment is repeated several times.

Fig. 6. Exercise for learning language rhythm.

Figure 6 shows several interface screens of the exercises focusing on repeatedly pronounced phrases within one exercise attempt. The screens demonstrate the model pitches (displayed in red) along with the learner's multiple attempts recorded and processed by the system. As we can see from the displayed graphs, following the timing of the model speaker is as important as following the intonation pattern.

C. Extended IPA combined with music notation

Traditional IPA phonetic notation is focused on independent sounds and lacks symbolic representations for phrasal rhythm and for changes between high and low pitches. That's why the exercise phonetic transcriptions are displayed using a kind of extended IPA containing special elements representing the pauses and showing the rhythmic structure using the parentheses-dash based organization. However, these extensions (which might still be helpful) provide no good way to show the differences between high and low pitches and are not much visual from the perspective of displaying the rhythmic patterns. We suggest to use the metaphor of music rhythm to make the system feedback more instructive and to better address the multimodality of learning environment. The idea is presented in Figure 7.

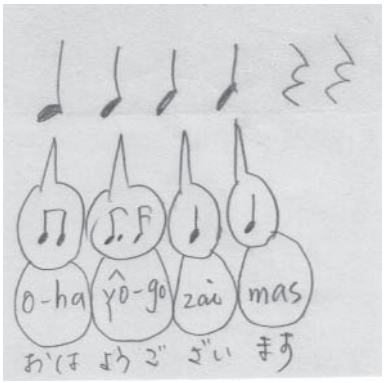


Fig. 7. Metaphor of music rhythm to illustrate Japanese language rhythm: initial conceptualization.

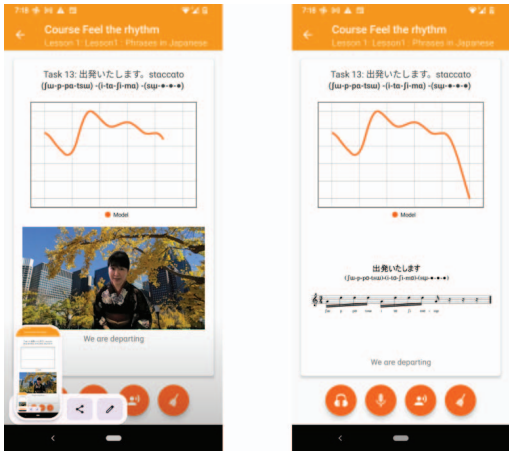


Fig. 8. Using music notation to complement other elements of application interface.

Fragments of music notation displayed along with other elements of exercise presentation, as shown in Figure 8, help to reflect the following "components" of speech melody:

1) **Time signature** – The constant rhythm assigned to each sentence can be naturally represented by a time signature, which helps learners to find an appropriate pace while repeating the exercise phrase.
2) **Pitch high and low tone** – Though Japanese does not belong to the class of tonal languages, modelling the pitch accent of the high and low tones within the utterance is important to achieve a well-sounding language melody.

3) **Mora-connected rhythmic patterns** – The notes help to visually represent the decomposition of the phrase to rhythmic units and also enable visualizing some additional Japanese speech features such as syncopated or staccato sounds.

The following examples from the set of exercises for Japanese pronunciation studies illustrate how music notation can complement the feedback of CAPT environment.

1) Time Signatures: A constant, pendulum-like rhythm is often assumed as an essential property of Japanese spoken language. This language characteristic is usually defined as *isochrony* [21]. In [22], it was suggested that isochronous rhythmic structures of Japanese language can be modelled mostly by using two- or four-bar rhythm, thus musically conforming to time signatures such as $\frac{2}{4}$, $\frac{4}{4}$, or even $\frac{6}{4}$ (as the first music score example presented in Figure 9 illustrates). Though in the majority of practical situations, the rhythmic patterns of Japanese language can be naturally mapped to a two-bar model, other time signatures can be helpful to introduce "meaningful" mapping to language learners (which does not only focus on the pure pronunciation aspects, but on the possible decomposition of the phrase with respect to its semantic elements). Figure 9 illustrates the possibility to model the rhythm of the same phrase by using different phrased decomposition mapped to different time signatures.

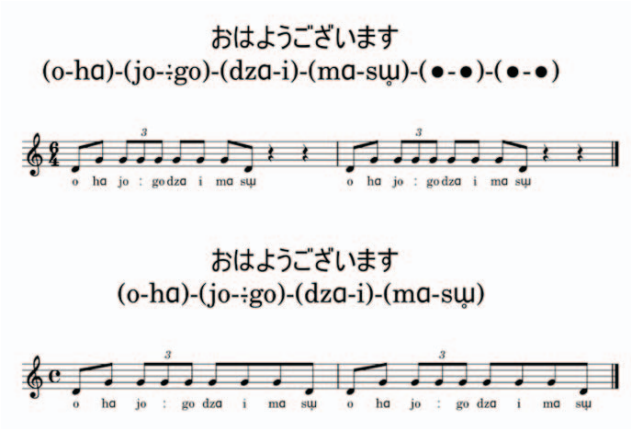


Fig. 9. Same phrases may be mapped to different time signatures.

2) Pitches: Figure 10 draws on musical representation of two modelling examples introduced in [23]. Two sample phrases have the same syllable structure, *kana* definition, and articulation, though the pitches are different. These differences could not be adequately represented by either *kana*, or standard IPA transcription. Thus, music notation can be a promising complementary feature to extend the system feedback to learners by demonstrating the essential rhythmic elements in a way enhancing learners' understanding of what to do to improve.

3) Rest duration and repetitions: Another interesting concept is the inclusion of rests at the end of the phrase in the model. Although after the phrase ends, no voiced elements

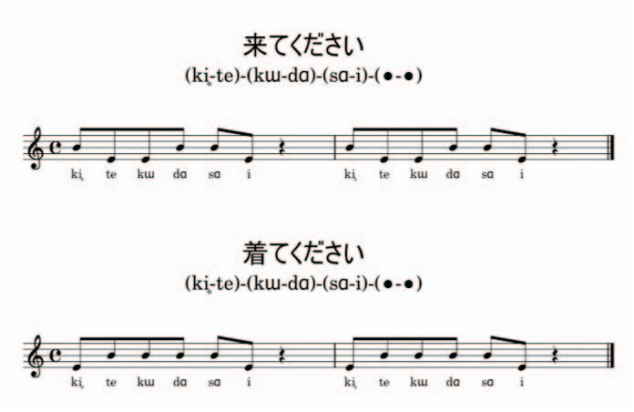


Fig. 10. Phonetically identical phrases need correct pitch accents for conveying the meaning.

are expected, preserving the pauses in the notation helps to implement repetitions. This creates a natural way to follow the suggested timing and to help learners feel the rhythm of the phrase as Figures 11–13 show.

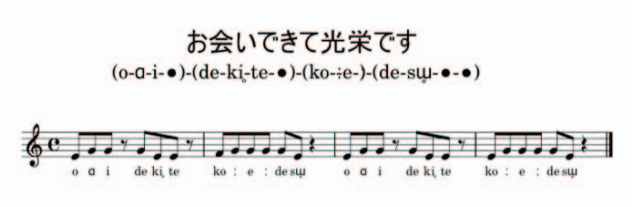


Fig. 11. Rests at the end of phrases are important.

Inserting the rest elements to the IPA notation makes it possible to see the exact moments of silence or "syncopation" while training. However, in order to avoid over-complication of phonetic description (which is already complex), we suggest not to map the phonetic symbols to a specific time duration, keeping the latter property for music score representation (as Figures 12–13 demonstrate).

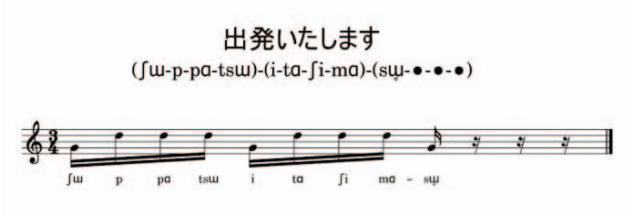


Fig. 12. Rest time in the extended IPA does not need to correspond to specific time duration.

VI. Conclusion

The studies on extending CAPT feedback production mode and manner exemplified by adopting *StudyIntonation* environment for Japanese L2 learning contribute to the broader

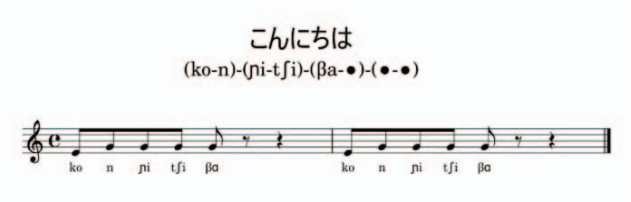


Fig. 13. Feel the rhythm with repetitions and find appropriate timing.

discourse on how feedback on prosody can be improved to address errors in the interlanguage of learners, and to understand the extent to which the current theoretical background supports the ways to improve and customize pitch quality evaluation with respect to speech dynamic assessment. Linking speech visualization to metaphors based on music notation can be considered as a method of visualizing speech dynamics to learners, especially those who are aware of elementary music models.

Using music notation naturally enables graphical representation of such features of mora-based languages as selecting appropriate timing for rhythmic constructions, rests, staccato-like sounds (called *soku-on* in Japanese phonology), short sounds, long vowels, and changes between higher and lower pitches; the latter is especially important in the cases when pitch is connected to meaning disambiguation.

Music notation provides one of additional components of multimodal feedback, which can be particularly helpful for learners with some knowledge in music. Further studies are required to consider other possible rhythm visualizations, for example, a moving timeline commonly seen in music rhythm games (as suggested by one of the reviewers of this paper), which could be an interesting option for learners without any music background.

ACKNOWLEDGEMENT

This work contributes to the project "Theory, Methodology and Tools for Tailoring CAPT Feedback" funded by the Japan Society for the Promotion of Science (JSPS), grant number 23000611.

We are infinitely grateful to our project team members, specifically, Iurii Lezhenin, Roman Svechnikov, Dmitrei Efimov, Andrei Kuznetsov, and Veronica Khaustova who have been working on the *StudyIntonation* project at its different stages. Though they have not been directly involved in this specific study, their efforts in developing, managing, and improving the CAPT environment algorithms and interfaces have been invaluable. Without their contribution to the creation of necessary fundamental project grounds, this particular work would never be possible.

We thank Ayako, Chie, Yuko, and Georgii for their help in recording the pronunciation exercises.

REFERENCES

[1] Basho, B. Yosa, I. Kobayashi, T. Saito, W. R. Nelson, and M. Sakaguchi, *1020 Haiku in translation: The heart of Basho, Buson and Issa*. BookSurge, 2006.

[2] M. C. Pennington and P. Rogerson-Revell, "Using technology for pronunciation teaching, learning, and assessment," in *English Pronunciation Teaching and Research*. Springer, 2019, pp. 235–286.

[3] G. Couper, "Teacher cognition of pronunciation teaching: The techniques teachers use and why," *Journal of Second Language Pronunciation*, vol. 7, no. 2, pp. 212–239, 2021.

[4] V. Mikhailava, E. Pyshkin, J. Blake, S. Chernonog, I. Lezhenin, R. Svechnikov, and N. Bogach, "Tailoring computer-assisted pronunciation teaching: Mixing and matching the mode and manner of feedback to learners," in *Proceedings of INTED2022*, vol. 7, 2022, p. 8th.

[5] D. M. Hardison, "Multimodal input in second-language speech processing," *Language Teaching*, vol. 54, no. 2, pp. 206–220, 2021.

[6] M. C. Pennington, "Teaching pronunciation: The state of the art 2021," *RELC Journal*, vol. 52, no. 1, pp. 3–21, 2021.

[7] M. Nespor, M. Shukla, and J. Mehler, "Stress-timed vs. syllable-timed languages," *The Blackwell companion to phonology*, pp. 1–13, 2011.

[8] A. Rilliard, A. Allauzen, and P. Boula de Mareüil, "Using dynamic time warping to compute prosodic similarity measures," in *12th Annual Conf. of the International Speech Communication Association*, 2011.

[9] Y. Permanasari, E. H. Harahap, and E. P. Ali, "Speech recognition using dynamic time warping (DTW)," in *Journal of Physics: Conference series*, vol. 1366, no. 1. IOP Publishing, 2019, p. 012091.

[10] E. Boitsova, E. Pyshkin, Y. Takako, N. Bogach, I. Lezhenin, A. Lamtev, and V. Diachkov, "Studyintonation courseware kit for efl prosody teaching," in *Proc. 9th International Conference on Speech Prosody 2018*, 2018, pp. 413–417.

[11] J. Blake, N. Bogach, A. Zhuikov, I. Lezhenin, M. Maltsev, and E. Pyshkin, "CAPT tool audio-visual feedback assessment across a variety of learning styles," in *2019 IEEE International Conferences on Ubiquitous Computing & Communications (IUCC) and Data Science and Computational Intelligence (DSCI) and Smart Computing, Networking and Services (SmartCNS)*. IEEE, 2019, pp. 565–569.

[12] N. Bogach, E. Boitsova, S. Chernonog, A. Lamtev, M. Lesnichaya, I. Lezhenin, A. Novopashenny, R. Svechnikov, D. Tsikach, K. Vasiliev *et al.*, "Speech processing for language learning: A practical approach to computer-assisted pronunciation teaching," *Electronics*, vol. 10, no. 3, p. 235, 2021.

[13] S. M. Montgomery and L. N. Groat, *Student learning styles and their implication for teaching*. Centre for Research on Learning and Teaching, University of Michigan, 1998, vol. 10.

[14] G. Molholt and F. Hwu, "Visualization of speech patterns for language learning," in *The path of speech technologies in computer assisted language learning*. Routledge, 2008, pp. 105–136.

[15] P. Martin, "Learning the prosodic structure of a foreign language with a pitch visualizer," in *Speech Prosody 2010*, 2010.

[16] C. Cucchiarini and H. Strik, "Second language learners' spoken discourse: Practice and corrective feedback through automatic speech recognition," in *Smart Technologies: Breakthroughs in Research and Practice*. IGI Global, 2018, pp. 367–389.

[17] E. Pyshkin, J. Blake, A. Lamtev, I. Lezhenin, A. Zhuikov, and N. Bogach, "Prosody training mobile application: Early design assessment and lessons learned," in *10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, vol. 2. IEEE, 2019, pp. 735–740.

[18] R. Gilbert and J. Yoneoka, "From 5-7-5 to 8-8-8: An investigation of japanese haiku metrics and implications for english haiku," *Language Issues: Journal of the Foreign Language Education Center*, vol. 3, no. 1, 2000.

[19] S. Bekku, "Rhythm of japanese language," 1977.

[20] M. Kono, ""haiku" in english education: Based on the presenter's writing of sunshine tm (2010 symposium debriefing)," *Akita Journal on English Literature*, no. 53, pp. 11–16, 2011.

[21] Y. Yamada, "Sakano, Nobuhiko, "unraveling the mystery of the seven-five chorus: Theory of Japanese rhythm"," *Bungei Kenkyu*, no. 143, pp. 131–132, 1997.

[22] T. Suzuki, "An objective analysis of japanese rhythms utilizing the "shibyoushi-ron"," *Hitotsubashi Japanese Language Education Research*, no. 2, pp. 95–106, 2014.

[23] T. Toda, "Acquisition of japanese special beats by foreign learners (second language acquisition)," *Phonetic Research,*, vol. 7, no. 2, pp. 70–83, 2003.