TAILORING COMPUTER-ASSISTED PRONUNCIATION TEACHING: MIXING AND MATCHING THE MODE AND MANNER OF FEEDBACK TO LEARNERS

Veranika Mikhailava[1], Evgeny Pyshkin[1], John Blake[1], Sergey Chernonog[2], Iurii Lezhenin[2], Roman Svechnikov[2], Natalia Bogach[2]

[1]University of Aizu (JAPAN)
[2]Peter the Great St. Petersburg Polytechnic University (RUSSIAN FEDERATION)

Abstract

This paper discusses computer-assisted pronunciation teaching from the perspective of enabling meaningful feedback to learners. We refer to our StudyIntonation project, which is a learning environment that provides feedback on pronunciation exercises to learners based on signal processing algorithms used to construct pitch graphs displayed in a mobile screen, with the support of an audio-visual content repository, and the extensible course developer's toolkit. Interactive mobile tools aim at providing multimodal tailored feedback according to learner preferences. Such feedback includes evaluative and actionable components. Instructive auditory and visual feedback is tailored using interactive personalized features so that learners can better understand where pronunciation is inappropriate and what to do to improve. The provision of visual speech representation in the form of interactive contours of model and learner's pitches has a positive effect on learner's pronunciation of the target language, the latter being an important part of language proficiency. The visual feedback is accomplished by the metrics of the distance between the graphs, based on a dynamic time warping (DTW) algorithm assuring tempo invariant estimation. Though DTW provides an objective primary estimation, we are working on matching the mode and manner of feedback to provide tailored feedback that meets or exceeds learner expectations.

Keywords: computer-assisted prosody training, L2 education, mobile technology, pitch graph, multimodal, tailored feedback, speech processing.

1 INTRODUCTION

In the broad context of human life, assuring relevant feedback to learners does not only mean evaluating the results fairly and putting the evaluation against the known or implied archetypes, but also providing an instructive value that would demonstrate the feasible steps towards future improvements. Though this thesis seems obvious, many evaluation systems used in society, research and education applications often lack this important trait. In turn, in education, successful approaches are (perhaps always) based on finding appropriate ways of interaction between tutors and learners with the help of technology.

Pronunciation instruction is a challenging domain of language learning, where exercises are often considered by learners as tedious and unconstructive [1], and contributing little to the visible progress in language proficiency. A focus on comprehensibility and intelligibility established in late 1990s has resulted in less discussion of segmental and suprasegmental aspects of language except for cases in which pronunciation problems impede effective communication [2][3][4]. Meanwhile, as pronunciation is the first thing that the interlocutor notices, the sociopragmatic outcome of mastering pronunciation has also to be considered. Fortunately, the latest research shows a positive shift of learners' attitude towards the inclusion of pronunciation in language practice [5].

In traditional classes, for pronunciation and conversation the audio materials are often used in isolation, thus, actuating only the auditory perception channel, which is not the leading perception channel for most people. It is commonly agreed and reported by many authors that computer-assisted solutions can help by harnessing higher levels of multimodality including the visual, audial, verbal, and even kinesthetic channels [6][7], which helps support the diversity of learning styles [8][9]. With the advancements of digital and mobile technology enabling better personalized solutions, the extensive use of visual perception linked to audial input and automated speech processing and evaluation has become easier to implement.

The system in focus is StudyIntonation (www.studyintonation.org) – a project on developing a prosody-based computer-assisted teaching environment for phrasal intonation training. It is powered by a stack of digital signal processing techniques for speech activity detection, pitch visual modeling and pronunciation quality evaluation along with the interactive learner interface based on mobile technology [10][11][12].

In terms of design and major features, we can cite other systems similar to StudyIntonation including recent projects, such as WinPitch, learning software based on pitch visualization [13], BetterAccent, English pronunciation training software using Praat [14], IntonTrainer, the system for suprasegmental training in English [15], TI_tobi, the distance L2 English intonation learning environment [16], KaSPAR, an approach to teaching English prosody and pronunciation to Italian speakers [17].

To the best of our knowledge, none of the existing solutions (including ours) completely encompasses the ideal model, which is presenting the feedback to learners in their preferred form, be it graphical, textual or audio feedback, along with the instructive connections between different models. Even very powerful CAPT tools are still lacking explicit multimodal feedback for acquisition and assessment of foreign language suprasegmentals [18]. Our goal is to provide multimodal tailored feedback according to learner preferences. Such feedback includes evaluative and actionable components. Instructive auditory and visual feedback is tailored using interactive personalized features so that the learners can better understand where pronunciation is inappropriate and what to do to improve. Based on our own experience, we want to systematize possible approaches to improve the CAPT system feedback so that it would fit the above-mentioned requirements and refer both to the implemented software available for learners and the research versions in progress involving the features not yet incorporated in the tools accessible by the end users.

2 PROJECT SCOPE AND METHODOLOGY

We have developed a learning environment that provides feedback on pronunciation exercises to learners based on signal processing algorithms. These are used to construct pitch graphs displayed on a mobile screen, with the support of an audio-visual content repository and the extensible course developer's toolkit enabling incorporation of different courses for different languages to the prosody training environment. Figure 1 shows the main elements of the StudyIntonation system architecture and the major steps of pitch processing, visualization and estimation.

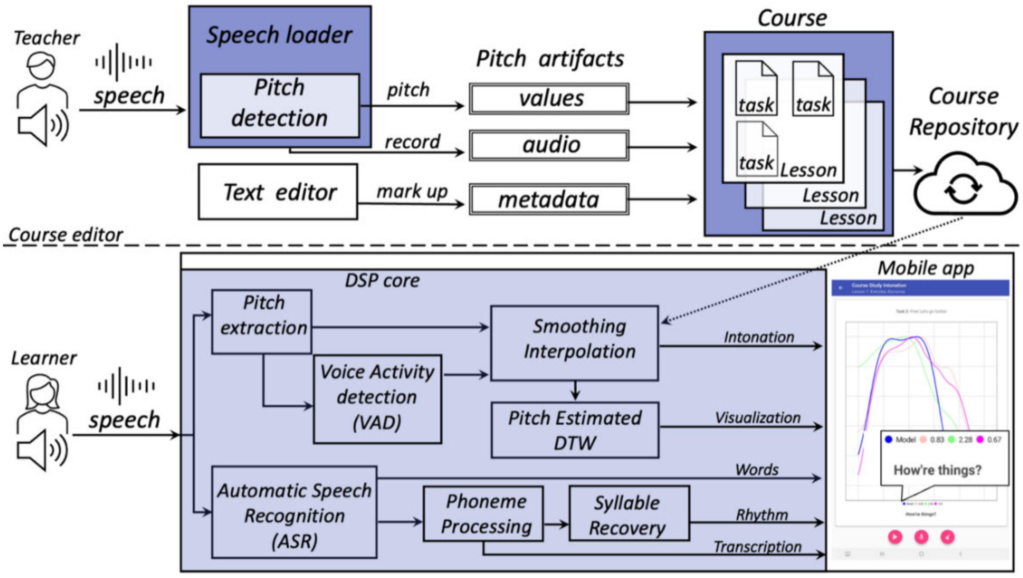


Figure 1. Project workflow.

The provision of visual speech representation in the form of interactive contours of model and learner's pitches has a positive effect on learner's pronunciation of the target language, the latter being an important part of language proficiency. For tonal languages, such as Chinese and Vietnamese, the correct intonation is important at phrasal, lexical and morphemic levels, since conveying the correct meaning is tightly connected to appropriate and accurate tone articulation. Even for non-tonal languages, such as English or Japanese, adequate modeling of tone movements within an utterance helps in achieving better connection to the very basic cognitive mechanisms of language.

Each pronunciation task is defined with model audio recorded by native speakers, its text; the plotted model pitch contour presented to the user on the screen. The app enables learners to try to record their attempts to replicate the pitch and rhythm of the model. The learner attempts are plotted alongside the model to show how closely attempts match the model.

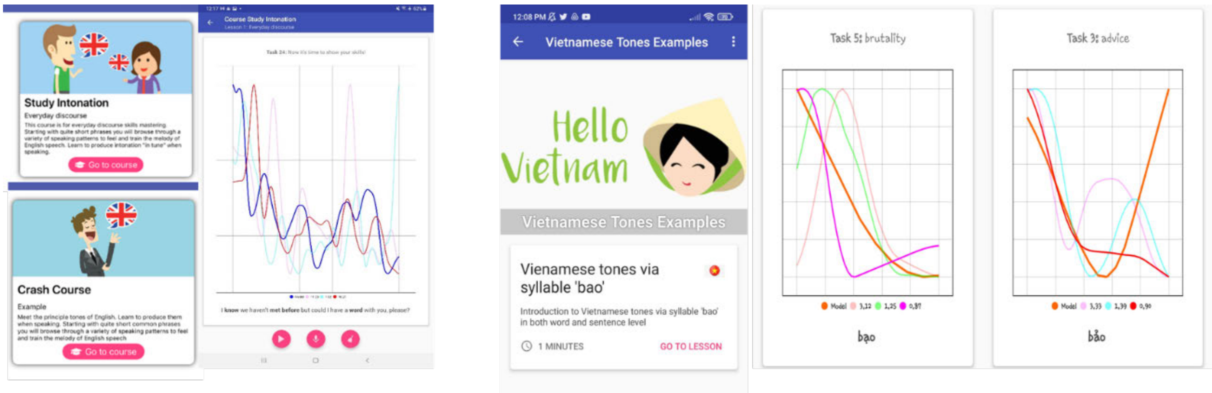
In the course of the project, we address the following aspects:

- The pedagogical soundness of StudyIntonation learning content.
- The learning style and prevalent learner modality influence.
- The adequacy of learner attempts display in the mobile application.
- The consistency between visual feedback and numeric evaluation metrics used.
- The developmental dynamics of learners.

3 FEEDBACK PRODUCTION: “The Mode”

This section discusses the modes of feedback production, which are already implemented and incorporated into the system.

3.1 Pitch Graph Contours with Multiple Attempts Cross-Check



Study Intonation

Everyday discourse

This course is for everyday discourse skills mastering. Starting with quite short phrases you will browse through a variety of speaking patterns to feel and train the melody of English speech. Learn to produce intonation "in tune" when speaking.

Go to course

Crash Course

Example

Meet the principle tones of English. Learn to produce them when speaking. Starting with quite short common phrases you will browse through a variety of speaking patterns to feel and train the melody of English speech

Go to course

Course Study Intonation

Lesson 1: Everyday discourse

Task 24: Now it's time to show your skills!

I know we haven't met before but could I have a word with you, please?

12:08 PM

Vietnamese Tones Examples

Hello Vietnam

Vietnamese Tones Examples

Vienamese tones via syllable 'bao'

Introduction to Vietnamese tones via syllable 'bao' in both word and sentence level

1 MINUTES

GO TO LESSON

Figure 2. Speech contour visualization in mobile app.

Using the intonation speech contour visualization enables the audio-visual feedback coupling that makes possible the significant progress in extending the prosody learning models using the natural multimodality of language learning. 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. Figure 2 shows examples of pitch contour visualization taken from the current exercises in intensive language courses for English and Vietnamese.

Playback of multiple attempts of users can be naturally processed and displayed on the same screen to demonstrate learner progress better.

3.2 Dynamic Time Warping

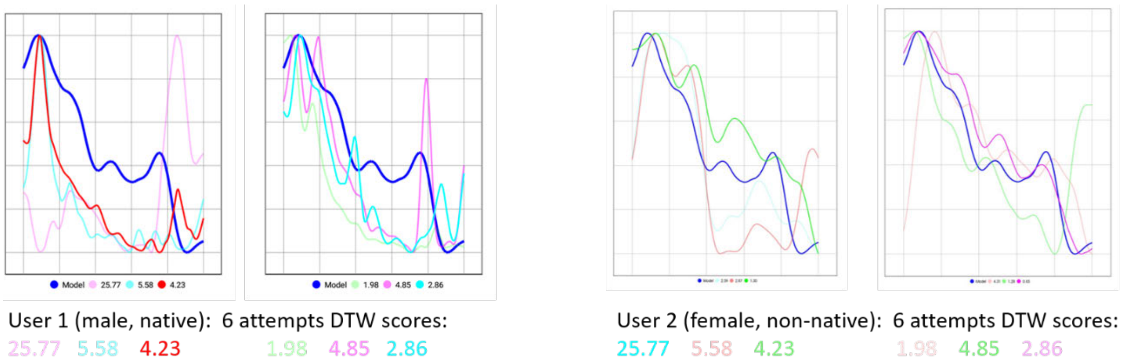
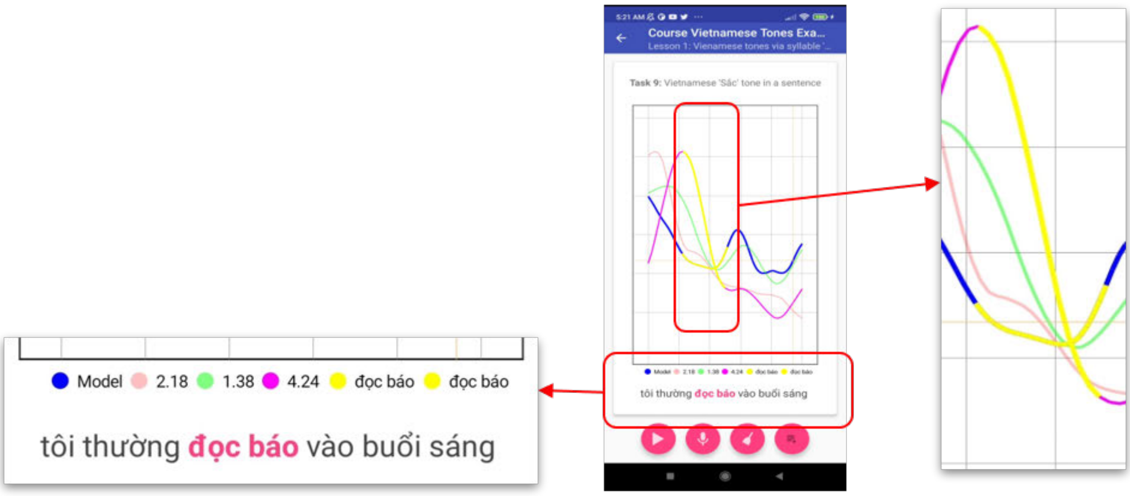


Figure 3. DTW makes the pitch quality estimation more robust.

The visual feedback is accomplished by the metrics of the distance between the graphs, based on a dynamic time warping (DTW) algorithm [19], which assures tempo invariant estimation, and therefore, it is more robust compared to other practical measures such as Pearson correlation coefficient and mean square error [10]. Figure 3 illustrates DTW scores for the series of user attempts for the phrase “How’s the conference going for you?”, how users can adopt their intonation with respect to their pitch graphs and the DTW scores obtained from the system. Though DTW provides an objective primary estimation, we are working on matching the mode and manner of feedback to provide tailored feedback.

3.3 Pitch Graph Segmentation and Segmented Visualization



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

Figure 4. Pitch graph segments: example from a course of Vietnamese.

Segmentation and highlighting the parts of pitch corresponding to relatively independent segments (e.g. particular tones in syllables, stressed word within the longer phrase, elements with higher degree of intonation variability, etc.) can be beneficial to allow users to focus on particular aspects as Figure 4 shows for an exercise from the Vietnamese course [20].

4 Feedback Production: “The Manner”

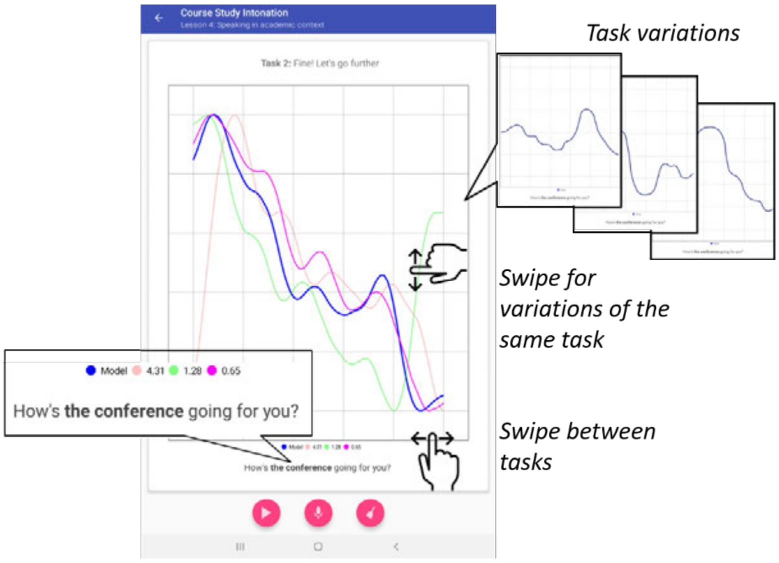
Though pitch graphs with DTW estimation can provide an objective primary estimation, they still lack corrective or instructive interpretation. Therefore, we need metrics that would enable the learners' progress and prosody production estimation. Larsen-Freeman pointed out that to promote L2 learners' spontaneous use of intonation, a systematic view is needed. Complexity Theory and Dynamic System Theory [21] act, at present, as our theoretical basis for second language development. Within a dynamic approach the relationships of phonological features are understood as an interrelated dynamic system, thus, giving a prospect of representation of individual language evolution dynamics.

Non-linear dynamics theory might be quantitatively incorporated in L2 pronunciation teaching using recurrence quantification analysis (RQA) and cross RQA (CRQA) [22][23]. These metrics can contribute to the further improvements in constructing an instructive and more personalized analysis of synchronization between the initial model, the learner, and the referential native speaker’s attempts.

4.1 Stacks of Attitudinal Intonation Deviations

As we mentioned in [11], speakers often choose between the tones (for example, between referring or proclaiming tones) depending on whether there is a known context or the referred information is completely new [24].

Modelling contrastive and attitudinal pronunciation in context can be very helpful in phrasal intonation training. We suggest to implement such a model in the form of presenting a stack of connected tasks referring to the same exercise but demonstrating possible intonation variations. Figure 5 sketches the possible user interface that could support such a feature.



Course Study Intonation

Lesson 4: Speaking in academic context

Task 2: Fine! Let's go further

● Model ● 4.31 ● 1.28 ● 0.65

How's **the conference** going for you?

Figure 5. Tasks with the stacks of alternative intonation exercises.

4.2 Portraying Rhythm

As we argued in [12], rhythm is used by native speakers to focus attention on certain keywords. When the patterns borrowed from a different language are used, the comprehensibility of the message might drop, forcing empathetic native speakers to concentrate intensively to attempt to decode the message. Thus, intonation is tightly connected to meaning, even for non-tonal languages [25]. The rhythmic patterns can be retrieved using energy deviations and displayed either jointly with the phonetic transcription or separately in the form of energy/duration patterns as shown in Figure 6.

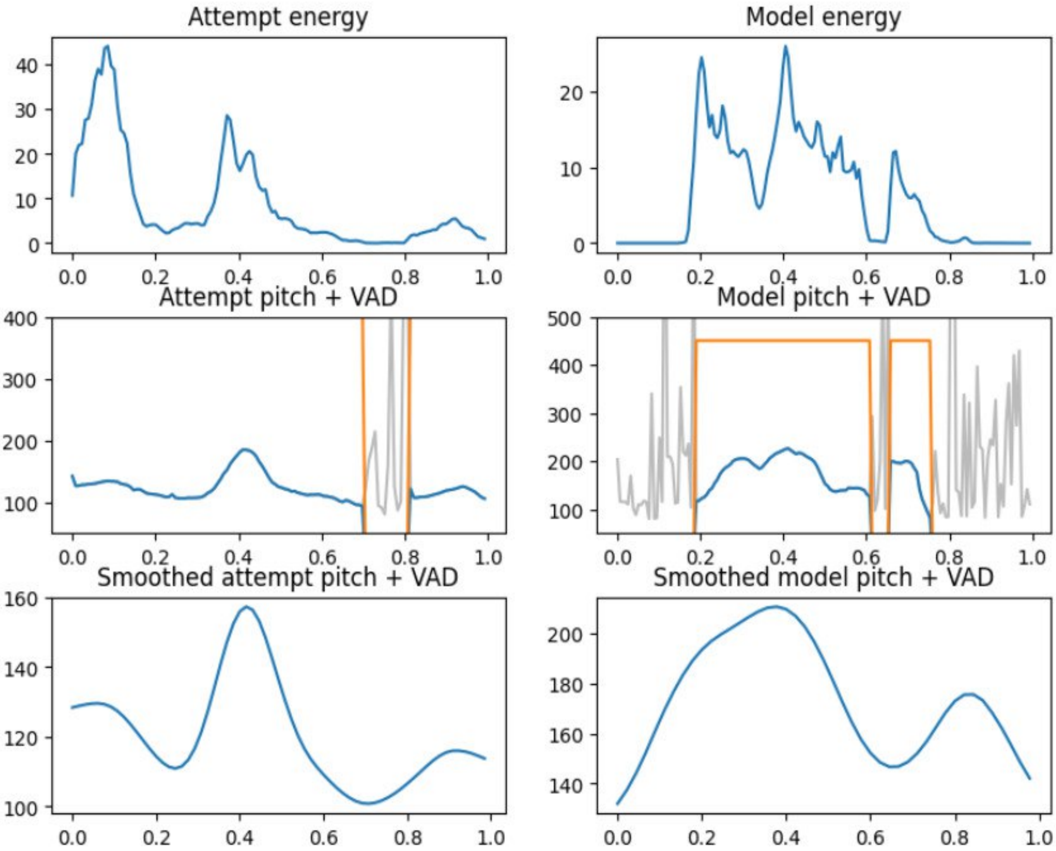


Figure 6. Pitch and energy signals used to reconstruct intonation and rhythmic patterns.

Intonation contours and rhythmic portraits of a phrase along with DTW and other numeral measures can provide learners with a better understanding of how they are successful in following the model patterns.

As mentioned in [26], full-fledged support for tonal languages might require more features in addition to intonation graphs and aggregated pitch quality measures. Visualization of pitch height, its intensity and rhythm may have a serious impact on tone production quality.

5 CONCLUSION

To sum up, this project addresses the domain of developing software for prosody learning using algorithms and data models for prosody recognition, classification, approximation, visualization, and estimation. Effective online learning requires more than the digitization of language learning materials; it requires learners to interact with the study materials. Rapid technology transformation requires further digitalization which is the use of digital technologies and of data in order to transform the processes and create an environment, where the digital technology brings completely new possibilities. The project in focus can serve as a small scale but a good example of such transformation. From the learner perspective, the StudyIntonation environment and the corresponding mobile tools provide an authentic speech context with real-time pitch graphs of speakers, perform learners' speech recording, display model speakers and learners pitch graphs to output a contrastive feedback and calculate speech quality measures aimed at improving the learners' speech progressively. By offering the activities different ways of perception, the system addresses a problem of supporting a diversity of user learning styles.

According to the review [27], an approach used in StudyIntonation is promising in its goal to serve as a mobile-assisted pronunciation training tool for classroom and individual learning purposes, however, there are still open issues to be addressed in the further studies and application releases. Such issues include providing more specific contrastive consistent feedback to users, so that to enable better problem segmentation, easier user self-correction, as well as clear positioning of the approach in scope of relevant pedagogical models. We consider the current contribution as a step in this direction.

ACKNOWLEDGEMENTS

The project has been supported by the Japan Society for the Promotion of Science (JSPS), grant 20K00838.

REFERENCES

[1] J. B. Gilbert, Teaching Pronunciation Using the Prosody Pyramid. Cambridge: Cambridge University Press, 2008.

[2] M.J. Munro and T.M. Derwing, "Foreign accent, comprehensibility, and intelligibility in the speech of second language learners," Language learning, 45(1), pp. 73–97, 1995.

[3] V.A. Murphy, Second Language Learning in the Early School Years: Trends and Contexts-Oxford Applied Linguistics. Oxford University Press, 2014.

[4] D. Liu and M. Reed, "Exploring the complexity of the L2 intonation system: An acoustic and eye-tracking study," Frontiers in Communication, 6, p. 51, 2021.

[5] D. Velázquez-López and G. Lord, G. "5 things to know about teaching pronunciation with technology," CALICO Infobytes, 2021. Retrieved from: https://calico.org/wp-content/uploads/2021/08/CALIB_0821.pdf.

[6] G. Molholt and F. Hwu, "Visualization of speech patterns for language learning," The path of speech technologies in computer assisted language learning, pp. 91–122, 2008.

[7] 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, pp. 367–389, 2018.

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

[9] 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,” 2019 IEEE International Conferences on Ubiquitous Computing Communications (IUCC) and Data Science and Computational Intelligence (DSCI) and Smart Computing, Networking and Services (SmartCNS), pp. 565–569, 2019.

[10] E. Boitsova, E. Pyshkin, T. Yasuta, N. Bogach, I. Lezhenin, A. Lamtev, and V. Diachkov, “StudyIntonation courseware kit for EFL prosody teaching,” In Proceedings of the 9th International Conference on Speech Prosody, pp. 413–417, 2018.

[11] E. Pyshkin, J. Blake, A. Lamtev, I. Lezhenin, A. Zhuikov, and N. Bogach, “Prosody training Mobile application: Early design assessment and lessons learned,” In Proceedings of the 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, IDAACS 2019 2, 735, 2019.

[12] N. Bogach, E. Boitsova, S. Chernonog, A. Lamtev, M. Lesnichaya, I. Lezhenin, A. Novopashenny, R. Svechnikov, D. Tsikach, K. Vasiliev, and E. Pyshkin, “Speech processing for language learning: A practical approach to computer-assisted pronunciation teaching,” Electronics, 10, 235, 2021.

[13] P. Martin, “Learning the prosodic structure of a foreign language with a pitch visualizer,” In Speech Prosody 2010 – Fifth International Conference, 2010.

[14] N. Hamlaoui and N. Bengrait, “Using Betteraccent Tutor and Praat for learning English intonation,” Arab World English Journal (AWEJ), Special Issue on CALL, 14(3), 2016.

[15] B. Lobanov, V. Zhitko, and V. Zahariev, “A prototype of the software system for study, training and analysis of speech intonation,” In Proceeding of the International Conference on Speech and Computer; Springer: Berlin/Heidelberg, Germany, pp. 337–346, 2018.

[16] E. Estebas-Vilaplana, “The teaching and learning of L2 English intonation in a distance education environment: TI_tobi vs. the traditional models,” Linguistica, 57 (1), pp. 73–91, 2017.

[17] F. Tallevi, “Teaching English prosody and pronunciation to Italian speakers: the Kaspar approach,” 2017.

[18] M.C. Pennington and P. Rogerson-Revell, “Using technology for pronunciation teaching, learning, and assessment,” In English Pronunciation Teaching and Research; Springer: Berlin/Heidelberg, Germany, pp. 235–286, 2019.

[19] A. Rilliard, A. Allauzen, and P. Boula de Mareüil, “Using dynamic time warping to compute prosodic similarity measures”, In Interspeech 2011, pp. 2021–2024, 2011.

[20] S. Luu Xuan, “Adopting StudyIntonation CAPT tools to tonal languages through the example of Vietnamese,” Graduation Thesis, University of Aizu, 2021.

[21] D. Larsen-Freeman, “Complexity and ELF,” The Routledge handbook of English as a lingua franca, pp. 51–60, 2017.

[22] D. R. Evans, “Bifurcations, fractals, and non-linearity in second language development: A complex dynamic systems perspective.” Diss. State University of New York at Buffalo, 2019.

[23] D. Liu, and M. Reed, "Exploring the complexity of the L2 intonation system: An acoustic and eye-tracking study," Frontiers in Communication, vol. 6, p. 51, 2021.

[24] R. J. Vilches, “Who is in charge? An L2 discourse intonation study on four prosodic parameters to exert the pragmatic function of dominance and control in the context of L2 non-specialist public speaking,” Complutense Journal of English Studies, vol. 23, pp. 33–58, 2015.

[25] D. Büring, Intonation and Meaning. Oxford University Press: Oxford, UK, 2016.

[26] N. Nguyen Van, S. Luu Xuan, I. Lezhenin, N. Bogach, and E. Pyshkin, “Adopting StudyIntonation CAPT tools to tonal languages through the example of vietnamese,” SHS Web of Conferences, vol. 102, EDP Sciences, 2021.

[27] A. Tan, “Study Intonation: A mobile-assisted pronunciation training application,” Teaching English as a Second Language Electronic Journal (TESL-EJ), 25(3), 2021. Retrieved from: https://tesl-ej.org/pdf/ej99/m3.pdf.