SHS Web of Conferences **102**, 01007 (2021) https://doi.org/10.1051/shsconf/202110201007
ETLTC2021

Adopting StudyIntonation CAPT Tools to Tonal Languages Through the Example of Vietnamese

Nhi Nguyen Van[1,*], *Son* Luu Xuan[1,**], *Iurii* Lezhenin[2,***], *Natalia* Bogach[2,****], and *Evgeny* Pyshkin[1,†]

[1]University of Aizu
[2]Peter the Great St. Petersburg Polytechnic University

Abstract. In tonal languages, tones are associated with both and phonological and lexical domains. Accurate tone articulation is required in order to convey the correct meaning. Learning tones at both word and phrase levels is often challenging for L2 learners with non-tonal language background, because of possible subtle difference between the close tones. In this paper, we discuss an adoption of StudyIntonation CAPT tools to the case of Vietnamese language being a good example of register tonal language with a complex system of tones comprising such features as tone pitch, its length, contour melody, intensity and phonation. The particular focus of this contribution is to assess the adoption of StudyIntonation course toolkit and its pitch processing and visualization algorithms in order to evaluate how the combined use of audio and visual perception mechanisms supported by StudyIntonation may help learners to improve the accuracy of their pronunciation and intonation with respect to tonal languages.

1 Introduction

According to linguists, a language is a synergistic system consisting of individual (but not completely independent) domains identified as phonology, morphology, syntax, semantics and pragmatics [1]. In particular, language intonation may be considered as an associated component of phonology (which is about appropriateness of using phonological patterns while speaking), but at the same type – a component of other domains as well. Language intonation mostly refers to pitch variations at the level of utterance, while language tones are usually discussed within the context of smaller units such as words and morphemes. There are languages (called *tonal* or *tone* languages (such as Vietnamese and Chinese), where tone differentiation is extremely important at both phonological, lexical levels, and pragmatic, hence, accurate articulation of tones is required in order to convey the correct meaning.

*e-mail: nvnhi1811@gmail.com
**e-mail: s1252014@u-aizu.ac.jp
***e-mail: lezhenin@kspt.icc.spbstu.ru
****e-mail: bogach@kspt.icc.spbstu.ru
†e-mail: pyshe@u-aizu.ac.jp

As Orie pointed out in [2], the mastery of language tone plays undeniably important role in tonal language acquisition. However, even for non-tonal languages (such as English), language intonation (particularly, phrasal intonation) is an important aspect of communication when people organize and manage their information *"into discrete pieces which are then worded and formulated grammatically and pronounced in intonation units"* [3]. Furthermore, as a part of his models for discrete intonations (introduced originally for English) Pike defined two major sub-classes of musical tones, namely, contour and register tones[4, 5]. In scope of Pike's classification of musical tones in their application to linguistics, Vietnamese may be considered as an example of register languages, where tones do not rely solely on pitches but also on length, contour melody, intensity and *phonation* as the constituent elements of language register complex [6].

Mastering tones in tonal languages can become a big obstacle for learners with non-tonal language background, because of the learner's unfamiliarity to such subtle differences of tones, their pitches, contours and intensity. Interference with their native language's pitch pattern may cause significant tonal perception errors.

In contrast to intonation, which is likely used in all languages, tones in tonal languages are used to distinguish lexical meaning. According to Yip, although many languages have occasional uses of pitch to change the meaning, in the overwhelming majority of cases, difference in pitches does not lead to the changes in the core meaning of words, unlike the highly tonal languages [2, 7, 8].

2 CAPT Tools for Language Pronunciation Training

Computer-Assisted Prosody Teaching (CAPT) tools integrated with various speech processing technologies make it possible to obtain pitch plots so that to provide a visual representation of the speech. From a number of research works, we learn that, within L2 learning activities, training enhanced by such a visualization of pitch contours has a positive effect on learner's pronunciation (e. g., [9, 10]). In particular, the authors of [11] conducted an interesting study, where they used speech analysis software (Praat) to present a visual display of the Chinese native speaker's pitch curves for learners, then asked learners to record themselves repeating the same words and compare their pitch contours with those of the native speaker.

Various efforts have been made to incorporate speech processing technologies and prosody teaching applications into language learning environments (e. g., [12, 13]), or to implement prosody visual and audio-visual learning environments such as [14, 15], including our own project originally introduced in [16] and described in more details in [17–19].

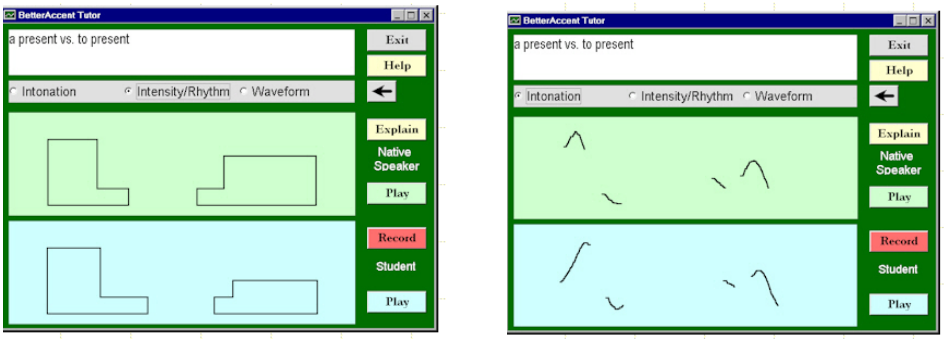
StudyIntonation project, at first, was focused on developing a teaching environment including mobile client apps, supporting audio-visual content repository, designing the course developer's toolkit (CDK) and the course inspector [16]. The approach used in StudyIntonation follows major CAPT principles and addresses different learning styles. Using pitch and speech analysis modules, StudyIntonation implements an idea to create a flexible CAPT tool with a learner-friendly interface. The next step was to suggest an approach to learning framework, where teachers can contribute by creating new courses and improving the contents of the existing ones [17]. An assessment on the early designs of the mobile app prototypes (for Android and iOS platforms) was described in [18], in order to discover the current drawbacks of the system and future necessary steps.

One of possible aspects of further StudyIntonation development and assessment is to investigate how the current features available for teacher and learner may be beneficial for supporting learning workflow for tonal languages such as Vietnamese, and what are necessary enhancements required in order to model various tone features rather than phrasal intonation only. Using our CAPT mobile environment, we are trying to continue assessment stage by arranging a number of experiments based on simple pronunciation tasks in Vietnamese characterized by rich diversity in tones. The objective of this study is to analyze whether the combined use of audio and visual perception help learners improve the accuracy of pronunciation and intonation with respect to tonal languages, and to suggest possible improvements in the approach to pitch/tone visualization.

3 Related Work: Existing Pitch Visualization Techniques

Visualization of prosody acoustic features is supported by a number of existing solutions, including such projects as WinPitch LTL [20], Windows Tool for Speech Analysis WASP, Praat, and BetterAccentTutor [21]. Since the original purpose of Winpitch LTL, WASP and Praat is for speech analysis, these tools are mainly used by phoneticians and speech processing scientists. Nevertheless, their features such as spectrograms, waveforms, and pitch contours display, may be used as reference models in process of designing tools for language learners.

Winpitch LTL enables such features as pitch intensity curves and pitch highlighting with a possibility to speech segment selection. The teacher can highlight the appropriate pitch curve sections for learners to direct their attention to particular parts of pitch, along with presenting main prosodic parameters such as fundamental frequency, intensity, syllable duration, and pauses. Praat is an ongoing open source project comprising many acoustic features similar to WinPitch LTL, except the highlighting function, but having interval segmentation and labelling instead. The latter feature along with a possibility to write scripts by using Praat internal scripting language creates a flexible and efficient environment for providing prosodic information or instructions to learners. In general, these applications still require support from teachers having specialized phonetic knowledge in order to interpret the results properly and to help learners improving their performance based on it.



(a) Intensity/Rhythm (b) Intonation

Figure 1: Pitch visualization in BetterAccentTutor

BetterAccentTutor is a commercially available software that is easier to study independently due to its simple and intuitive user interface. It visualizes the native speaker model's intonation in the form of pitch contours, while intensity and rhythm are shown in the forms of steps, where the length represents the duration and the height serves as the energy of a corresponding syllable's vowel (as shown in Figure 1). After the learner records her utterance, the software provides an audio-visual feedback with the analysis of produced intonation, stress and rhythm in comparison against that of the native speaker's. These visual explanations help learners to identify the most pertinent segments they need to match. Unfortunately, this application is currently limited by American English prosody.

In sum, among the known prosody visualization techniques, pitch contours are considered to be the most common way in terms of intonation representation: it is easy to connect the voice tone rises and falls to the ups and downs on the plot. However, complementary ways to the user feedback production should be considered to implement a prosody visualization model that can help to get a better tailored feedback from the learning tools.

4 StudyIntonation Approach At a Glance

At large, StudyIntonation is a learning environment comprising the Android and iOS mobile applications as well as the course development kit contains the mobile application (MA) and the course development kit (CDK). The CDK allows teachers to develop a series of courses specifically designed for a certain pronunciation training goal. From the learner's perspective, the client mobile tools are aimed at providing a convenient and intuitive tool for tonal and prosodic training. In each exercise, a pitch sample is presented to the user. This sample presents a model audio recorded by a native speaker, along with the text contents and the plotted model pitch contour. After recording the user's attempt, the application displays the user's pitch graphs on top of the model's in order to provide output a contrastive visual feedback. The fundamental frequency f_0 estimation algorithm is used in order to generate the graphical images of the pitch contour (this is a common approach for intonation representation used in many systems). In brief, the pitch processing and visualization process can be summarized as follows:

1. Raw audio input is being pre-processed to eliminate unnecessary points and possible noise.

2. Pre-processed signal is being captured by the pitch detection algorithm YIN, which outputs the pitch reading with timestamps and pitch probabilities.

3. Pitch is being smoothed and interpolated in order to obtain more learner-friendly curves which could allow the user perceiving the visual representation properly.

4. After the interpolation stage, two spline functions: $S^r(t)$ (for the user's recorded pitch) and $S^m(t)$ (for the model pitch) are obtained.

The visual feedback is enforced by displaying the distances between the model and the learner attempts based on dynamic time warping algorithms (DTW), which implements is a conventional approach to measure the distance between two time

(a) Bào (b) Bạo

Figure 2: Example of user's attempts to pronounce (a) *Bào* and (a) *Bạo* words with pitch graphs of model and different attempts as well as DTW score of each attempt

series. Although in [22] it was shown that even simple metrics such as Pearson correlation coefficient (PCC) and Root Mean Square (RMS) may fit the requirements of calculating the difference between two given pitch contours, in the last decade, prosodic similarities were successfully evaluated using DTW [23] shown to be effective at capturing the similar intonation patterns being tempo invariant. Therefore, DTW algorithm is particularly useful if the signals have a similar shape but different argument scale, for example, because of speech rates of the model and the learner.

5 Case study: Vietnamese Tones in StudyIntonation

As a tonal language, Vietnamese has six specific tones that can be assigned to each syllable. For example, the syllable "bao" coupled with different tones correspond to different words (and meanings) such as *bao* "bag", *bão* "storm", *báo* "newspaper", *bạo* "bold". Tones are most easily distinguished by their pitch contour, which can be shown in a clear and friendly manner using our application. Figure 3 demonstrate the model pitches of the syllable *bao* in all different tones shown on the screen of StudyIntonation application.

For tonal languages, pitch graphs are beneficial for learning both phrasal intonation and particular tones and words. One of obvious challenge of tone pronunciation

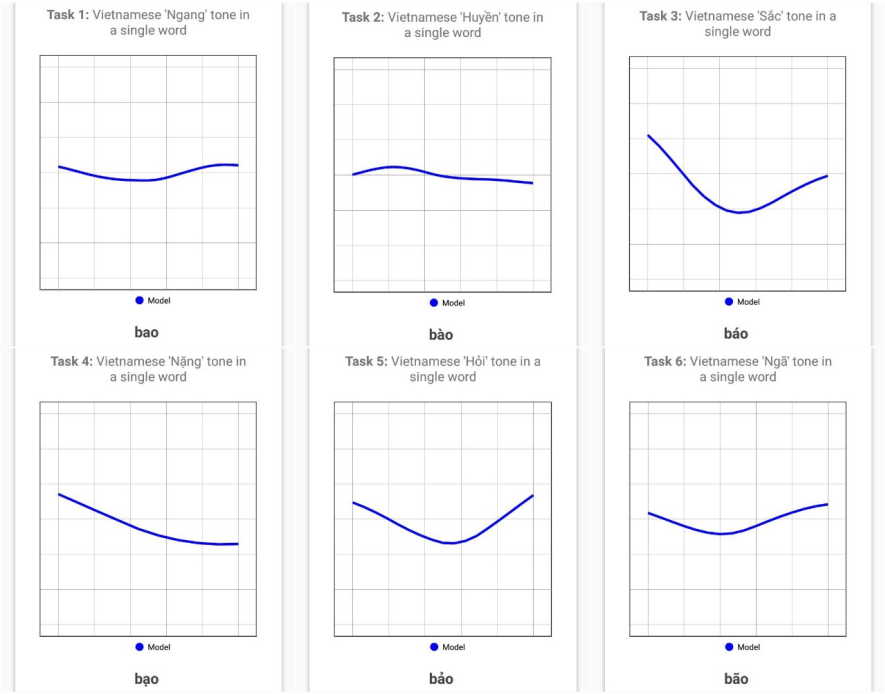


Figure 3: The syllable *bao* with different tones

teaching is that it may be difficult for instructors to provide sufficient verbal explanations on correct pronunciation so that to explore the reasons of the differences between students' tonal production compared to native speakers. Also, without using proper pitch visualization, it may be hard to explain subtle details such as "how low the dip should be in the *bão* tone" or "how flat the neutral tone should be in *bao*".

Learners of Vietnamese need to pay more attention to intonation of an utterance. Vietnamese intonation is generally more "cautiously" used compared to non-tonal languages, so that to avoid possible conflicts between phrasal intonation and lexical tones. Tones in each syllable must remain comprehensible; their relative pitches and shapes must be retained regardless of the phrasal intonation. Furthermore, similar intonation patterns in different languages do not necessarily convey the same meaning. For examples, questions in English usually end with a rising tone, however, in Vietnamese, questions can also end with a falling tone, for example, the questions ending with the sentence particle *à* that has "thanh huyền" tone. Such difficulties may lead to common mistakes of learners with non-tonal language background.

The model's pitch graphs in Figure 4 illustrates examples of different tones in a sentence, with *tôi thường đọc báo vào buổi sáng* ("I usually read newspapers in the morning") for the tone *báo*, *Việt Nam phải hứng chịu rất nhiều cơn bão trong một năm* ("Vietnam experiences many storms a year") for the tone *bão*, and *anh ấy bạo dạn một mình đi vào rừng* ("He went boldly into the woods alone") for the tone *bạo*. In particular, the consistency of each tone's relative shapes within a sentence is demonstrated in the first sentence's pitch graph. The line begins at a considerably high position because of the tone *ngang*'s pitch, then falls down markedly at the tone

huyền, and slightly further with *nặng*. The line later rises as a result of the tone *sắc*'s shape characteristic and falls again for the same aforementioned reason. A little fluctuation was created due to the tone *hỏi* and continue to rise till the end along with the tone *sắc*.

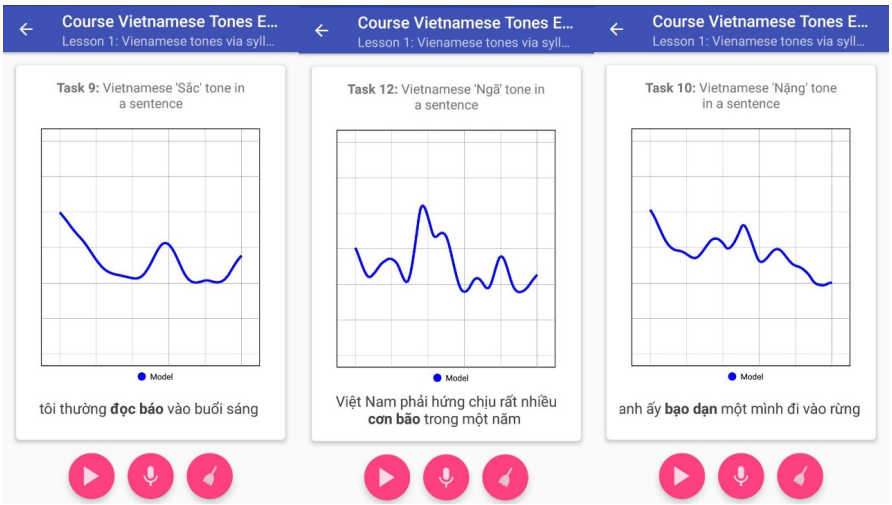


Figure 4: Example sentences of the syllable *bao* with different tones

With the help of visual (and potentially interactive) pitch representation enforced by the pitch quality evaluation (using PCC, MSE or DTW), the learners can have a more objective and, perhaps, more accurate estimation of their language production, thus, having a complementary view in addition to the subjective opinion of their teachers or the learners themselves.

6 Conclusion

Early design assessments demonstrate both the high potential of StudyIntonation environment and the improvements required to create a convenient, intuitive and interactive CAPT environment [18]. Examples that we developed for supporting a crash course of Vietnamese pronunciation demonstrated applicability of existing tools for creating the courses oriented to tonal languages. However, we understand that more experiments must be conducted in order to have a thorough analysis over the full adoption of StudyIntonation to the purposes of learning tonal languages, specifically with respect to assuring production of adequate feedback that would pinpoint the learner's errors and suggest the method to fix it. For full-fledged support for tonal languages, intonation graphs and aggregated pitch quality measures may not be good enough. For example, pitch height and intensity can impact the quality of tone production, therefore incorporating some additional forms of pitch visualization and evaluation into the system may be helpful to produce a robust tailored user feedback.

Another possible improvement regarding tonal production is that the users might have particular problems with certain tone compared to others tones. Thus we need to investigate, how to help learners to overcome such particular problems, for example by creating manually constructed artificial curves in addition to the native speaker patterns.

References

[1] J.B. Gleason, N.B. Ratner, *The development of language* (Merrill Columbus, OH, 1989)

[2] O. Orie, *L2 Acquisition and Yoruba Tones: Issues and Challenges.*, Selected Proceedings of the 36th Annual Conference on African Linguistics,ed. Olaoba F. Arasanyin and Michael A. Pemberton, 121-128. (2006)

[3] P. Tench, *Intonation Unit Boundaries* (Bloomsbury Publishing, 2015), pp. 49–52

[4] K.L. Pike, *The Intonation of American English.* (1945)

[5] K.L. Pike, *Tone Languages: A Technique for Determining the Number and Type of Pitch Contrasts in a Language, with Studies in Tonemic Substitution and Fusion.* (1964)

[6] A.H. Pham, *Vietnamese tone: a new analysis* (Routledge, 2004)

[7] M. Yip, *Introduction* (Cambridge University Press, 2002), p. 1–16, Cambridge Textbooks in Linguistics

[8] W.C. Lin, *Teaching Mandarin Tones to Adult English Speakers: Analysis of Difficulties with Suggested Remedies*, RELC Journal **16**, 31 (1985)

[9] R. Delmonte, *Prosodic tools for language learning*, International Journal of Speech Technology **12** (2009)

[10] M. Eskenazi, *Using a Computer in Foreign Language Pronunciation Training: What Advantages?*, CALICO Journal **16**, 447 (1999)

[11] D.M. Chun, Y. Jiang, J. Meyr, R. Yang, *Acquisition of L2 Mandarin Chinese tones with learner-created tone visualizations*, Journal of Second Language Pronunciation **1**, 86 (2015)

[12] M. Holland, *The Path of Speech Technologies in Computer-Assisted Language Learning*, 1st edn. (Routledge, USA, 2007), ISBN 0415960762

[13] R. Delmonte, *Exploring Speech Technologies for Language Learning* (2011), ISBN 978-953-307-322-4

[14] B. Kröger, P. Birkholz, R. Hoffmann, H. Meng, *Audiovisual Tools for Phonetic and Articulatory Visualization in Computer-Aided Pronunciation Training* (2010), pp. 337–345

[15] O. Niebuhr, M. Alm, N. Schümchen, K. Fischer, *Comparing visualization techniques for learning second language prosody: First results*, International Journal of Learner Corpus Research **3**, 250 (2017)

[16] Y. Lezhenin, A. Lamtev, V. Dyachkov, E. Boitsova, K. Vylegzhanina, N. Bogach, *Study Intonation: Mobile Environment for Prosody Teaching* (2017), pp. 1–2

[17] E. Boitsova, E. Pyshkin, Y. Takako, N. Bogach, I. Lezhenin, A. Lamtev, V. Diachkov, *Studyintonation courseware kit for EFL prosody teaching*, Proceedings of the International Conference on Speech Prosody **2018-June**, 413 (2018)

[18] E. Pyshkin, J. Blake, A. Lamtev, I. Lezhenin, A. Zhuikov, N. Bogach, *Prosody Training Mobile Application: Early Design Assessment and Lessons Learned*, Proceedings of the 2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, IDAACS 2019 **2**, 735 (2019)

[19] J. Blake, N. Bogach, A. Zhuikov, I. Lezhenin, M. Maltsev, 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)* (2019), pp. 565–569

[20] P. Martin, *Learning the prosodic structure of a foreign language with a pitch visualizer* (2010)

[21] N. Hamlaoui, N. Bengrait, *Using Betteraccent Tutor and Praat for Learning English Intonation*, SSRN Electronic Journal (2016)

[22] D.J. Hermes, *Measuring the perceptual similarity of pitch contours*, Journal of Speech, Language, and Hearing Research **41**, 73 (1998)

[23] A. Rilliard, A. Allauzen, P. Boula de Mareüil, *Using Dynamic Time Warping to Compute Prosodic Similarity Measures*, in *INTERSPEECH* (2011), pp. 2021–2024