様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ５ 月 １９ 日現在

機関番号：２１６０２
研究種目：基盤研究(C)（一般）
研究期間：2020～2022
課題番号：２０Ｋ００８３８

研究課題名（和文）Cross-disciplinary approach to prosody-based automatic speech processing and its application to computer-assisted language teaching

研究課題名（英文）Cross-disciplinary approach to prosody-based automatic speech processing and its application to computer-assisted language teaching

研究代表者

Ｐｙｓｈｋｉｎ　Ｅｖｇｅｎｙ（Pyshkin, Evgeny）

会津大学・コンピュータ理工学部・上級准教授

研究者番号：５０７９４０８８

交付決定額（研究期間全体）：（直接経費）　　2,900,000円

研究成果の概要（和文）：この研究では、信号・音声認識と音声処理アルゴリズムに基づくCAPTシステムの進化を探求。デジタル信号処理コアを開発し、ピッチ抽出、音声活動検出、ダイナミックタイムワープに基づくピッチ推定を含む。DTWベースのメトリクス、拡張音声転写、音声およびビデオ出力を含むマルチモーダルフィードバックを提供。評価的、指導的、行動的要素を含むフィードバックを可能にし、さらなる多言語化のための基盤となり得るシステムを評価。

研究成果の学術的意義や社会的意義
The project advances a prosody-based CAPT system using signal and speech processing algorithms for speech visualization and providing a multimodal feedback to learners. Applying the approach to different language groups has a strong impact to improving communication skills of language learners.

研究成果の概要（英文）：We completed a study on the potential of CAPT system advancement based on signal and speech recognition and speech processing algorithms and their customization via computer-aided prosody modeling and visualization instruments.
We developed the digital signal processing core comprising pitch extraction, voice activity detection, pitch graph interpolation, and pitch estimation, the latter based on using dynamic time warping algorithm.
The current implementation supports the transcription and phrasal intonation visualization shown by model and user pitch curves accompanied by a multimodal feedback including DTW-based metrics, extended phonetic transcription, and audial and video output, thus, providing a foundation for further feedback tailoring with evaluative, instructive, and actionable components. The system has been assessed for several languages representing different language groups, thus, creating good ground for further multilingual setup of personalizable CAPT environment.

研究分野：Human-centric software

キーワード：CAPT　prosody　speech visualization　pitch estimation　multimodal feedback

１．研究開始当初の背景 (Background at the beginning of research)

Speech prosody is a complex phenomenon involving suprasegmentals, such as intonation and stress, which, according to cognitive models, exist at the base level of speech production. Adequate mastery of prosody is one of the major factors determining language proficiency and communication culture. Prosody is a superordinate term that encompasses all suprasegmental aspects of speech, including pitch, amplitude, duration and voice quality.

Computer-assisted prosody training (CAPT), a subdomain of computer-assisted language learning (CALL), is a relatively new topic of interest for computer scientists and software developers. Automatic speech recognition (ASR) for the purposes of pronunciation instruction is an important technology making a measurable impact on CALL by enabling the identification of particular parameters of the learner's output and harnessing artificial intelligence to language and speech processing, thus, transforming traditional CALL to intelligent CALL (iCALL). Present-day innovative technology-enhanced CAPT solutions does not only rely on signal processing to transform sound waves using algorithms to enable the sounds to be visualized, but incorporate artificial intelligence, speech processing, and natural language processing in a single CAPT pipeline; and, thus, contribute to the growing domain of iCALL.

However, combined usage of ASR and speech visualization is still considered as a challenging CALL application because of the **complexity of matching wave forms and spectrograms** in a way that enables language learners to act on the **quantitative, qualitative, and instructive feedback**. Stability of prosodic markers enables their measurement, classification and usage for creating prosodically annotated corpora for the purposes of ASR and phonology studies. In sum, our approach addresses three target groups of users (language learners, language teachers, and researchers). It connects four basic ideas.

1. Harnessing **multimodality** in language learning (e.g. perception of interactive audio-visual channels) enforced by multimedia features of mobile devices;
2. Supporting **different learning styles** of students (e.g. visual, auditory and kinesthetic);
3. Advancing signal and **speech processing, visualization and estimation** algorithms to enhance technology-driven education and research;
4. Developing **mobile tools** leveraging rich existing experience of portable device users.

Intonation contour and rhythmic portrait of a phrase provide learners with a better understanding of how they follow the recorded patterns of native speakers. However, such graphs do not presume innate corrective or instructive value. Conventional score-based approach cannot tell the second language learners why their mispronunciations occur and how to correct them. Consequently, adequate metrics to estimate learning progress and prosody production should be combined with CAPT development, while giving more intuitive and instructive feedback.

２．研究の目的 (Purpose of research)

The practical purpose of the project is twofold: first, to develop and assess a technology-driven language learning environment including a course toolkit with end-user mobile applications; and second, to develop tools for speech annotation and semantic analysis based on intonation patterns and digital signal processing algorithms.

The course development kit (which is part of StudyIntonation system) allows teachers to develop a series of courses specifically designed for a certain pronunciation training goal. 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 can be accompanied by a model audio and video recorded by a native speaker, along with the text explanation and the plotted pitch contour. After recording the user's attempt, the application displays the user's pitch graphs against the model, thus, forming a contrastive visual feedback.

In process of project development, the initial model of L2 English pronunciation training has been repurposed support a possibility to extend the number of languages, and to elicit specific features necessary to support adequate interfaces for the languages belonging to different language groups.

３．研究の方法 (Method)

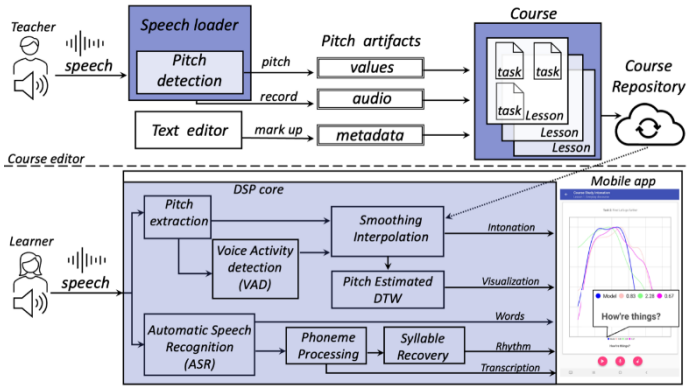

Figure 1. System architecture.

The research builds on the grounds of developing a multimodal CAPT environment comprising a toolkit that manages mobile applications using the cutting-edge 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 modeling speech intonation.

Learners listen to and shadow contextualized model utterances and record their attempt. The pitch curves of both are plotted enabling learners to compare and contrast their performance with the model. Feedback is generated from pitch similarity metrics using dynamic time warping (DTW) as shown in Figure 2. Using a generic model based on sound signals rather than particular languages enabled further adaptation of the system from L2 English learning (original goal) to incorporation of other L2 languages representing major language groups, including non-tonal (English), tonal (Vetnamese), and mora-timed (Japanese) languages.

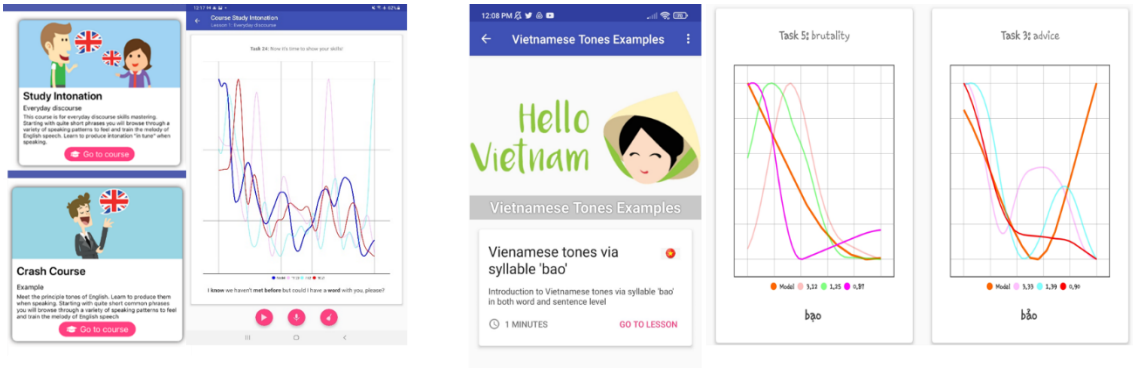

Figure 2. Pitch visualization in mobile app.

The practical implementation (StudyIntonation CAPT environment with mobile end-user interfaces) is the most complex component of the project, involving the following major design, development, and assessment stages:

1. Development of digital signal processing (DSP) core algorithms;
2. Development of audio-visual content repository of pronunciation exercises;
3. Design of extensible course developer's toolkit enabling incorporation of different CAPT courses for different languages;
4. Implementation of proof-of-concept CAPT environment prototypes comprising the server-side software and end-user mobile app components;
5. Assessment of the feedback loop;
6. Analysis of the implemented modes of feedback production for formulating further goals on tailoring CAPT feedback to language learners with a specific focus on actionable and instructive feedback production models.

４．研究成果 (Results)

We completed a study on the potential of CAPT system advancement based on signal and speech recognition and speech processing algorithms and their customization via computer-aided prosody modeling and visualization instruments.

In practical perspective, we developed the digital signal processing core comprising pitch extraction, voice activity detection, pitch graph interpolation, and pitch estimation, the latter based on using dynamic time warping algorithm. We applied voice activity detection (VAD) before pitch processing and instrumented StudyIntonation with a third-party automatic speech recognition (ASR) system. ASR internal data obtained at intermediary stages of speech to text conversion provide phonetic transcriptions of the input utterances

of both the model and learner. The rhythmic pattern is retrieved from phonemes and their duration and energy. Transcription and phrasal rhythm are visualized alongside with phrasal intonation shown by pitch curves. CAPT courseware is reorganized to represent each task as a hierarchical phonological structure which contains an intonation curve, a rhythmic pattern (based on energy and duration of syllables) and IPA transcription. The validity of dynamic time-warping (DTW), which is currently applied to determine the prosodic similarity between models and learners was estimated over native and non-native speakers' corpora using IViE stimuli. DTW results were compared and contrasted against CRQA metrics which measured synchronization and coupling parameters in the course of CAPT operation between model and learner; and were, thus, shown to add to the accuracy of learner performance evaluation through the experiments with two automatic binary classifiers.

Additional developed features include pitch graph contours with multiple attempts cross-check, pitch graph segmentation and segmented visualization (particularly important for tonal languages), as well as a prototype of interface enabling the exercises on attitudinal intonation training.

The project advances the approach to developing a personalized prosody-based CAPT environment based on signal processing and speech processing algorithms for pitch visualization and evaluation aimed at providing a multimodal feedback to learners. Applying the approach to different language groups including non-tonal (English), tonal (Vietnamese) and mora-timed (Japanese) languages has a strong social impact to societal development and improving communication skills of language learners.

As major theoretical contribution, this project helps to understand better how teaching individual segmental and suprasegmental features can positively influence the global construct of L2 pronunciation proficiency. Maximum sensitivity to particular contents of developmental levels means that experiences at those levels yields a maximal effect. The major outcome of our work is our ability to demonstrate how to perform dynamic analysis during the process of phrasal intonation teaching with a CAPT system and how to determine learners' movement from one developmental level to another. Longitudinal and microgenetic analysis of L2 pronunciation development has been conducted based on Vygotskian sociocultural theory concepts, thus, providing rationale for scaffolding learners through their zone of proximal development, which is the region through which learners improve from their actual level to their potential level under guidance and through feedback. The results of these studies have been presented at a highly authoritative Speech Prosody 2022 conference (Figure 3).

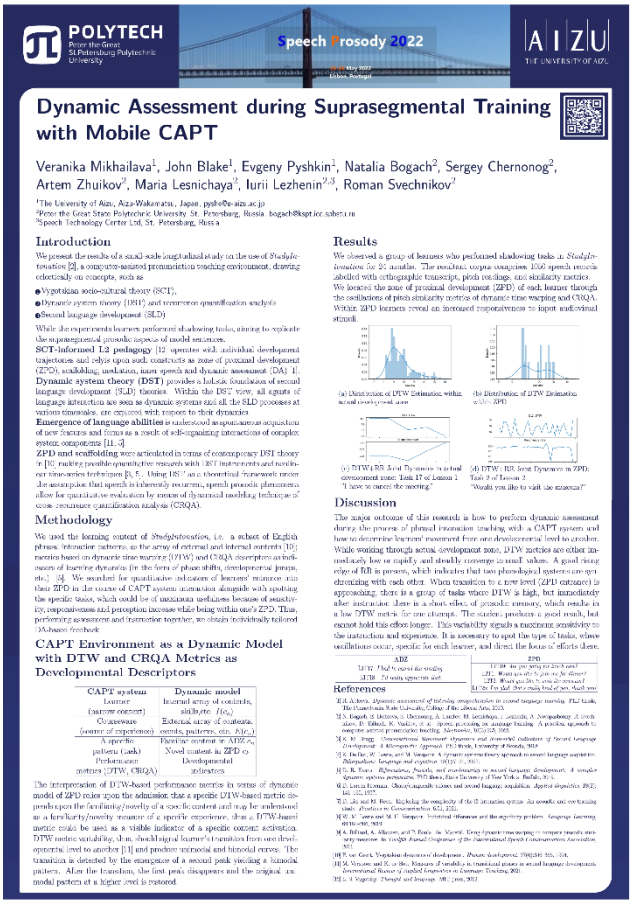

Figure 3. Project presentation at Speech Prosody 2022.

The current results provide grounds for further feedback tailoring with evaluative, instructive, and actionable components. The system has been assessed for several languages representing different language groups, thus, creating good ground for further multilingual setup of personalizable CAPT environment.

Although StudyIntonation enables provisioning the feedback in the form of visuals and some numeric scores, there are still open issues in our design such as (1) metric adequacy and sensitivity to phonemic, rhythmic and intonational distortions; (2) feedback limitations when learners are not verbally instructed what to do to improve; (3) rigid interface when the graphs are not interactive; and (4) the effect of context which produces multiple prosodic portraits of the same phrase which are difficult to be displayed simultaneously. Mobile CAPT tools are supposed to be used in an unsupervised environment, when the interpretation of pronunciation errors cannot be performed by a human teacher, thus an adequate, unbiased and helpful automatic feedback is desirable.

５．主な発表論文等

〔雑誌論文〕　計4件（うち査読付論文　4件／うち国際共著　4件／うちオープンアクセス　4件）

１．著者名 N. Nguyen Van, S. Luu Xuan, I. Lezhenin, N. Bogach, and E. Pyshkin	４．巻 102
２．論文標題 Adopting StudyIntonation CAPT Tools to Tonal Languages Through the Example of Vietnamese	５．発行年 2021年
３．雑誌名 SHS Web Conf.	６．最初と最後の頁 Article 01007
掲載論文のDOI（デジタルオブジェクト識別子） 10.1051/shsconf/202110201007	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 N. Bogach, E. Boitsova, S. Chernonog, A. Lamtev, M. Lesnychaya, I. Lezhenin, A. Novopashenny, R. Svechnikov, D. Tsikach, K. Vasiliev, J. Blake, and E. Pyshkin	４．巻 10 (3), 235
２．論文標題 Speech Processing for Language Learning: A Practical Approach to Computer-Assisted Pronunciation Teaching	５．発行年 2021年
３．雑誌名 Electronics	６．最初と最後の頁 1 - 22
掲載論文のDOI（デジタルオブジェクト識別子） 10.3390/electronics10030235	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 E. Pyshkin and J. Blake	４．巻 11 (3)
２．論文標題 A Metaphoric Bridge: Understanding Software Engineering Education through Literature and Fine Arts	５．発行年 2020年
３．雑誌名 Society. Communication. Education	６．最初と最後の頁 59 - 77
掲載論文のDOI（デジタルオブジェクト識別子） 10.18721/JHSS.11305	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 V. Mikhailava, M. Lesnichaia, N. Bogach, I. Lezhenin, J. Blake, and E. Pyshkin	４．巻 10
２．論文標題 Language accent detection with CNN using sparse data from a crowd-sourced speech archive	５．発行年 2022年
３．雑誌名 Mathematics	６．最初と最後の頁 2913
掲載論文のDOI（デジタルオブジェクト識別子） 10.3390/math10162913	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

〔学会発表〕　計5件（うち招待講演　1件／うち国際学会　5件）

１．発表者名
V. Mikhailava, E. Pyshkin, J. Blake, S. Chernonog, I. Lezhenin, R. Svechnikov, and N. Bogach,

２．発表標題
Tailoring Computer-Assisted Pronunciation Teaching: Mixing and Matching the Mode and Manner of Feedback to Learners

３．学会等名
INTED-2022（国際学会）

４．発表年
2022年

１．発表者名
E. Pyshkin

２．発表標題
“Tailored Fit”: Shaping CAPT Tools Feedback to Language Learners

３．学会等名
ICSEB-2021（招待講演）（国際学会）

４．発表年
2021年

１．発表者名
V. Mikhailava, J. Blake, E. Pyshkin, N. Bogach, S. Chernonog, A. Zhuikov, M. Lesnichaya, I. Lezhenin, and R. Svechnikov

２．発表標題
Dynamic assessment during suprasegmental training with mobile CAPT

３．学会等名
11th International Conference on Speech Prosody 2022（国際学会）

４．発表年
2022年

１．発表者名
M. Lesnichaia, V. Mikhailova, N. Bogach, I. Lezhenin, J. Blake, and E. Pyshkin

２．発表標題
Classification of accented English using CNN model trained on amplitude mel-spectrograms

３．学会等名
Interspeech 2022（国際学会）

４．発表年
2022年

１．発表者名
E. Pyshkin and J. Blake

２．発表標題
Increasing inclusivity: Catering to the needs of socially inactive learners

３．学会等名
Diversity and Inclusivity in English Language Education 2022（国際学会）

４．発表年
2022年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

Study Intonation: English Intonation Training
http://studyintonation.org/

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	Ｍｏｚｇｏｖｏｙ　Ｍａｘｉｍ (Mozgovoy Maxim) (60571776)	会津大学・コンピュータ理工学部・准教授 (21602)	
研究分担者	ＢＬＡＫＥ　Ｊｏｈｎ (Blake John) (80635954)	会津大学・コンピュータ理工学部・上級准教授 (21602)	

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関