CAPTuring Accents: An Approach to Personalize Pronunciation Training for Learners with Different L1 Backgrounds

Veronica Khaustova[1](✉), Evgeny Pyshkin[1](✉), Victor Khaustov[2], John Blake[1], and Natalia Bogach[3]

[1] The University of Aizu, Aizuwakamatsu, Japan
{d8231106,pyshe}@u-aizu.ac.jp
[2] Eyes, JAPAN Co. Ltd., Aizuwakamatsu, Japan
[3] Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia

Abstract. This paper presents a novel approach to addressing the often-overlooked issue of pronunciation instruction in language learning through a Computer-Assisted Pronunciation Training (CAPT) system. While traditional CAPT systems are based on Automatic Speech Recognition (ASR) models trained on native speakers, we argue that this approach results in low accuracy when applied to non-native speakers. To address this limitation, we propose integrating advancements in ASR and accent recognition technology to create a more tailored and effective system. Specifically, our innovation lies in incorporating an accent recognition model into our mobile applications, allowing us to identify learners' first language (L1) backgrounds and subsequently provide personalized exercises and feedback. By doing so, we enable course content creators to design exercises that are linguistically context-aware, and we employ ASR technology to enhance the accuracy of speech detection and accelerate transcription generation during the content creation phase. Furthermore, we make use of neural style transfer techniques to adapt learners' accents before comparing them to reference pronunciations. The evaluation scores are then generated using the Dynamic Time Warping (DTW) algorithm. The key contribution of this paper lies in demonstrating how the integration of ASR-based and accent-targeted solutions can significantly enhance the effectiveness of CAPT systems. This integrated approach offers learners a more precise and personalized learning experience, thereby optimizing pronunciation training.

Keywords: CAPT · ASR · Accent recognition · Personalized feedback

1 Introduction

The study of pronunciation is an essential part of learning to speak a language. However, it has often been a neglected area of focus, leading to a significant

A. Karpov et al. (Eds.): SPECOM 2023, LNAI 14339, pp. 59–70, 2023.
https://doi.org/10.1007/978-3-031-48312-7_5

negative impact on the overall effectiveness of language education. From the learner's perspective, pronunciation exercises are often considered as tedious and nonconstructive [8]; thus, contributing little to the measurable progress of the learner in language proficiency. From the teacher's perspective, studies on speech comprehensibility and intelligibility known since the 1990s have been partially contextualized in a discourse on the segmental and suprasegmental aspects of language and on how pronunciation problems impede effective communication [13,17,18].

Prosody teaching systems, by definition, focus on the suprasegmental features, such as intonation and rhythm patterns, and ignore the segmental features, such as the pronunciation of individual phonemes, e.g. consonant and vowel sounds. To address this problem, Computer-Assisted Pronunciation Training (CAPT) environments integrate Automatic Speech Recognition (ASR) systems, which capture both suprasegmental and segmental pronunciation features to "understand" the words and their component phonemes pronounced by the learner. Although the application of ASR in language learning tools has gained popularity in recent years, its primary limitation is that most ASR models are trained predominantly on data from native speakers. Consequently, its accuracy drops substantially when applied to non-native speakers, diminishing the effectiveness of the feedback provided to learners [5]. This paper seeks to address this gap in accuracy by integrating recent advances in accent recognition and applying transfer learning to state-of-the-art ASR models within a CAPT system.

We address an ongoing project on developing a CAPT system, originally oriented toward English pronunciation learning, but which, nevertheless, has demonstrated sufficient robustness and built-in flexibility to accommodate content creation and interface adjustment for instantiating the system for a variety of target second languages (L2).

One of the key innovations presented in this paper is an approach to integrating the accent recognition and modification models into a mobile application, the latter being the end-user component of our CAPT environment. The idea is to "teach" the system to identify the first language (L1) background of a specific learner, thus creating grounds for personalization of pronunciation exercises. The knowledge of the user's L1 background can not only help to apply learner-specific ASR models, leading to a significant improvement in phoneme detection accuracy, but also allow pronunciation course content creators to deliver more personalized teaching material. For that purpose, we introduce the ability to include L1-tailored exercises into the course, powered by an integrated ASR model for content metadata generation. The inclusion of an accent neutralization model (by means of neural style transfer [21]) modifies the learner's accent to facilitate a more accurate comparison with reference pronunciation using the Dynamic Time Warping (DTW) algorithm [23].

2 Background and Related Work

Computer-Assisted Language Learning (CALL) systems have undergone a remarkable transformation, fueled by advances in computational technology and technology-aware pedagogy [20]. Early versions of CALL systems provided basic drill-and-practice exercises, such as typical listen-and-repeat activities in language laboratories; the limited interaction and feedback capabilities often led to less than satisfactory learning outcomes. As CALL was applied to pronunciation, platforms integrating specialized technologies like ASR were developed to analyze and provide feedback on the pronunciation of learners. These CAPT systems provide more effective and personalized learning experiences.

ASR technology has evolved significantly over the years [12]. Initial attempts at ASR were based on simpler models and relied heavily on handcrafted features. The emergence of deep learning has opened up a new period for ASR, resulting in models that can learn more intricate and abstract representations from data. Advances in applying transfer learning to ASR models for non-native speakers further improved the applicability of ASR to a variety of tasks [25]. These models, trained on carefully curated datasets, provided a marked improvement in performance for language learners with different native languages [27]. Datasets, such as L2-ARCTIC [29], have played a pivotal role in this progression.

Accent recognition and classification is an expanding field in speech technology. The ability to identify and classify a speaker's accent may have multiple applications, ranging from personalized language learning to sociolinguistic research [26]. Recognizing that each learner has a unique speech profile influenced by their native language, researchers have developed methods to adapt ASR models to individual accents, providing more accurate feedback on pronunciation.

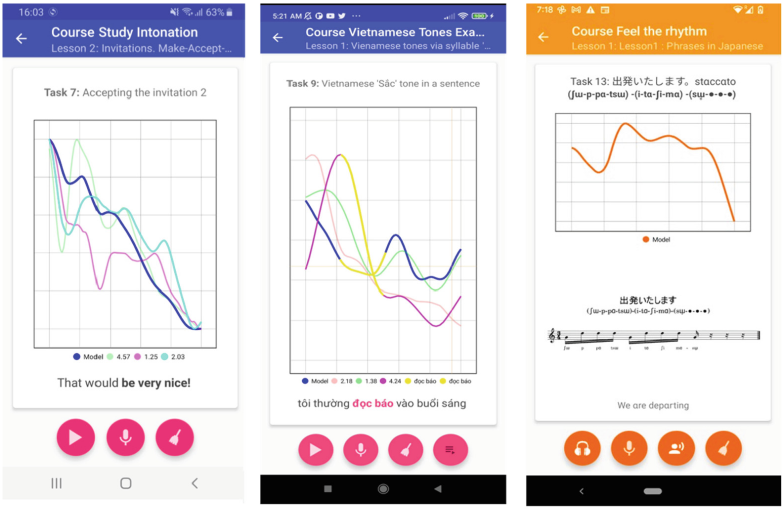
In the literature on language education, the mother tongue L1 has a dominant influence on the accent of the target language L2. But in a broader sense, the personal accent when learning and speaking some language could be significantly influenced by environmental and other factors such as teacher and learning materials, friends and colleagues, country of living, and previously learned languages, all contributing to varying degrees to the formation of an individual accent.

Accent modification [1] and voice conversion [16] have become another challenging area of study, focusing on the modification of a speaker's accent to facilitate better comprehension and evaluation of speech. One pioneering technique in this field is the use of neural style transfer to modify learners' accents [22]. When such modifications are applied to CAPT learner speech, they improve comparison results with reference pronunciation [6]. This, in turn, leads to significant improvements in pronunciation training, allowing learners to receive more accurate feedback and better understand their pronunciation errors.

This study builds on the available technological and pedagogical advancements to enhance the CAPT system in focus by utilizing ASR, accent recognition, and accent neutralization techniques to provide personalized pronunciation feedback to CAPT system users [24].

3 StudyIntonation—CAPT Environment in Focus

This research aims at improving the current components of the existing multimodal multilingual CAPT environment *StudyIntonation* described in detail in


16:03
63%
Course Study Intonation
Lesson 2: Invitations. Make-Accept-...
Task 7: Accepting the invitation 2
● Model ● 4.57 ● 1.25 ● 2.03
That would be very nice!
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
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) -(sɰ-•-•-•)
● Model
出発いたします
(ʃɯ-p-pa-tsɯ)-(i-ta-ʃi-ma)-(sɰ-•-•-•)
ʃɯ p pa tsɯ i ta ʃi ma - sɰ
We are departing

(a) L2: English
Pitch graphs and DTW scores

(b) L2: Vietnamese
UI for tone segmentation and highlight

(c) L2: Japanese
Music notation and extended IPA

Fig. 1. Multimodality in feedback production for different L2.

our previous works [4,15]. The system is a computer-assisted instructional environment that aims to improve the pronunciation skills of learners, with a key focus on prosodic elements, including intonation, stress, and rhythm. The system uses a variety of digital signal processing techniques, such as speech activity identification, pitch visualization modeling using pitch graphs, and evaluation of pronunciation quality.

As shown in Fig. 1, the key interface of the end-user mobile application enables learners to compare their pronunciation pitch graph with a reference model pronunciation (recorded by native speakers). This pitch visualization is accompanied by a number of feedback models that address different levels of system multimodality and various learning styles. In the frame of instantiating the system for a number of target L2 languages representing different language groups (currently, English, Vietnamese, and Japanese), we experimented with the following interface components aiming at tailoring the CAPT feedback to language learners:

- Pitch quality score based on using DTW algorithms [21,23] (known as a robust model to measure the distance between the graphs, which is tempo and scale-invariant);
- Demonstrating a short contextualized video of the exercise (helpful for exercises connected to real-life conversational scenarios);
- Stack of exercise variations (assuming that the same phrase can be trained using a variety of context-dependent intonation patterns);
- Repeated exercise pronunciation patterns (specifically important for mora-timed languages such as Japanese);

- Music notation (helpful for learners with a musical background, especially to represent the language rhythm and higher and lower pitches in mora-timed languages);
- Extended IPA-transcription (with respect to rhythmic units and necessary fragments of silence).

Figure 1 illustrates some of the above-mentioned multimodal feedback interfaces as implemented while instantiating the CAPT environment for different target L2 languages in the process of its multilingual setup. Simultaneously, it is underpinned by an interactive interface designed primarily for mobile devices, tapping into the ubiquity and accessibility offered by today's advanced technology. The application provides a flexible and user-friendly interface designed with learners' needs in mind.

4 Enhancing Pitch Processing Pipeline with ASR Algorithms

In our commitment to offering tailored feedback for learners from diverse L1 backgrounds, we are integrating accent recognition and ASR components into our system. These additions are designed to significantly enhance the personalization of our platform, ensuring a more targeted and effective learning experience for each individual.

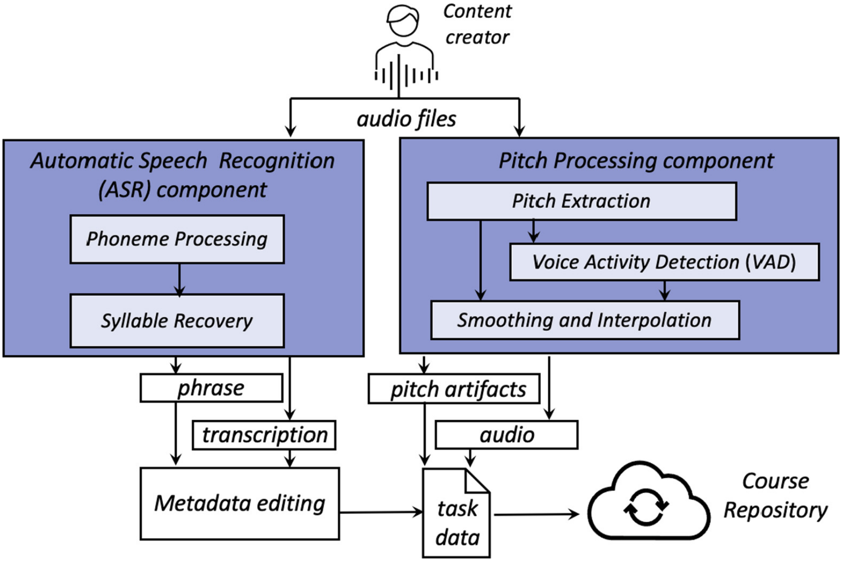


Fig. 2. Workflow of the Course Editor Module.

The key elements of our approach include components such as pitch processing and automatic speech recognition, which are shown in Fig. 2. Pitch processing utilizes the DTW algorithm to provide a tempo-invariant evaluation of the learner's performance. The changes we are implementing concern the neutralization of the accent to produce a more relevant evaluation, as described below. By integrating these different components, our system aims to provide comprehensive and personalized feedback that addresses the specific needs and challenges of each learner.

4.1 Automatic Speech Recognition Component

To achieve efficient processing and robust understanding of diverse accents, we employ transfer learning techniques on an advanced multilingual XLS-R model [2]. This model incorporates the self-supervised learning approach of Wav2Vec2.0 [3] and its capability to learn rich representations from raw audio. Its training on an extensive and varied set of languages enhances its ability to generalize across a wide range of linguistic contexts. It learns contextualized speech representations by randomly masking feature vectors before passing them to a transformer network in the course of self-supervised pre-training. The model is trained to predict the correct speech unit for masked parts of the audio while also learning what the speech units should be. This allows us to capture the nuanced variations of pronunciation across different accents.

To tailor the model to specific accents, we fine-tune the XLS-R model using Connectionist Temporal Classification (CTC) [10] on speech recordings and corresponding transcriptions from the L2-ARCTIC dataset [29], which contains English speech recordings from 24 non-native speakers of six different L1 backgrounds: Arabic, Hindi, Korean, Mandarin, Spanish, and Vietnamese. We resample the audio from 44.1 kHz to the same sampling rate of 16 kHz that was used to pretrain the XLS-R model. We leverage the power of PyTorch [19], an open-source machine learning framework, in tandem with Hugging Face's Transformers library [28], a state-of-the-art natural language processing tool, that provides *Wav2Vec2FeatureExtractor* to process the speech signal to the model's input format, and *Wav2VecCTCTokenizer* to process the model's output into text.

For the training stage, we implement a data collator that dynamically pads training batches to the longest sample in the batch, and use a word error rate (WER) metric, which is common in ASR, to compute the performance of the model. We load a pretrained checkpoint of XLS-R from Hugging Face Hub, freeze the feature extractor that consists of a stack of CNN layers, and add a linear layer on top of the transformer block to classify each context representation into a token class. For the training configuration and hyperparameter tuning, we follow the recommendations from the XLS-R and Wav2Vec2 papers [2,3] by employing the tri-state learning rate schedule: warm-up, constant stage, and decay stage. This approach helps us to refine and optimize the model, making it more attuned to the idiosyncrasies of different spoken accents, ultimately enhancing its precision and usability for our diverse range of learners. Furthermore, the Transformers library allows us to efficiently execute our language models on mobile devices [9], thus ensuring the wide accessibility and seamless operation of our system for users anywhere and anytime.

4.2 Accent Recognition Component

Building on the foundations laid in our previous research [14], the accent recognition module forms a crucial part of our system. During the initial setup of the application, the learner may read a phrase from the Speech Accent Archive [7], so the application can discern the user's native language.

Subsequent to the recording, the system's analysis yields an accent classification which is then presented to the user for verification. Upon user confirmation of the identified accent, the system may suggest downloading the respective fine-tuned ASR model that has been tailored specifically for that accent. This fine-tuned model will facilitate more precise words and phoneme recognition, enabling the system to provide more accurate and personalized feedback to the learner. If the user decides not to record the phrase or select the L1 background from the list, the generic ASR model is used.

This novel approach to incorporating accent recognition into the initial setup process not only enhances the personalization of our system, but also significantly improves the effectiveness of subsequent pronunciation training exercises. It acknowledges the reality of linguistic diversity and responds by ensuring that the application is adapted to the needs of each individual learner right from the outset.

4.3 Course Editor Module Setup

The Course Editor Module (as shown in Fig. 2) is a purpose-built tool for educators, enabling them to create and structure pronunciation courses tailored to their learners' needs. These custom-designed courses are stored in the Course Repository, where they become readily accessible for students to engage in personalized practice. Each course consists of lessons, and each lesson, in turn, comprises a series of tasks (see Fig. 3).

In our quest to continually enhance the Course Editor and expand its capabilities, we are introducing several key features to empower content creators in developing more personalized and effective language learning courses.

Understanding the profound impact that a learner's L1 can have on their English pronunciation and intonation, we are equipping content creators with the means to design specific exercises that cater to learners from a wide variety of L1 backgrounds. This personalized teaching approach aims to tackle the unique pronunciation hurdles each learner might face due to their L1 influence, paving the way for more effective learning outcomes.

In addition, we are harnessing the power of ASR technology to streamline the content creation process. This innovation automates the generation of transcriptions for recorded tasks, significantly reducing the manual workload of content creators. In case of inaccuracies in the automatically generated transcriptions,

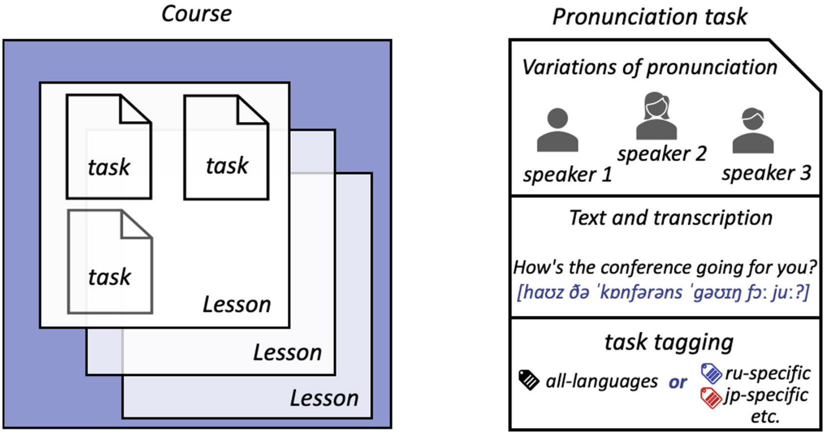


Fig. 3. Course structure.

creators have the flexibility to manually review, edit, and save the corrected version, ensuring the quality and accuracy of the learning materials.

5 Mobile Application

Based on practical possibilities to incorporate the ASR and accent recognition modules into the CAPT environment, we anticipate further extensions of the mobile app interfaces to fit the workflow presented in Fig. 4.

We employ accent recognition and customized ASR models for non-native speakers to help optimize the learning experience tailored to the specific needs of non-native English speakers. This technology helps us to determine the most suitable ASR model to use for each user. Typically, generic ASR models are developed based on the pronunciation patterns of native speakers. However, non-native speakers often exhibit unique pronunciation traits influenced by their L1. The mobile application can use ASR-recognized transcription to highlight the difference between the reference transcription and a recognized one from the learner's recording.

To address this, we refine open-source ASR models for improved accuracy in handling non-native speakers, specifically focusing on those from the most commonly represented L1 backgrounds among the users of the application. By tuning these models to better recognize and understand the pronunciation quirks of different L1 backgrounds, we ensure more accurate and personalized feedback for our users.

One of the challenges in teaching suprasegmental pronunciation is to make sure that the learner not only repeats the intonation correctly but also pronounces the phonemes appropriately. For intonation, the user compares a visualized pitch of their attempt with the reference pitch graph. ASR technology can be deployed directly on the user's mobile device, generating transcriptions of the

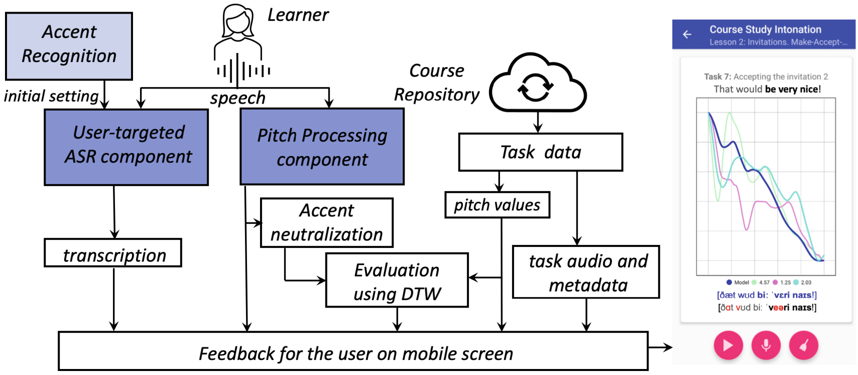

Accent Recognition
Learner
initial setting
speech
Course Repository
User-targeted ASR component
Pitch Processing component
Task data
pitch values
Accent neutralization
transcription
Evaluation using DTW
task audio and metadata
Feedback for the user on mobile screen
Course Study Intonation
Lesson 2: Invitations. Make-Accept-...
Task 7: Accepting the invitation 2
That would be very nice!
● Model ● 4.57 ● 1.25 ● 2.03
[ðæt wʊd bi: ˈvɛri naɪs!]
[ðɑt vʊd bi: ˈveəri naɪs!]

Fig. 4. Workflow of the learning process in the mobile app.

user's spoken utterances. Then, these transcriptions are compared to the reference pronunciation. Any phonemes that differ from the reference are highlighted in red, offering a clear visual indication to the learner of areas where improvement is needed. Figure 4 illustrates the visual feedback of the pitch graphs and the evaluation score extended with the reference and the actual phonetic transcription of the user speech.

Furthermore, accent recognition is employed not only in fine-tuning the ASR but also in categorizing a user's L1 accent. This allows for an even more personalized learning experience, as the system can suggest targeted exercises and provide custom-tailored feedback based on common pronunciation challenges associated with specific L1 accents. By leveraging this technology, we are improving the efficacy of our application, making it an even more powerful tool in the hands of language learners. The courses that provide support for different L1 backgrounds show additional content to practice the corresponding tasks.

We are currently developing a dynamic adaptation module for our system that provides learners with personalized tasks based on their individual performance. This performance is quantified through the use of the DTW evaluation metric—a lower DTW score signifies a higher alignment between the learner's pronunciation and the reference pronunciation, thus indicating better performance.

In cases where a significant discrepancy is detected between the student's pronunciation and the reference standard (indicated by a high DTW score), the system could "intelligently" suggest some additional practice tasks. Such tasks, provisionally added by content creators, are designed to help the learner improve their pronunciation skills in a targeted manner, addressing the areas of difficulty identified through the DTW evaluation. In this way, we aim to foster an adaptive learning environment that tailors instruction to each individual learner's needs, optimizing their language learning journey.

In addition to the possibility of replaying their attempts, users may find helpful an option to adjust the speed of playback. This feature gives learners the opportunity to slow down the audio, making it easier to dissect and understand the intricate phonetic components of the language.

6 Discussion

Enhancing interactive features of a CAPT system, we adapt the learning environment in a way to make it more friendly to users, since the learning process is better tailored to match the individual pace and proficiency level of learners. At the same time, we create more opportunities for teachers (course creators), enabling the promotion of more efficient and learner-centric language acquisition.

Although this contribution is focused on accent-targeted models and L1-specific ASR, it is important to note that these two are not the only feasible ways to produce better personalized CAPT feedback for language learners.

Automatic recognition of the influence of the mother tongue on the target language not only enables CAPT systems to tailor feedback and practice activities to the users but also does so in a discrete yet targeted manner. Accents are often linked to stereotypes, prejudices, and cultural identity, so bypassing the declaration of the influence of an accent helps avoid issues related to, inter alia, privacy, identity, and self-esteem.

The training of ASR models on non-native speaker rather than traditional native speaker datasets enables CAPT systems to move away from the one-size-fits-all model in which all language learners, regardless of mother tongue or language family, are treated the same, viz. as non-native speakers. However, the transition toward targeting specific sets of learners by language family or mother tongue signals a shift to personalized pronunciation training for learners with different L1 backgrounds. We should note that applying transfer learning to the XLS-R model using the L2-Arctic dataset, which contains phrases from out-of-copyright texts from Project Gutenberg books, may decrease the performance of the ASR model for everyday conversations. For that reason, we are working on employing spoken language datasets, such as ICNALE [11].

Although accent recognition plays a valuable role, it is necessary to take into account the complex linguistic landscape in which learners may be exposed to multiple linguistic factors affecting their linguistic repertoire, including the influence of L1 on the target language of the learner.

7 Conclusion

In this paper, we introduce an approach that integrates accent recognition and customized ASR into a CAPT pipeline with the use of ASR-based accent recognition and accent neutralization techniques, along with an approach to design L1-specific exercises and utilizing ASR for transcription generation. A system that incorporates such accent-reflected language-family-specific feedback adjustments could be particularly beneficial for learners whose accents are more heavily influenced by their mother tongue by providing them with implicit but targeted hints on pronunciation improvement.

In conclusion, it is important to note that our CAPT environment is mainly aimed at creating better conditions for the evolution of learners' conversational skills by replicating modeled pronunciation rather than by focusing on the mistakes of the learners. From this point of view, adopting accent recognition techniques to a CAPT system is considered a promising component to be used in conjunction with other approaches towards better CAPT feedback customization, including contextual feedback, enhanced visualization techniques, and multimedia integration.

More experimental and analytical studies are required to evaluate and assess the suggested models in pedagogical practices, including using independent techniques to evaluate learners' progress.

Acknowledgements. This research was funded by the Japan Society for the Promotion of Science (JSPS), grant number 23K00679.

References

1. Aryal, S., Gutierrez-Osuna, R.: Can voice conversion be used to reduce non-native accents? In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7879–7883. IEEE (2014)
2. Babu, A., et al.: Xls-r: self-supervised cross-lingual speech representation learning at scale. arXiv preprint arXiv:2111.09296 (2021)
3. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: a framework for self-supervised learning of speech representations. Adv. Neural Inf. Process. Syst. **33**, 12449–12460 (2020)
4. Bogach, N., et al.: Speech processing for language learning: a practical approach to computer-assisted pronunciation teaching. Electronics **10**(3), 235 (2021)
5. Chakraborty, J., Sinha, R., Sarmah, P.: Influence of accented speech in automatic speech recognition: a case study on Assamese L1 speakers speaking code switched Hindi-English. In: Prasanna, S.R.M., Karpov, A., Samudravijaya, K., Agrawal, S.S. (eds.) Speech and Computer: 24th International Conference, SPECOM 2022, Gurugram, India, 14–16 November 2022, Proceedings, pp. 87–98. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-20980-2_9
6. Felps, D., Bortfeld, H., Gutierrez-Osuna, R.: Foreign accent conversion in computer assisted pronunciation training. Speech Commun. **51**(10), 920–932 (2009)
7. George Mason University. Speech accent archive (2021). https://accent.gmu.edu/
8. Gilbert, J.B.: Teaching Pronunciation: Using the Prosody Pyramid. Cambridge University Press (2008)
9. Gondi, S.: Wav2vec2. 0 on the edge: Performance evaluation. arXiv preprint arXiv:2202.05993 (2022)
10. Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: Proceedings of the 23rd International Conference on Machine Learning, pp. 369–376 (2006)
11. Ishikawa, S.: The ICNALE Guide: An Introduction to a Learner Corpus Study on Asian Learners' L2 English. Taylor & Francis (2023)
12. Karpagavalli, S., Chandra, E.: A review on automatic speech recognition architecture and approaches. Int. J. Signal Process. Image Process. Pattern Recogn. **9**(4), 393–404 (2016)

13. Liu, D., Reed, M.: Exploring the complexity of the l2 intonation system: an acoustic and eye-tracking study. Front. Commun. **6**, 51 (2021)
14. Mikhailava, V., Lesnichaia, M., Bogach, N., Lezhenin, I., Blake, J., Pyshkin, E.: Language accent detection with CNN using sparse data from a crowd-sourced speech archive. Mathematics **10**(16), 2913 (2022)
15. Mikhailava, V., et al.: Tailoring computer-assisted pronunciation teaching: mixing and matching the mode and manner of feedback to learners. In: INTED2022 Proceedings, pp. 767–773. IATED (2022)
16. Mohammadi, S.H., Kain, A.: An overview of voice conversion systems. Speech Commun. **88**, 65–82 (2017)
17. Munro, M.J., Derwing, T.M.: Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. Lang. Learn. **45**(1), 73–97 (1995)
18. Murphy, V.A.: Second language learning in the early school years: trends and contexts. Oxford University Press (2014)
19. Paszke, A.E.A.: Pytorch: an imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems, vol. 32, pp. 8024–8035. Curran Associates, Inc. (2019)
20. Pennington, M.C., Rogerson-Revell, P.: English Pronunciation Teaching and Research, vol. 10, pp. 978–988. Palgrave Macmillan, Londres (2019)
21. Permanasari, Y., Harahap, E.H., Ali, E.P.: Speech recognition using dynamic time warping (DTW). J. Phys. Conf. Ser. **1366**, 012091 (2019). https://doi.org/10.1088/1742-6596/1366/1/012091. IOP Publishing
22. Radzikowski, K., Wang, L., Yoshie, O., Nowak, R.: Accent modification for speech recognition of non-native speakers using neural style transfer. EURASIP J. Audio Speech Music Process. **2021**(1), 1–10 (2021)
23. Rilliard, A., Allauzen, A., Boula_de_Mareüil, P.: Using dynamic time warping to compute prosodic similarity measures. In: Twelfth Annual Conference of the International Speech Communication Association (2011)
24. Rogerson-Revell, P.M.: Computer-assisted pronunciation training (CAPT): current issues and future directions. RELC J. **52**(1), 189–205 (2021)
25. Sullivan, P., Shibano, T., Abdul-Mageed, M.: Improving automatic speech recognition for non-native English with transfer learning and language model decoding. In: Analysis and Application of Natural Language and Speech Processing, pp. 21–44. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-11035-1_2
26. Thandil, R.K., Basheer, K.M.: Accent based speech recognition: a critical overview. Malaya J. Matemat. **8**(4), 1743–1750 (2020)
27. Viglino, T., Motlicek, P., Cernak, M.: End-to-end accented speech recognition. In: Interspeech, pp. 2140–2144 (2019)
28. Wolf, T., et al.: Huggingface's transformers: state-of-the-art natural language processing. arXiv preprint arXiv:1910.03771 (2019)
29. Zhao, G., et al.: L2-arctic: a non-native English speech corpus. In: Interspeech, pp. 2783–2787 (2018)