5-2024

Exploring Asynchronous Pronunciation Training Through Context-Aware Pronunciation Applications

Claire L. Schweikert
University of Nebraska at Omaha, cschweikert@unomaha.edu

University of Nebraska at Omaha
College of Information Science & Technology
Department of Computer Science
Supervisor: Dr. Harvey Siy

Honors Capstone Report

in partial fulfillment for the degree
Bachelor of Science in Computer Science (Honors Distinction)
in Spring 2024

Exploring Asynchronous Pronunciation Training Through Context-Aware Pronunciation Applications

—

Alongside development of CAPT application "Pronunciation Pal" for CSCI capstone

Submitted by:

Submission date: May 2024

Claire Schweikert
Email: cschweikert@unomaha.edu
B.S. Computer Science

Abstract

This paper provides a survey of various research articles on context-aware asynchronous pronunciation training applications. First, a set of seven articles is reviewed and summarized. Next, they are synthesized over the three main topics of 1) automated speech recognition, 2) non-native speaker considerations in language learning, and 3) future directions for research and development within computer-assisted pronunciation training (CAPT). Research in the areas of acoustic and pronunciation modeling (both implicit and explicit), pedagogical considerations for CAPT application design, Goodness of Pronunciation algorithm scoring, accent recognition and neutralization, and more are discussed.

Contents

List of Figures

List of Tables

1 Introduction

Language learning is among the most complex cognitive tasks we face as humans. With technology being found in nearly every aspect of life today, we have also begun to see how instrumental of a role tech could play in language learning, especially in the particularly difficult area of pronunciation training. This area has developed significantly in recent years, giving rise to an entire field of research and application development: computer-assisted pronunciation training (CAPT). The motivation for this paper was to compare, contrast, and synthesize various pieces of literature relating to the topic of CAPT in order to inform development of a pronunciation aid application. Various end users for a pronunciation aid application were considered, including: 1) people who are deaf/hard of hearing, 2) people learning English as a second language, 3) children not meeting developmental communication milestones, and 4) anyone else requiring additional, targeted instruction for pronunciation of specific phonemes or words. In implementation of an application to serve any member of these various user groups, it is important to keep in mind a set of numerous and widespread factors on the user's language-learning ability. As seen in the various articles explored in this paper, there are various internal and external effects on the way people learn to pronounce sounds in a new language. In this research, I focus mainly on the following two research questions: 1) What tools for CAPT are needed to support coarticulation-related issues, and how should they take these issues into account? and 2) What research has been done about the effects of first-language accents on acquistion and accurate pronunciation of a new language, and how is accurate pronunciation defined?

2 Summaries

2.1 Acoustic and Pronunciation Model Adaptation for Context-Independent and Context-Dependent Pronunciation Variability of Non-Native Speech

The paper by Oh, Kim, and Kim (Oh et al., 2008) proposes a hybrid model adaptation method for context-dependent (CD) and context-independent (CI) pronunciation variability. This model is intended to improve performance of an automatic speech recognition (ASR) system, specifically one for non-native speech by using both acoustic model adaptation and pronunciation model adaptation, and is evidenced in experiments on Korean and English.

The proposed method consists of a three-step process. First, analysis of non-native speech is conducted, forming an n-best list of phoneme sequences and resulting in a set of pronunciation variant rules from the identified sequences with the help of a decision tree. Second, these rules are decomposed into CI and CD pronunciation variation (PV) with the help of context dependency. Finally, the two types of adaptation are completed: acoustic model adaptation via a state-tying step through an indirect, data-driven method for CI PV, and the pronunciation model adaptation via construction of a multiple pronunciation dictionary using CD PV.

The paper details how the PV for non-native speech is decomposed as well as how the two distinct model adaptations (acoustic and pronunciation) are combined to form the desired hybrid method. The approach uses various assistive tools and softwares, including C4.5 for decision tree creation and the Carnegie Mellon University (CMU) Pronouncing Dictionary for the pronunciation model adaptation. As for results, this approach using both model adaptations shows word error rate (WER) reductions by over 16% when compared with the baseline ASR system trained on native speech. It achieves WER reductions by 8.95% and 3.67% in comparison to model adaptations for acoustic and pronunciation, respectively, alone.

2.2 Computer Assisted Pronunciation Training: Targeting Second Language Vowel Perception Improves Pronunciation

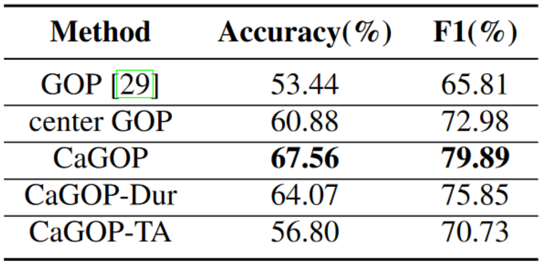
The paper by Thomson (Thomson, 2011) explores application of high-variability pronunciation training (HVPT) principles, particularly through development of a CAPT application with training features reflective of a research-based understanding of second language (L2) development. The article covers constraints on attainment of L2 pronunciation, current approaches to CAPT, and a study where L2 English speakers discriminated Canadian English vowels. Research questions include: 1) "Can computer-mediated training in the perception of L2 English vowels improve speech intelligibility without explicit pronunciation practice?", 2) "Does perceptual training generalize to L2 productions elicited by an unfamiliar voice?", 3) "Can perceptual training in one phonetic environment improve speech intelligibility in new phonetic environments?", and 4) "Do differences in the quality of training stimuli lead to differences in learner outcomes?" (Thomson, 2011).

For adults learning an L2, a set of challenges unique from those faced while learning their first language (L1) often arise. For example, learning may be difficult in the case of L2 sounds that are similar but not identical to L1 sounds, due to wrong associations of sounds in the L2 with a similar L1 category. With the help of tools like HVPT as discussed in this paper, learners can combat these challenges and learn correct pronunciation.

Within development of L2 pronunciation, three factors are identified as commonly constraining L2 pronunciation accuracy: 1) interactions between L1 and L2 phonological systems, 2) the age the learner acquired the L2, and 3) the learner's level of experience with the L2. As interactions with the L2 tend to be lacking in classroom-based language learning, the article also aims to increase the quantity and quality of phonetic experience beyond what is typically available for adult language learners. In addition, the authors hope to provide language instructors with tools to critically evaluate CAPT applications they may use in their instruction.

2.3 Context-aware Goodness of Pronunciation for Computer-Assisted Pronunciation Training

In Shi, Huo, and Jin's paper (Shi et al., 2020), they devise a Goodness of Pronunciation (GOP)-based scoring approach for CAPT mispronunciation detection. They propose a context-aware GOP (CaGOP) scoring model that involves both transition and duration factors. Of the acoustic, decoding, and scoring modules, this paper focuses on scoring (based on GOP) and its evaluation. This module converts each phonetic segment into a score based on the reference phoneme and computes the transition and duration factors into this score.



Method	Accuracy(%)	F1(%)
GOP [29]	53.44	65.81
center GOP	60.88	72.98
CaGOP	**67.56**	**79.89**
CaGOP-Dur	64.07	75.85
CaGOP-TA	56.80	70.73

Figure 1: Model performance on mispronunciation detection – reproduced from Table 2 of (Shi et al., 2020).

Unlike similar methods, the CaGOP method outperforms comparable methods in both F1 and mispronunciation detection accuracy by considering context information among phonetic segments, as shown in Figure 1. For example, the transition between phonemes with regards to the time domain, which can be tracked via observing entropy, or disorder, is involved by paying attention to frames and posterior probability. This variable reaches its height at the centermost point of a transition between phonemes and settles to zero exactly when there is a definite event (in this case, a specific phoneme being produced without effects from surrounding phonemes). Duration, on the other hand, relates to how long a phoneme sound is sustained; it is computed in a two-step process. First, a context-dependent model is applied to compute the duration for the given sequences. Then, this duration is compared to the duration of reference utterances to find the final duration factor of the test utterance. Duration of phonemes is strongly correlated with the general speed of speech.

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

The paper by Khaustova (Khaustova et al., 2023) discusses an approach to personalized CAPT which is cognizant of first-language accents and their effects on pronunciation training. The paper also reviews a tool called StudyIntonation, a multimodal and multilingual CAPT environment that focuses on improving prosodic elements of pronunciation, such as intonation, stress, and rhythm. The application uses visual elements, including charts and videos, to make this pronunciation improvement maximally accessible for a variety of language learners.

The researchers aimed to improve various aspects of common CAPT applications through their new approach. One particularly notable element of the research referenced was the intentionality the researchers brought to respecting L1 accents. This was done in a variety of ways, including keeping any accent recognition done within the app internal and not public-facing. Also, instead of targeting mistakes that users make, the new approach has a central focus on replicating the modeled correct pronunciation. An example of this model pronunciation is provided via pitch visualizations comparing the user's pronunciation pitch graph with a reference model pronunciation by a native speaker. Not only does the application draw users' attention to their pitch, but it also keeps in mind other factors on pronunciation such as environmental factors, friends and colleagues, country of living, previously learned languages, and more. The approach also targets accent detection and recognition with the help of ASR, which is trained on recordings of many non-native speakers in this case, in order to more effectively personalize training.

The paper has an additional focus on providing eduational content creators with tools and information necessary to ensure classroom instruction is most beneficial to students' pronunciation improvement through a Course Editor Module. Instructors can custom-design pronunciation courses made up of lessons and tasks, each personalized for students.

2.5 Audiovisual Tools for Phonetic and Articulatory Visualization in Computer-Aided Pronunciation Training

The paper by Kröger (Kröger et al., 2010) discusses the use of audiovisual tools as a way to improve pronunciation training and make it accessible to a wider range of learners. Three of the important research issues addressed in this paper include 1) “Can phonetic errors be detected by machine, for example by using speech recognition algorithms?”, 2) “Can learners become aware of their phonetic errors by using a human-machine interface exclusively?” and 3) “Is it possible to develop computer-aided self-learning environments for advising the learner an efficient way to overcome phonetic problems with respect to the target language?” (Kröger et al., 2010). In order to explore these areas, the paper reviews a set of interactive methods for improving pronunciation both in cases of L2 learners and speech therapy clients suffering from hearing and articulation disorders.

Auditory feedback is an important aspect of language learning and pronunciation training in specific. As we acquire language skills while growing up, most people receive regular auditory input so that we can compare our own perceived pronunciation against that of others who we hear talking. However, those with hearing impairments can be signficantly delayed in this early speech development due to the lack of auditory feedback; the same can happen with blind students, due to the lack of visual feedback when learning pronunciation. This paper introduces a few different approaches to bridge these gaps through intentional and personalized pronunciation instruction via CAPT applications that involve audiovisual tools.

One issue that must be considered is that not all users will be technically savvy, meaning any application used to teach skills in these areas should be intuitive and engaging to use. Even the display of visual comparisons such as acoustics-related parameters like spectral energy distribution and formant trajectories can be stylized in a particular way in order for users to easily understand the data being compared and how they can put the lessons learned by this comparison into practice.

2.6 A Study of Implicit and Explicit Modeling of Coarticulation and Pronunciation Variation

The paper by Dupont (Dupont et al., 2005) discusses various approaches to modeling acoustic variation among coarticulation and pronunciation. It focuses on ASR, which often utilizes context-dependent phoneme models and multiple pronunciation lexicons. This article explores acoustic models' ability to handle coarticulation and pronunciation variation among different English words. It analyzes the phonetic-level performance of both context-dependent and context-independent acoustic models while also tracking the impact of different time contexts within this modeling.

In the article, the term "coarticulation" is defined as the strong influence of phonetic context on the acoustic realization of phonemes in fluent speech. Rephrased, coarticulation refers to the process through which neighboring phonemes join together to affect each others' pronunciation. For example, the *n* sound is produced in a different mouth location in the word *tenth* than in *never*; since the *n* in tenth is followed by a *th* sound, its pronunciation takes place closer to the teeth than the *n* in *never*, where the sound is pronounced a bit farther back in the mouth since it is only followed by an open *e* sound. This coarticulation can be attributed to the "instrinsic inertia of the human speech articulatory system" (Dupont et al., 2005), and it has a notable effect on pronunciations across the English language.

In explicit modeling of pronunciation variants, phonemes are visualized by a set of articulatory features. These features may include their degree of aperture, whether they're voiced, location of articulation, etc. A set of coarticulation rules can be applied to sequences of these phonemes, resulting in a prediction of what pronunciation variants may be present from them. With the help of standard techniques such as triphone modeling and multiple pronunciation dictionaries, a complete approach to explicit modeling can be achieved.

One notable result of this study was that it demonstrated generic acoustic models as being capable of implicitly handling pronunciation variation. With this in mind, it is more important for research in this area to focus on handling varation in coarticulation effects.

2.7 Computer-Assisted Pronunciation Training (CAPT): Current Issues and Future Directions

The paper by Revell-Rogerson (Rogerson-Revell, 2021) explores the tension between technology and pedagogy when it comes to design of CAPT resources. It reports that many of these resources are less pedagogically innovative than could be expected, yet there is great potential for expansion in this area. The authors call for more intentional combination of the technological design of these applications with the pedagogical purpose behind them.

There is a great nuance to the pronunciation of words in the English language — often, the specific pronunciation of a given word by a speaker can give the listener information about the speaker's ethnic background, exposure to the English language, and their general region in the United States. With this level of complexity within each pronunciation variation for a single word, it is obvious that designers of tools for pronunciation training must treat this task with a similar level of complexity. Today, many CAPT applications are still designed with a main goal of conforming the users' pronunciation as closely to that of native speakers as possible. However, this paper proposes that it is essential that application design takes into consideration the users' personal language goals and first language background, catering their instruction to specific ways for users to use their prior language knowledge, especially similar phonetic patterns among both languages, to apply it to the new language.

The paper also investigates future directions for research and development in this area. Rather than focusing on upcoming technology developments, the authors orient the discussion toward future directions within pedagogy. They discuss areas such as ubiquitous learning; intelligent tutoring and authentic interaction; and goal-oriented, meaningful, task-based learning. One takeaway is that approaches combining intentional pedagogy with current technology may end up overlapping with virtual reality (VR). This combination could be very beneficial for CAPT users with anxiety if they're allowed to assume the form of an avatar or character, as the anonymity provided by an alternate personality decreases the pressure on that person to perform well in their pronunciation training.

3 Synthesis and Discussion

3.1 Synthesis Matrix

The synthesis matrix is shown in two parts; see Table 1 and Table 2.

Table 1: Synthesis Matrix (Part 1)

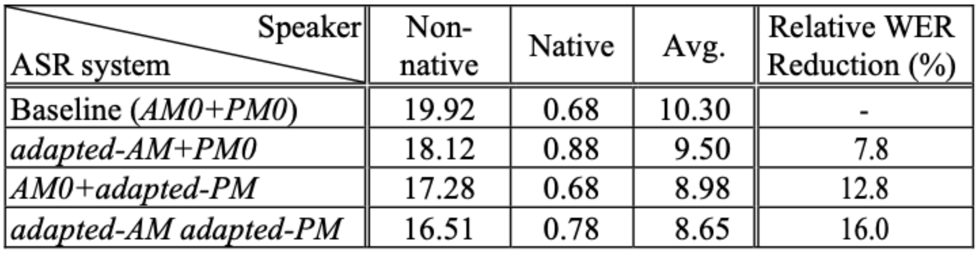
Sources	**ASR in CAPT**	**Consideration of Non-Native Speaker (L1) Backgrounds**	**Future Directions for CAPT R&D**
(Oh et al., 2008)	· An ASR system for non-native speech underperforms a comparable one focusing on native speech in recognition tasks (pg. 1)	· CI pronunciation variabilities come from a different pronunciation space than the speaker's L1 (pg. 1) · For non-native speakers, models with adaptations on either acoustic modeling, pronunciation modeling, or both, outperform reference models (pg. 3)	· A combination of adaptations of acoustic and pronunciation models can improve performance of a non-native ASR system (pg. 4)
(Thomson, 2011)	· ASR should be used in place of spectrograms to provide segmental feedback for language learners looking to improve pronunciation (pg. 5)	· Many CAPT approaches are not based in a current understanding of L2 accents (pg. 2) · Much variation in the degree of L1 accents comes from L1 influence and quantity/quality of L2 phonetic input (pg. 2)	· Current approaches to CAPT need to be better grounded in L2 accents (pg. 2) · CAPT can and should offer specific language feature instruction in ways that traditional classrooms can't (pg. 17) · Expansion of learning outside of controlled, research-based environments, especially to full words rather than just specific phones (pg. 17)
(Shi et al., 2020)	· Modern CAPT systems are based on an ASR-like architecture whose scoring module's Goodness of Pronunciation algorithm could be improved with increased context awareness (pg. 2)		· Scoring strategies to evaluate speech outside the context of forced alignments (pg. 2)

Table 2: Synthesis Matrix (Part 2)

Sources	**ASR in CAPT**	**Consideration of Non-Native Speaker (L1) Backgrounds**	**Future Directions for CAPT R&D**
(Khaustova et al., 2023)	· ASR is often used to improve both accuracy of speech detection and speed of transcription generation in content creation (pg. 2) · ASR models trained on native speech underperform when applied to non-native speech (pg. 3) · ASR could be integrated into applications with accent recognition models to improve accuracy (pg. 9)	· CAPT effectiveness can be improved through integration of ASR-based solutions that target accents (pg. 3) · Most data to train ASR models comes from native speakers, decreasing accuracy when applied to L2 speakers (pg. 3) · The CAPT experience is more personalized to the user when combining ASR, accent recognition, and accent neutralization (pg. 4) · Modeling adequate pronunciation rather than focusing on users' mistakes (pg. 12)	· ASR models to teach both segmental and suprasegmental aspects of pronunciation, trained on not just native, but also non-native speech (pg. 2) · Combining ASR, accent recognition, and accent neutralization to form a comprehensive CAPT tool (pg. 4) · Involving pronunciation training models in applications with pedagogical focuses (pg. 12)
(Kröger et al., 2010)		· Linguistic-level problems are more easily identifiable by language learners than phonetic-level problems (pg. 1) · L2 foreign accents affecting pronunciation are often more noticeable to L1 speakers than L2 speakers (pg. 1)	
(Dupont et al., 2005)	· Many ASR systems explicitly model coarticulation and pronunciation variation through CD phoneme models and multiple pronunciation lexicons, but this paper explores the benefit of studying longer time segments (pg. 2)		· Expansion of research past triphone as context-dependent phonetic unit, instead to multiple phonemes on each side of the target phoneme (pg. 2) · Building CD models and multiple pronunciation dictionaries for more spontaneous speech (pg. 5) · Effects of task and language perplexity on word-level performance (pg. 5)
(Rogerson-Revell, 2021)	· ASR is great for immediate, personalized feedback (pg. 5) · ASR models do not naturally handle accented or L2 speech very well (pg. 5)	· Many CAPT resources evaluate users' speech accuracy by comparing to native speech rather than focusing on individual users' mistakes and goals (pg. 4) · Teachers and researchers debate the need for acquiring all the phonological features of a target language (pg. 4) · Learners often don't realize their L1 interference in L2 pronunciation targets (pg. 4)	· A combination of speech synthesis, speech recognition, and AI to remove the need for language learning (pg. 12) · Improvements in real-time, robust, easily interpretable, and automated feedback (pg. 13) · Optimization of pedagogical effectiveness within CAPT applications (pg. 13)

3.2 Automated Speech Recognition (ASR)

ASR is a tool integrated into various CAPT applications. As shown in Tables 1 and 2, many papers claim that it is a helpful tool but that it does have a few areas demanding improvement. In the context of CAPT, ASR is especially beneficial for personalized and efficient feedback, as it can quickly and easily process recorded speech, but it is only as good as the data it was trained on. As discussed by Oh (Oh et al., 2008), Khaustova (Khaustova et al., 2023), and Revell-Rogerson (Rogerson-Revell, 2021), ASR models trained on native speech have been shown to have low accuracy rates when applied to non-native speech pronunciation evaluation. Figure 2 specifies word error rates (WERs) for the project's evaluation set, showing that the systems based on adapted acoustic and pronunciation models reduced error rates as compared to the baseline system (Oh et al., 2008).



Speaker ASR system	Non-native	Native	Avg.	Relative WER Reduction (%)
Baseline (*AM0+PM0*)	19.92	0.68	10.30	-
adapted-AM+PM0	18.12	0.88	9.50	7.8
AM0+adapted-PM	17.28	0.68	8.98	12.8
adapted-AM adapted-PM	16.51	0.78	8.65	16.0

Figure 2: Comparison of the average WERs (%) of the baseline ASR system and ASR systems with a different combination of adapted models for the evaluation set – reproduced from Table 2 of (Oh et al., 2008).

Weighed against comparable approaches, ASR provides a great baseline. It can be paired with accent recognition and/or accent detection algorithms, as proposed by Khaustova (Khaustova et al., 2023), to improve accuracy. In addition, its output is more useful than spectrograms in providing users specific feedback on how to improve their pronunciation (Thomson, 2011).

3.3 Consideration of Non-Native Speaker (L1) Backgrounds

There are many factors that affect a person's experience with and ability to learn a new language, especially with regards to pronunciation. One of these factors is the learner's L1 experience. Unlike problems while learning vocabulary or grammar, which tend to require more cognitive processes, problems that arise while learning the pronunciation of a new language at the phonetic level aren't always readily evident to L2 learners. Often, learners will automatically apply patterns and rules present in their L1 when attempting pronunciations in the L2, and they may not even realize these phonetic-level faults (Kröger et al., 2010).

Many modern CAPT models don't consider L1 assumptions and accents in pronunciation of the L2 because much of the data the ASR models were trained on was extracted from native speech, as mentioned by Khaustova (Khaustova et al., 2023) and Rogerson-Revell (Rogerson-Revell, 2021). With an upgrade to models trained on both native and non-native speech, effectiveness of ASR models in CAPT tools for non-native speakers could be greatly improved.

3.4 Future Directions for Computer-Assisted Pronunciation Training (CAPT) Research and Development

As the previously mentioned articles have demonstrated, CAPT is a widespread and complex area of research. With the landscape of pre-existing research in mind, there are many areas of potential into which to expand in order to improve the impact of CAPT to an even greater degree. Ideas to improve effectiveness of CAPT include: combining adaptations of acoustic and pronunciation models into one tool; grounding current approaches to CAPT for L2s in L1 accents; focusing on the services CAPT provides that traditional classroom instruction cannot; exploring deeper into pronunciation training for full words rather than just phonemes; expanding scoring strategies to account for coarticulation rather than just forced alignments among phonemes; ASR teaching both segmental and suprasegmental pronunciation lessons; combining ASR, accent recognition, and accent neutralization into one tool; expanding the context from just immediate right and left phonemes; and keeping pedagogical integration in mind alongside technological functionality. Within these areas and through the combinations therein, there is potential for much more development and sharpening of tools to help language learners improve their pronunciation.

4 Conclusion

One of the most difficult aspects of language acquisition can often be learning how to pronounce words correctly in the target language. With the help of computer-assisted pronunciation training (CAPT) applications, users would ideally have a tool to help them improve targeted sounds and pronunciations in their sedcond language. However, the effectiveness of these applications varies widely, especially how much the designers accounted for first-language effects on the acquisition of the second language. In this paper, I have explored the following two research questions: 1) What tools for Computer-Assisted Pronunciation Training are needed to support coarticulation-related issues, and how should they take into account these issues? and 2) What research has been done on the effects of first-language accents on acquistion and accurate pronunciation of a new language, and how is accurate pronunciation defined?

Both these questions relate to the wider context in which the learner is working on their pronunciation, whether accounting for the words/phonemes surrounding the word/phoneme in question or the background knowledge that learner themself is bringing to the language acquisition process. As for the first research question, I have found that various automated speech recognition (ASR) approaches can help to support coarticulation-related issues, especially those trained on triphones or a scope even wider than mere triphones. By considering coarticulation effects resulting from nearby sounds on either side of the word/phoneme in question, the model has a greater chance at understanding the word/phoneme in context and recognizing it as pronounced correctly even if it deviates from the precise pronunciation that the model expects. The second question considers learners' prior language experiences, and I have found that there is still much research to be done in this area. Research has proven that it is fairly cognitively straightforward for users to learn vocabulary and grammar in a new language, but pronunciation tends to be the area lagging behind. This "accent" coming from the learner's first language appears because the learner has established patterns and pronunciation rules in their mind, and these are hard to break in order to mold to a new language (especially if they overlap heavily with the new language). Currently, “accurate” pronunciation is often defined as being substantially similar to reference speech by a native speaker, but this isn’t always the learner’s personal goal in language acquisition. Research debates the importance of uniformity with native speech and, instead, questions whether the focus should be more on fixing specific issues the user encounters with regards to pronunciation.

Like any area of research, there still exist unanswered questions in the field of asynchronous pronunciation training. However, it has been demonstrated that there has been substantial exploration into some of the sub-areas within this field, especially those related to personalizing the experience for each language learner and considering their personal language history. This topic will continue to hold high importance in effecting positive change on pronunciation skills through CAPT.

References

Dupont, S., Ris, C., Couvreur, L., and Boite, J.-M. (2005). A study of implicit and explicit modeling of coarticulation and pronunciation variation. In *INTERSPEECH*, pages 1353–1356.

Khaustova, V., Pyshkin, E., Khaustov, V., Blake, J., and Bogach, N. (2023). CAPTuring accents: An approach to personalize pronunciation training for learners with different L1 backgrounds. In *International Conference on Speech and Computer*, pages 59–70. Springer.

Kröger, B. J., Birkholz, P., Hoffmann, R., and Meng, H. (2010). Audiovisual tools for phonetic and articulatory visualization in computer-aided pronunciation training. *Development of Multimodal Interfaces: Active Listening and Synchrony: Second COST 2102 International Training School, Dublin, Ireland, March 23-27, 2009, Revised Selected Papers*, pages 337–345.

Oh, Y. R., Kim, M., and Kim, H. K. (2008). Acoustic and pronunciation model adaptation for context-independent and context-dependent pronunciation variability of non-native speech. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4281–4284. IEEE.

Rogerson-Revell, P. M. (2021). Computer-assisted pronunciation training (capt): Current issues and future directions. *RELC Journal*, 52(1):189–205.

Shi, J., Huo, N., and Jin, Q. (2020). Context-aware goodness of pronunciation for computer-assisted pronunciation training. In *INTERSPEECH*.

Thomson, R. I. (2011). Computer assisted pronunciation training: Targeting second language vowel perception improves pronunciation. *Calico Journal*, 28(3):744–765.