A DISSERTATION
SUBMITTED IN FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
IN COMPUTER SCIENCE AND ENGINEERING

Approaches to Speech Prosody Visualization and Evaluation for Improving Tailored Feedback in Computer-Assisted Pronunciation Training Systems

by

Veronica Khaustova

March 2024

The dissertation titled

Approaches to Speech Prosody Visualization and Evaluation for Improving Tailored Feedback in Computer-Assisted Pronunciation Training Systems

by

Veronica Khaustova

is reviewed and approved by:

Chief Referee

Senior Associate Professor

Evgeny Pyshkin Pyshkin Evgeny ピシキン

Senior Associate Professor

John Blake John Blake 武礼克

Senior Associate Professor

Jeremy Perkins Perkins

Senior Associate Professor

Maxim Mozgovoy M. Mozgovoy モズゴヴォイ

THE UNIVERSITY OF AIZU

March 2024

Contents

List of Figures

List of Tables

List of Abbreviations

ADZ	Actual Development Zone
ASR	Automatic Speech Recognition
CALL	Computer-Assisted Language Learning
CAPT	Computer-Assisted Pronunciation Training
CNN	Convolutional Neural Network
CRQA	Cross-recurrence quantification analysis
DA	Dynamic Assessment
DL	Deep Learning
DSP	Digital Signal Processing
DST	Dynamic System Theory
DTW	Dynamic Time Warping
EMD	Earth Mover's Distance
IPA	International Phonetic Alphabet
MFCC	Mel-frequency cepstral coefficients
ML	Machine Learning
MSE	Mean Squared Error
NLP	Natural Language Processing
PCC	Pearson Correlation Coefficient
RQA	Recurrence quantification analysis
SCT	Sociocultural Theory
SLA	Second Language Acquisition
SLD	Second Language Development
VAD	Voice Activity Detection
WER	Word Error Rate
ZPD	Zone of Proximal Development

Acknowledgment

I am profoundly grateful to several individuals whose support and guidance have been invaluable during my research for this dissertation.

First, I express my gratitude to my research advisor, Professor Evgeny Pyshkin, whose expertise and insightful guidance have been instrumental in shaping my research trajectory. His willingness to participate in this work and provide continuous, invaluable advice has been an inspiration that has helped me grow both professionally and personally. Thank you for your unwavering support and mentorship.

I also thank my esteemed referees, Professor John Blake, Professor Jeremy Perkins, and Professor Maxim Mozgovoy, whose constructive feedback, evaluations, and suggestions have been greatly appreciated and have contributed significantly to this research.

Special thanks go to Natalia Bogach and her dedicated team at Peter the Great St. Petersburg Polytechnic University. Our collaboration on the *StudyIntonation* project has offered me an excellent opportunity to apply practical knowledge and skills in applying and evaluating the concepts discussed within this work.

On a more personal note, my heartfelt thanks go to my family, who have provided an unwavering support system throughout the demanding process of pursuing a Ph.D. I am particularly grateful to Victor Khaustov, whose support went beyond measure. His patience, understanding and belief in my work have been a constant source of strength and motivation.

I am also grateful to the Japanese Government for giving me the opportunity to do research in Japan through the MEXT (Ministry of Education, Culture, Sports, Science, and Technology) Monbukagakusho Scholarship.

Thank you to all who have contributed to my academic journey.

Abstract

This dissertation presents a comprehensive enhancement of the Computer-Assisted Pronunciation Training (CAPT) system, focusing on suprasegmental training, dynamic assessment, accent recognition, and personalized pronunciation training for learners with diverse first language (L1) backgrounds. Research integrates advanced digital signal processing, machine learning algorithms, and sociocultural language pedagogy theories to develop innovative methodologies in language learning. This research uses *StudyIntonation* system, a CAPT environment designed to improve prosody skills of language learners, including rhythm, stress, and intonation, as a test bed for the proposed enhancements. Initially, the system was designed for the English language, but gradually expanded to Vietnamese and Japanese languages.

This work starts by explaining the design and structure of the *StudyIntonation* system, detailing its components and underlying technology. It highlights the system's goals in addressing global challenges in prosody training and the deficiencies in existing CALL software. It provides analysis of usage of various metrics, including Dynamic Time Warping (DTW) algorithm for measuring similarity between two temporal sequences, before exploring the usage of Cross-recurrence quantification analysis (CRQA) for the dynamic assessment of L2 suprasegmental training and applying concepts from Dynamic System Theory (DST) and Sociocultural Theory (SCT) in language pedagogy. The proposed approach offers a way to find the developmental trajectories of learners, particularly in the context of their Zone of Proximal Development (ZPD), and evaluates the effectiveness of dynamic models in providing personalized feedback in *StudyIntonation* system.

Another area of proposed enhancements is to integrate advances in Automatic Speech Recognition (ASR) and accent recognition technology to create a more personalized and effective system based on the L1 background of the user. We detail how to implement accent classification using deep learning techniques and demonstrate that a Convolutional Neural Network (CNN) model, trained on a diverse set of accents, is effective in classifying a range of accents, surpassing existing models in accuracy. We propose improvements to the system's course editor module to allow educators to create customized pronunciation courses based on the identified L1 background. The other direction of using accent-related information is to prepare tailored ASR models to improve the recognition of accented speech and provide the learner with textual feedback on the pronounced phrase. To achieve efficient processing and robust understanding of diverse accents, we applied transfer learning techniques to an advanced multilingual model that can be used on mobile devices. We used a dataset of accented English speech and showed the improvements in the Word Error Rate (WER). We describe how to integrate these models into mobile application and course editor module of the *StudyIntonation* system, and suggest improvements to the interface, including pitch graph segmentation and segmented visualization, portraying rhythm, providing context-rich exercises, and improving the interface for task variations.

This dissertation contributes to the field of CAPT by presenting novel approaches to dynamic assessment and personalized pronunciation training. This integration of advanced technologies with pedagogical theories offers learners a personalized learning experience, optimizing suprasegmental pronunciation training in the language acquisition process.

Chapter 1

Introduction

With the rapid globalization of our world, effective communication is crucial, especially between different linguistic backgrounds. On the path to becoming proficient in a foreign language, there are certain milestones that must be reached to be able to communicate effectively. Efficient and meaningful communication assumes the ability to be understood, speak fluently, sound close to a native speaker, and comprehend spoken dialogue. However, many traditional language courses and tools have prioritized other language skills, such as grammar or writing, often neglecting the nuanced aspects of speech, such as pronunciation and prosody. Although sounding like a native speaker is not a vital goal to everyone, it is important to learn how to control correct pronunciation and intonation so that communication is not hindered.

Advances in speech technology have released new methodologies that substantially improve voice recognition and speech synthesis capabilities. Therefore, state-of-the-art pronunciation systems can now be seamlessly integrated into pedagogical devices dedicated to improving pronunciation. These applications are customized for smart devices to further empower learners, giving them the autonomy to engage in continuous self-directed practice.

The focus of this dissertation is on the approaches and methodologies that can be applied to enhance the feedback of the Computer-Assisted Pronunciation Training (CAPT) systems. As a test-bed for these investigations, we use *StudyIntonation* [1] CAPT environment, the innovative system created in the scope of an ongoing project to develop prosody-based phrasal intonation training tools grounded on supporting methods and algorithms of signal processing, visualization, and estimation.

In the following sections, we examine the importance and practicality of pronunciation instruction and familiarize the reader with the context and motivations, while also outlining the distinct challenges that have formed the research goals and questions that are essential to this study.

1.1 Purpose of the Study

There is growing interest and demand for Second Language Acquisition (SLA) in the world. According to the report [2], the online language learning market is projected to grow at a compound annual growth rate of 12.79% between 2023 and 2030 and more and more people around the world are learning a second language. Furthermore, communication between people from various global regions has been greatly facilitated by technological advances. However, the diversity of languages and different cultural contexts can pose challenges to effective communication.

CAPT systems bring new methods and ideas that can change the way we learn languages. These systems allow us to rethink old concepts and question long-held views. The primary motivation behind this research comes from the limitations observed within current CAPT systems to effectively address prosodic elements of speech. Although many CAPT tools can pinpoint and correct segmental pronunciation errors, such as individual phonemes, they often fall short when training learners in the mastery of suprasegmental elements. Other CAPT systems that allow practicing intonation have many limitations to provide personalized feedback to the user. This study aims to explore innovative approaches to visualize and evaluate speech prosody, with an emphasis on enhancing and tailoring the feedback that a CAPT system can provide to learners.

1.2 Significance of the Research

Providing effective personalized feedback to learners is not an easy task due to many reasons including the difference in learners' native languages (L1), the latter might require exercise personalization fitting the learner's level and language background. Research challenges that arise must be addressed from a multidisciplinary perspective: learning methodologies, design of learning tools, evaluation and visualization techniques, etc.

To provide a complete understanding of how speech prosody can be effectively visualized and evaluated, this study can inform the development of next-generation CAPT tools that offer more targeted and nuanced feedback to learners. Such advances would not only improve the learning experience but would also accelerate the proficiency of learners in achieving native-like prosody in their speech.

1.3 Research Questions and Objectives

This research seeks to improve the feedback mechanism of the existing CAPT systems by proposing and applying such changes to the *StudyIntonation* system. Given the complexity and multifaceted nature of CAPT systems and their potential impact on language learning, this dissertation seeks to answer the following research questions.

- How can modern CAPT systems, including *StudyIntonation*, better incorporate multi-modal feedback to cater to diverse learner preferences?
- How to perform dynamic assessment of learners performance based on the concepts of Dynamic System Theory and Sociocultural Theory?
- How to improve the CAPT systems to better recognize and cater to non-native speakers with different L1 backgrounds?
- How to implement accent-related components for the mobile-based CAPT system?

In line with these questions, the primary objectives of this dissertation are the following:

- Assess the efficacy of the CAPT systems, particularly the *StudyIntonation* system, in delivering tailored, multimodal feedback for pronunciation training.
- Evaluate the challenges posed by varying accents in CAPT systems, exploring the potential of modern techniques, such as accent recognition and modification, and user-targeted ASR to enhance feedback accuracy.
- Provide recommendations for the design and implementation of future CAPT systems, ensuring they are robust, effective, and tailored to individual learner preferences.

By exploring these areas, this dissertation aims to contribute to the growing knowledge surrounding CAPT systems, offering insights and recommendations that can shape the future of pronunciation instruction in an increasingly digital age.

1.4 Outline

This dissertation is composed of 7 chapters. The following chapters of this dissertation include Chapter 2, that provides a background and literature review on the history of Computer-Assisted Language Learning (CALL) and the evolution of the CAPT systems, current methods for evaluating and visualizing speech prosody, and the role of prosody in language learning. Chapter 3 describes the multimodal and multilingual CAPT environment *StudyIntonation* that is used as a testbed for the experiments and enhancements described in this dissertation. Chapter 4 focuses on performing dynamic assessment based on the concepts of Dynamic System Theory and Sociocultural Theory during the teaching of phrasal intonation with a CAPT system and determining learners' movement from one developmental level to another. Chapter 5 examines approaches to implement accent recognition that can be used for personalizing feedback in CAPT systems. Chapter 6 discusses how to tailor feedback to users of *StudyIntonation* by improving visual feedback and catering to learners with different L1 backgrounds. Chapter 7 concludes the dissertation.

Chapter 2

Background

Learning a new language is a complex task that involves many different components of speech, thought patterns, and cultural details. With the rise of new technology tools for learning, it is important to understand the depth of what has been studied before. In this section, we take a closer look at how pronunciation plays a major role in language learning, the history and importance of computer tools designed for pronunciation, and how technology is helping to pinpoint and improve accent differences. We will explore key research ideas from experts and the main findings where more study might be needed. This gives us a clear map of where language teaching is now and where it could head in the future.

2.1 Pronunciation and Its Role in Language Learning

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 negative impact on the overall effectiveness of language education. From the learner's perspective, pronunciation exercises are often considered as tedious and nonconstructive [3]; thus, contributing little to the measurable progress of the learner in language proficiency. 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 [4–6]. The conversation partner's first impression is based on pronunciation, so the social and practical benefits of mastering pronunciation must be taken into account. Fortunately, recent studies have revealed a more positive attitude among learners toward the inclusion of pronunciation in language learning [7].

Speech features associated with distinctive ways of pronunciation connected to the speaker's gender, age, family, social class, geographic location, and mother tongue are instantiated in the form of different language accents. Specifically, foreign accents can be considered as a compound effect of contact between two L1 and L2 phonological systems, where L1 is derived from the speaker's native language, while L2 refers to the second language [8]. 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.

It is known that speakers with heavy accents tend to make more errors in terms of standard L2 pronunciation. Speakers from regions with the same accent have been observed to have similar trends in mispronunciation [9, 10]. Therefore, for systems that aim to provide feedback on L2 pronunciation, it makes sense to determine the speaker's speech accent in L2.

In speech production, significant variation is observed within the same dialect or language among different speakers. Despite these variations, comprehension between speakers generally remains unaffected. However, the degree to which variations in speech are acceptable can differ for certain linguistic elements.

Take, for example, the pronunciation of the /p/ sound in English words like ”pie” and ”spy”. In ”pie,” the /p/ is aspirated, meaning it is pronounced with a burst of air. In ”spy,” the /p/ is unaspirated and sounds closer to a /b/. Yet, English speakers typically do not notice this difference, and understanding remains clear even if these sounds are interchanged.

This contrasts with languages such as Thai, where aspirated and unaspirated sounds are distinct and crucial for meaning. Mispronouncing these sounds in Thai can lead to miscommunication as they differentiate words.

This example illustrates that while some phonetic variations are overlooked in one language, they can be significant in another. These variations are not only between languages, but also within a single language, depending on factors like the position in a sentence or the surrounding sounds. Recognizing these variations is important in linguistics, as they affect both how speech is analyzed and how effectively we communicate.

2.1.1 Segmental and Suprasegmental features

Languages consist of two main phonological elements: segmental features, such as individual consonant and vowel sounds, and suprasegmental features, such as intonation, stress, rhythm, pitch, tempo, loudness, tone, and voice quality. While segmental pronunciation focuses on specific sound units like vowels, consonants, or syllables, suprasegmental pronunciation looks at broader intonation or prosody patterns that cover words or phrases. The term ”prosody” broadly covers all suprasegmental aspects of speech, including pitch, volume, duration, and the quality of voice. In instances where pitch variations can distinguish between word meanings or grammatical contexts, these variations are referred to as tones [11].

Segmental features refer to individual phonemes that serve as the building blocks of words. These features often constitute the primary focus of language instruction, as they are readily identified and isolated for practice. However, suprasegmental features operate over stretches of speech units such as syllables, words, or phrases [12]. These features are critical for conveying meaning and emotion, and their mastery can significantly influence the comprehensibility and fluency of a non-native speaker. In a multilingual context, the intricacies of segmental and suprasegmental features can vary considerably across languages, making it particularly challenging for learners to adapt to the phonetic and prosodic norms of the target language. Therefore, an understanding of both feature types is essential for non-native learners aiming to achieve a natural and intelligible pronunciation in any given language.

2.1.2 Prosody

One area that merits considerable attention in the field of CAPT is the integration of prosody, the rhythmic and melodic aspects of speech, especially within a multilingual context. Prosody encompasses an array of linguistic features such as pitch, tempo, and stress, which serve as critical constituents of meaning and communicative efficacy [13]. In a multilingual setting, prosodic patterns often vary widely across languages, thereby constituting a notable challenge for non-native speakers who strive to achieve high levels of proficiency. Understanding and mastering the prosodic elements of a language can impact the intelligibility and naturalness of spoken discourse. As such, incorporating prosody in CAPT applications substantially enriches the learning experience by offering language-specific guidance on these often overlooked but crucial elements of speech.

The significance of prosody is demonstrated in an example sentence concerning dogs. Figure 2.1 shows the speech for two examples: “No. Dogs are here.” (on the left) and “No dogs are here.” (on the right) [14]. The upper part of the two examples displays the variation in sound pressure over time; the lower part shows the pitch contour. Poor prosody could make speech unintelligible and incomprehensible, and it can also be used to disambiguate meaning.

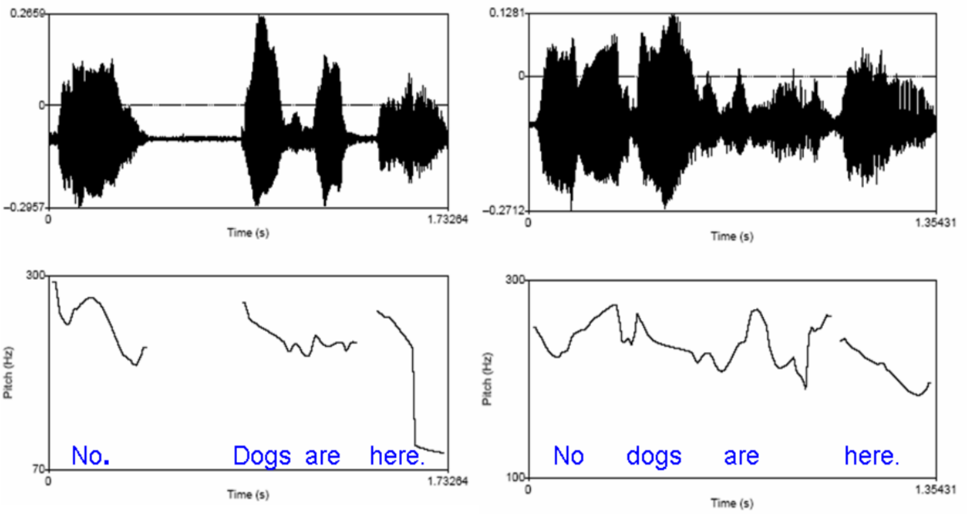


Figure 2.1: Prosody helps to identify sentence boundaries and disambiguate meanings.

Stress, Intonation, and Tone

Stress, intonation, and tone are linguistic features that operate on the suprasegmental level [15], shaping the prosodic landscape of speech. Stress refers to the emphasis placed on certain syllables or words. In English, stress can change the meaning of a word (e.g., [ˈrekərd] as a noun versus [rəˈkôrd] as a verb) or indicate the importance of a word in a sentence. Stress-timed languages such as English, German, Swedish, and Russian use the amount of time between stressed syllables, which tends to be relatively consistent, while the number of unstressed syllables between them can vary. While the concepts of stress-timing and syllable-timing offer useful frameworks for understanding the rhythmic properties of languages, they are simplifications. Many languages do not fit neatly into these categories and may exhibit mixed properties or vary depending on dialect or regional differences. Emphasis often relates to stress in that it is a way to highlight certain syllables, words, or phrases within speech. By emphasizing a particular word or syllable, a speaker can change the meaning, tone, or emotional content of a sentence. This feature isn't confined to individual sounds but spans over larger units of speech, making it suprasegmental. Emphasis can be achieved through a combination of increased loudness, changes in pitch, elongated duration, and other prosodic elements.

Intonation, as a form of prosody, exists in both tonal and pitch-accent languages. However, the function and prominence of intonation can differ between these types of languages. In tonal languages, where pitch contours at the syllabic level are crucial for distinguishing lexical or grammatical meaning, intonation must be integrated carefully so as not to conflict with these tonal distinctions. As a result, the methods for incorporating intonation in tonal languages may be more constrained, leading to different conventions for conveying prosodic meaning compared to pitch-accent languages. For example, a study by [16] found that Chinese speakers relied on native lexical tones to produce English prosody. The importance of intonation varies by language. As a rule of thumb, when learning tonal languages, in the early stages mastering the individual tones has a greater impact on intelligibility and is therefore more of a priority than working on intonation. However, intonation is much more important when learning English, since using intonation inappropriately may create a negative impression. For example, when making a polite request, such as *Could you help me?* using falling intonation rather than the conventional rising intonation will most likely cause the listener to assume that the speaker is disappointed or angry.

Tone, particularly relevant in tonal languages, involves the use of pitch variations at the syllable level to distinguish lexical or grammatical meaning. Since tone operates at the level of entire syllables or words rather than individual sounds, it is considered suprasegmental. In contrast, pitch is a perceptual parameter related to the frequency of the vocal fold vibrations, which listeners interpret as either high or low. While stress, intonation, and tone are linguistic features that utilize variations in pitch to convey meaning, pitch itself is a purely acoustic property.

Pitch

Pitch refers to the perceived frequency of the sound, basically, how high or low a note is. In speech, pitch can vary to create intonations. The pitch is used to convey different emotions, ask questions, and add melody to the speech. For example, in English, the pitch often rises at the end of a sentence when asking a question. Additionally, in tonal languages like Thai and Chinese, pitch can change the meaning of a word. Pitch can be varied throughout the sentence or just a part of it. It can add expressiveness to speech and also help to maintain the listener's interest and attention.

Pitch is often conflated with the fundamental frequency (f_0) [17], although the two are distinct [18]. The fundamental frequency reflects the rate of vocal fold vibrations, whereas pitch is the perceptual correlate of this frequency and is subjectively interpreted by listeners as being high or low. As pitch is a perceptual property, it cannot be directly extracted from speech recordings; what can be measured is the fundamental frequency or f_0 contour [17], which serves as an acoustic indicator of what listeners perceive as pitch. This distinction is crucial for non-native speakers seeking mastery in the pronunciation of a target language, as it underlines the difference between the physical and perceptual aspects of speech sounds.

Therefore, understanding the relationship between pitch and fundamental frequency can provide valuable insights for mastering linguistic features such as stress, intonation, and tone. In stress-based languages like English, for example, syllabic emphasis is often marked by a higher fundamental frequency, which listeners perceive as stress due to a higher pitch. In the domain of intonation, the modulation of fundamental frequency over a stretch of speech — often colloquially referred to as the "pitch contour" can significantly impact the meaning of an utterance. For example, the rising fundamental frequency at the end of the English question "You're coming?" signals a question intonation.

For tonal languages like Vietnamese, the manipulation of fundamental frequency at the syllabic level results in different lexical or grammatical meanings, which are then perceived as different tones or pitch contours by the listener. Meanwhile, pitch-accent languages like Japanese employ variations in fundamental frequency to distinguish words that are otherwise phonemically identical. For example, the word *hashi* can signify *bridge* or *chopsticks* depending on its pitch pattern, which in turn is dictated by its fundamental frequency.

Hence, an understanding of pitch and fundamental frequency is indispensable for non-native speakers striving to perfect their pronunciation and to interpret the spoken language more accurately. Whether the target language is stress-timed like English, pitch-accented like Japanese, or tonal like Vietnamese, mastering these acoustic and perceptual aspects of speech will significantly improve communicative efficacy.

Pitch accent languages are those that assign a tone to one syllable within a word. This syllable may be pronounced with a higher pitch to distinguish it from the other syllables. The pitch accent languages include Japanese and Swedish. English may use pitch to emphasize a syllable, but the stressed syllable may also be emphasized by saying it louder or elongating the sound. Syllables that stand out are known as stressed syllables while the other syllables are unstressed. Therefore, English does not clearly match all the criteria needed to classify it as a true pitch accent, since the method of stressing the syllable differs. In English, up to two syllables may be stressed in multi-syllable words. The main stress is called primary, while secondary stress may also occur in the initial syllable of a word. Mistakes in both the placement of accent or stress may impact intelligibility.

Voice quality

Voice quality is the character or quality of a voice that distinguishes it from others even if they say the same thing at the same pitch and loudness. It can be affected by factors such as the tension of the vocal cords and the shape of the speaker's vocal tract.

Tempo and loudness

Tempo is the speed with which someone speaks. Speaking too quickly or too slowly can affect comprehension. Loudness refers to the volume or amplitude of speech. Like intonation, changes in volume can convey different emotions or emphases.

2.1.3 Isochrony

Isochrony is a concept in the field of phonetics and linguistics that refers to the rhythmic division of time into equal intervals in spoken language. It serves as a useful paradigm for categorizing languages based on their rhythmic structures [19, 20]. By focusing on specific pronunciation features, namely: stressed syllables, all syllables, or mora; this method allows linguists to broadly classify languages into stress-timed, syllable-timed, and mora-timed categories. These categories are not without criticism, though [21, 22]. While useful for comparing languages, teaching them, and in speech technology, isochrony has its critics and limitations. The perception of speech rhythm can vary greatly depending on the listener's language background and the context in which something is said. The actual rhythm might not always match what we perceive. Despite these issues, using isochrony for classification has educational benefits. It can help language learners grasp pronunciation nuances, fluency, and comprehension, particularly when learning languages with different rhythm types. However, the isochrony concept has downsides; it can oversimplify language prosody, forcing complex rhythmic patterns into basic categories. It also often ignores how different dialects within the same language can have unique rhythms, which challenges simple classification.

Stress-based languages

In languages like English, which are stress-timed, the time between stressed syllables tends to be fairly regular. This often leads to the shortening of unstressed syllables, resulting in language patterns such as contractions and elisions. For those learning the language, understanding and pronouncing these unstressed syllables, particularly in words like articles and prepositions, can be challenging.

In our *StudyIntonation* project [3], we have focused primarily on this type of stress-timed English. However, it is important to recognize that not all varieties of English spoken around the world fit neatly into this category. For instance, English as spoken in some regions of the Caribbean, India, or Singapore might lean more towards being syllable-timed or exhibit a mix of rhythmic structures. Thus, the way we categorize spoken languages can be more nuanced and varied than it might seem at first.

Syllable-timed languages

Vietnamese is a good example of a syllable-timed language, where each syllable is given about the same duration. This equal timing often makes each syllable stand out more clearly, which can simplify pronunciation for those learning the language and result in more uniform rhythmic patterns. Languages like Vietnamese and Italian, both syllable-timed, tend to have steady rhythmic flows because each syllable takes a similar amount of time to pronounce. A key difference, however, is that Vietnamese is both syllable-timed and tonal, meaning the tone of each syllable changes its meaning, while Italian is syllable-timed but non-tonal.

Mora-timed languages

A mora is a unit in phonology that influences the weight of the syllable, and each mora takes approximately the same amount of time in speech. This equal duration of morae makes mora-timed languages unique in their rhythmic and prosodic features. In these languages, it is the count of morae, rather than syllables alone, that shapes the linguistic rhythm. This moraic system has a significant impact on various linguistic elements, including phonetics, phonology, and how speech is processed.

Japanese serves as a prime example of a mora-timed language. Take the word *Tokyo*: in Japanese, it is not just two syllables ('To-kyo'), but four morae ('To-u-kyo-o'). This emphasis on morae is integral to the rhythm and flow of the language, as evidenced by research indicating the importance of the mora in Japanese speech patterns [23, 24].

2.2 Computer-Assisted Pronunciation Training Systems

2.2.1 Brief History of CALL and Evolution of CAPT Systems

Computer-Assisted Language Learning (CALL) reveals a significant evolution from the early days of pre-recorded audio samples to the present state-of-the-art AI-driven systems. Over time, there has been an increasing focus on personalization, accuracy, and context-specific feedback. Early CALL systems (1960s-1970s) used pre-recorded audio samples to help language learners practice pronunciation [25]. The students listen to and repeat these samples, often comparing their pronunciation with the samples. This approach was limited in terms of personalization and context-specific feedback. The development of text-to-speech synthesis and speech recognition technologies in the 1980s allowed for more interactive pronunciation practice [26]. Learners could receive immediate feedback on their pronunciation, but the accuracy of the feedback was still limited due to the early stages of speech recognition technology.

The rise of the Internet and multimedia technologies in the 1990s marked a turning point for pronunciation feedback in CALL. Online resources and software offered a wider range of pronunciation exercises and allowed for more detailed feedback. Although still limited by speech recognition technology, some systems started to provide visual representations of pitch, stress, and intonation patterns to help learners improve their pronunciation [27].

The widespread adoption of mobile devices and the Internet in the 2000s further revolutionized pronunciation training in CALL. Learners can now access pronunciation resources anytime, anywhere and receive real-time feedback on their pronunciation skills. Additionally, speech recognition technology saw significant improvements, enabling more accurate analysis of learner pronunciation and better personalized feedback.

Advancements in Machine Learning (ML) and Natural Language Processing (NLP) techniques in 2010 further enhanced pronunciation feedback in CALL. These technologies enabled the development of adaptive feedback systems. Software such as Rosetta Stone [28] and Carnegie Speech NativeAccent [29] began offering pronunciation feedback, targeting specific areas of improvement for individual learners.

Today, CALL continues to evolve, with the development of more advanced technologies such as Artificial Intelligence and Machine Learning. These technologies have led to the development of more sophisticated and personalized language learning experiences. AI-powered CALL programs can analyze learner data to provide tailored feedback, and can also offer interactive language learning experiences through chatbots and other conversational interfaces.

Throughout the evolution of pronunciation feedback in CALL, a consistent focus has been on improving personalization, accuracy, and context-specific feedback. As technology continues to advance, we can expect more innovations to drive improvements in pronunciation training and feedback for language learners.

The evolution of CAPT systems owes its inception to CALL. CALL systems have witnessed significant advancements due to technological progress and pedagogical innovations [30]. Initially, these systems primarily offered simple exercises like 'listen and repeat' activities. However, these exercises provided limited interaction and feedback, resulting in unsatisfactory learning outcomes. As the focus of CALL shifted towards pronunciation, the platforms began to incorporate advanced technologies such as signal processing and ASR to better analyze and guide learners on their pronunciation.

Today, CAPT systems, a specialized branch of CALL, employ modern computational methods and artificial intelligence to offer learners a more customized and effective learning experience. One of the cornerstone technologies in this transformation has been speech recognition, which converts spoken language into text. In CALL, this allows for a precise comparison of learner pronunciation with that of native speakers or set linguistic benchmarks. Although technology has advanced significantly, recognizing diverse accents and speech patterns remains a challenge.

Speech recognition and synthesis technologies have transformed the landscape of language learning. The former, by converting spoken language into text, offers real-time feedback, allowing learners to refine their pronunciation by comparing it with native benchmarks. The latter, called text-to-speech, turns written content into audible speech, providing learners with clear models for pronunciation and intonation. The infusion of ML and NLP has further elevated these technologies. Advanced ML models, especially within Deep Learning (DL), have improved the accuracy of recognizing diverse accents and speech nuances. At the same time, ML facilitates a deeper understanding of the structure and semantics of the language, allowing more nuanced feedback and creating interactive platforms like chatbots. Together, these technologies, bolstered by ML and NLP, offer a comprehensive and personalized approach to language education.

At the same time, text-to-speech technology, known as speech synthesis, transforms written text into spoken words. This can be used in CAPT systems, enabling the generation of native-speaker-like samples for learners to emulate. As text-to-speech has evolved, it now produces more natural and precise speech samples tailored to the specific accent or dialect desired by the learner, facilitating a more individualized and context-driven approach to pronunciation practice.

2.2.2 Existing CAPT Systems

This thesis aims to enhance the personalized feedback capabilities of the *StudyIntonation* system, which is thoroughly discussed in Chapter 3. In terms of design and key features, we can cite other systems in some way similar to *StudyIntonation* or share similar ideas:

- IntonTrainer [31]. The system focuses on suprasegmental training in English and includes a rich acoustic database to cover a wide range of intonation patterns. This system engages users through interactive auditory and visual means and quantitatively assesses their intonation proficiency. Currently, the system has exercises to train the intonation of Russian, British English, and basic exercises to improve singing skills. Moreover, there are several demos for Belarusian, American English, Chinese, and German, as well as the intonation of emotional speech. Adding other languages is possible by creating appropriate acoustic databases.

- Another notable system is designed for Hungarian children with profound hearing loss, with the aim of effectively teaching speech prosody. [32]. It leverages automatic scoring and various visual feedback methods to be both user-friendly and educational. This system, while tailored to Hungarian prosody, has the flexibility to be adapted to other languages. It demonstrates its efficacy through a subjective listening test where students who used the software for three months showed improved prosody compared to those who did not use the system.

- BetterAccent Tutor [33] is a software tool designed to aid in the learning and teaching of English pronunciation and accent reduction. It utilizes advanced speech recognition and analysis technologies to provide real-time visual feedback on a user's pronunciation, intonation, rhythm, and stress patterns compared to native-speaker norms. This software typically features a variety of interactive exercises and drills that adjust to the learner's proficiency level, allowing for personalized training. It aims to improve the learner's speaking skills by offering a structured approach to accent training, helping users develop a more natural-sounding English accent and thereby enhance their communicative abilities.

- The FLUENCY system [34], is an educational technology designed to help language learners improve their spoken English proficiency. FLUENCY integrates speech recognition and natural language processing technologies to offer interactive dialogue-based exercises that adapt to the user's level of proficiency. By engaging learners in conversations, the system focuses on improving their fluency, listening comprehension, and spoken grammar. It also provides immediate feedback on pronunciation and linguistic choices, fostering an immersive learning environment that encourages continuous practice and self-improvement. The ultimate goal of the FLUENCY system is to support learners in achieving a higher degree of comfort and naturalness in their use of the English language in real-life communicative contexts.

- The EduSpeak system [35], is a comprehensive educational software suite designed to assist language learning through speech recognition technology. It integrates advanced spoken-language processing features to facilitate interactive language learning and personalized feedback. EduSpeak supports multiple languages and is adaptable to various curricula, allowing for a broad application in language education. The system focuses on enhancing pronunciation, fluency, and comprehension by providing learners with an automated and real-time assessment of their speech. It is tailored to improve language acquisition efficiency in both classroom and individual settings, capitalizing on its technology to deliver a flexible and engaging user experience.

- WinPitch [36] is a versatile software tool designed for phonetic and phonological analysis, with a strong emphasis on pitch visualization. It allows linguists, language educators, speech therapists, and researchers to record, upload, and play back speech, providing detailed visual representations of pitch contours and spectrograms for in-depth analysis. The software's capability to segment and annotate speech makes it ideal for detailed phonetic study and language instruction, offering comparative analysis features that are invaluable for accent reduction and pronunciation training. Additionally, WinPitch's user-friendly interface facilitates the creation of tailored exercises for teaching intonation, while customizable settings cater to specific research or pedagogical needs, solidifying its reputation as a comprehensive tool for both educational and research applications in spoken language analysis.

- KaSPAR [37] is an innovative project designed to help Italian speakers with dyslexia improve their English prosody and pronunciation to emulate the speech patterns of native English speakers. The core of the project is the development and implementation of a system that uses visual feedback through intuitive graphs, thereby strengthening users' expressive skills. By comparing their own speech with that of native speakers, students

> can receive valuable insights into their pronunciation. The system cleverly integrates existing tools such as Praat [38] for sound manipulation, SPPAS for text alignment, and a pronunciation evaluation mechanism based on phonetic alignment and automatic speech recognition technology. While currently focused on English, the system holds the potential for adaptation to other languages, provided they are supported by the selected alignment tool and the pronunciation system is retrained accordingly. Preliminary results have shown promise, indicating that the system performs well, despite some limitations due to data noise and training size. A comprehensive evaluation of its effectiveness, particularly for people with dyslexia, is expected to be completed by long-term testing. The project invites further research to expand its capabilities and refine its evaluative precision.

Some of the mentioned systems are more focused on practical aspects of language learning, such as improving intonation, rhythm, and pronunciation, and are typically user-friendly for learners of various ages and needs. Others, like WinPitch and KaSPAR, provide more technical analysis and are used by professionals for detailed phonetic study and research. These tools can be adapted for educational purposes, but are generally more feature rich and offer more in-depth analysis.

There are also annotation software like Praat [38] and TI_TOBI (Tones and Break Indices) [39] that serves specialized purposes in the field of speech analysis and linguistic research.

Praat is a versatile software tool widely used in the fields of phonetics and linguistics. It provides advanced capabilities to analyze, synthesize, and manipulate speech sounds. With Praat, users can record audio, visualize it in various forms such as waveforms and spectrograms, label and segment recordings, extract acoustic features, and conduct statistical analyses. Its broad functionality also extends to the study of both spoken and sung voice quality and intonation, allowing researchers to examine fine-grained phonetic details. Praat's flexibility and comprehensive feature set make it an essential resource for researchers, educators, and students engaged in the detailed study of speech and its properties. The software is open-source and maintained through an active community, ensuring its ongoing adaptation and enhancement in response to the evolving needs of the field.

TI_TOBI, on the other hand, is an annotation system used for marking intonation and prosodic features in spoken language corpora. It follows a set of conventions for transcribing the pitch and intonational features of speech, providing a framework for researchers to analyze and compare prosodic elements across different languages and dialects. The system is built upon the TOBI (Tone and Break Index) framework, which divides intonation into pitch accents and boundary tones, allowing linguists and speech technologists to systematically study and document the intonational patterns that occur in natural conversation. TI_TOBI is often used in the fields of speech synthesis, recognition, and linguistics to train models and improve the naturalness of synthesized speech.

These tools are less about interactive language learning or pronunciation practice, as in the first group of systems we categorized, and more about the detailed scientific analysis and annotation of speech. They provide foundational support for research that could inform the development of language learning tools, but they are primarily used in a research context rather than for direct language acquisition.

On the basis of our research and understanding, it appears that none of the solutions currently available in the market, including the one we have developed, perfectly embody the optimal model for delivering feedback. Ideally, feedback should be tailored according to the learner's preferences, whether they prefer it in a graphical, textual, or auditory format. Additionally, this feedback should be interwoven with clear instructional links that connect different modes of learning. Despite the advances and capabilities of many state-of-the-art CAPT tools, there remains a noticeable void when it comes to providing clear and explicit multimodal feedback, especially in the context of acquiring and assessing suprasegmental elements of foreign languages [40].

2.3 Personalized Feedback in CAPT Systems

In the world of Computer-Assisted Language Learning (CALL), the importance of personalized feedback for pronunciation cannot be overstated. Effective feedback is essential for learners to develop accurate pronunciation and improve their communication skills. CAPT systems have emerged as a promising tool for language learners who seek to improve their speaking skills. These systems offer learners the ability to receive immediate feedback on their pronunciation, allowing them to practice and improve their speaking skills in a structured and effective way. However, one of the key challenges of CAPT systems is providing tailored feedback on speech prosody, including rhythm and intonation, which are essential aspects of effective communication. Errors in prosody can lead to misunderstandings and communication breakdowns. In recent years, researchers have explored different approaches to visualization and evaluation of speech prosody in CAPT systems in order to better tailor feedback to individual learners. In this section, we will review and analyze these approaches to speech prosody visualization and evaluation, exploring their strengths and weaknesses, and how they can be used to provide tailored feedback in CAPT systems, ultimately leading to improved pronunciation skills for language learners.

2.3.1 Speech Prosody Visualization and Evaluation

In traditional classes, for pronunciation and conversation, the audio materials are often used in isolation, thus, actuating only the auditory perception channel, which is not the main perception channel for most people. Many authors have commonly agreed and reported that computer-assisted solutions can help harness higher levels of multimodality, including visual, audial, verbal, and even kinesthetic channels [41], [42], which helps support the diversity of learning styles [43], [44]. With the advancements of digital and mobile technology enabling better personalized solutions, the extensive use of visual perception linked to audial input and automated speech processing and evaluation has become easier to implement.

Displaying the fundamental frequency (the primary acoustic element associated with stress and intonation) alongside auditory feedback, as seen in the *StudyIntonation* environment, proves beneficial [1]. The value of such visual feedback is enhanced when it concurrently showcases the learner's pitch contour with a structured interpretation of the variance between the reference and the student's pitches. There are numerous conventional speech processing algorithms to determine pitch [45,46]. However, selecting the most appropriate algorithm for pitch estimation and similarity assessment based on the specific application remains a point of contention. Various studies employ metrics like Mean Squared Error (MSE), Pearson Correlation Coefficient (PCC), Earth Mover's Distance (EMD), and DTW to compare reference and student pitches. DTW, in essence, evaluates the similarity of patterns across different time sequences [47]. Research by [48] has shown that DTW is adept at recognizing congruent intonation patterns, making it more resilient than the Euclidean distance due to its tempo invariance.

However, using DTW scores has its shortcomings. The first is its static nature, offering only a snapshot of a learner's progress at a particular moment. The second drawback is its broad scope; the score gives a comprehensive assessment without granular details. This absence of specific measurements can make score interpretation challenging, potentially leaving students uncertain about the adjustments needed to enhance their pronunciation. An alternative approach involves leveraging metrics from the CRQA [49,50], enhancing the quality and clarity of CAPT feedback. In scenarios where students are required to align their prosodic features with a reference model, CRQA metrics can shed light on their learning trajectory. Ideally, CRQA metrics should correspond both to the successes and challenges faced by learners. The impact of prosodic synchronization has been explored in emotion detection applications [51] and in dissecting conversational dynamics [52].

2.3.2 Dynamic Assessment based on Dynamic System Theory

DST) has gained attention as a holistic foundation of Second Language Development (SLD) theories [53–55]. Within the DST view, all agents of language interaction are seen as dynamic systems, and all the SLD processes, which occur for all language competencies [56, 57] at various timescales, are explored with respect to their dynamics [58]. A wide range of research on language development considers language acquisition as the emergence of language abilities over time and through language use, and not just as a process of acquiring abstract rules [57–60]. Emergence is understood as the spontaneous acquisition of new features and forms as a result of self-organizing interactions of complex system components [61, 62].

The idea of learning as an emergent process was described within SCT [63–65], which considers L2 development as a process of social mediation, thus sharing an emphasis on the role of environmental contexts with DST [66, 67]. SCT-informed L2 pedagogy pays much attention to individual development trajectories and operates with constructs such as ZPD [66], scaffolding [68], mediation [69], inner speech [70] and Dynamic Assessment (DA) [71]. ZPD is the region through which learners improve from their actual level to their potential level under the guidance and through feedback [54, 66]. DA is understood as a way to infer the abilities of the learner and move beyond performance assessment towards understanding of the processes underlying the individual learning dynamics [64, 66]. ZPD and scaffolding were articulated in terms of contemporary DST theory in [53, 68, 72], making quantitative research possible with DST instruments and non-linear time series techniques [62, 67, 73].

The field of DST for L2 pronunciation teaching and learning is still under-addressed currently [74, 75], but there is evidence that DST applied to L2 pronunciation teaching and longitudinal L2 pronunciation assessment is an insightful tool to explore the individual progress and motivation of learners [76, 77]. Using DST as a theoretical framework could bridge the gap between L2 phonology and pronunciation teaching [6, 78]. Under the assumption that speech is inherently recurrent [79–81], speech prosodic phenomena allow quantitative evaluation using the dynamical modeling technique of CRQA [82, 83].

2.3.3 Accent Recognition and Classification

Accent recognition and classification is an expanding field in speech technology. The ability to identify and classify the accent of a speaker can have multiple applications, ranging from personalized language learning to sociolinguistic research [84]. 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.

Automatic recognition of foreign accents (i.e., detection of the speaker's L1 based on L2 samples) can improve the robustness of ASR-based software and CAPT systems. Accent detection can help overcome the unwanted variability of speaker-independent speech recognition models [85–87]. The knowledge gained from accent classification can improve the overall performance of an ASR system and make it more reliable; since, in the case of preliminary accent identification, the speech recognizer can be further trained for a specific accent group [85, 88, 89]. Conventional acoustic language models adapted to fit the standard language corpus cannot satisfy the recognition requirements for accented speech. Solving the problem of accented speech recognition by adding more pronunciation samples to the dataset used for training is inappropriate, since such an approach increases processing time and creates additional noise that degrades performance [88].

Research shows accent detection can contribute to significant improvements in algorithms, models, and interfaces of other human-centric systems, including but not limited to:

- Analysis and modeling of speakers' variability in the frame of speech recognition [9];
- Development of user interaction scenarios in video-games [90];

- Analysis of phonetic particularities and related personal behavior [91];
- Using accent-related information as components of biometric data [92];
- Mitigating accent influence in voice-control systems [93];
- Improving personalization of exercises and feedback in CAPT systems [94].

2.3.4 ASR addressing speakers with different L1 backgrounds

As reported in [94], the accuracy of the ASR software can be high for native speakers but drops significantly for L2 speakers with advanced level proficiency but accented speech. Accent-aware modeling has recently been reported to be an efficient approach to improve mispronunciation detection and diagnosis systems [95, 96]. However, it is generally assumed that the information about the accent of an utterance is known in both the training phase and the testing phase, although, in real-life scenarios, the accent might be *a priori* unknown.

ASR technology has evolved significantly over the years [97]. Initial attempts at ASR were based on simpler models and relied heavily on handcrafted features. The emergence of DL 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 [98]. These models, trained on carefully curated datasets, provided a marked improvement in performance for language learners with different native languages [99]. Datasets, such as L2-ARCTIC [100], 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 [84]. 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.

Accent modification [101] and voice conversion [102] 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 [103]. When such modifications are applied to CAPT learner speech, they improve comparison results with reference pronunciation [104]. This, in turn, leads to major improvements in pronunciation training, allowing learners to receive more accurate feedback and better understand their pronunciation errors.

In addition, some existing studies [105, 106] suggest that second-language learners show greater success when practicing with a voice model that closely resembles their own. Unlike previous methods [107], which relied on predetermining speaker-related cues for each second-language (L2) speaker and training a separate model for each speaker, it would be interesting to develop a speaker-independent acoustic model trained on corpora for the first language (L1). This model extracts bottleneck features that encapsulate the linguistic content of utterances. These features then be transformed into mel-spectrograms through a sequence-to-sequence model, influenced by both the L2 speaker's embeddings and the L1 accent characteristics. This approach relies on findings of [108], but Gaussian mixture attention mechanism has been shown to be more robust in generating longer utterances [109]. By using this method, it would be possible to facilitate a zero-shot learning system capable of generating accent conversion for any learner using just a few seconds of their speech, without the need for retraining or fine-tuning the model.

Our 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 [94].

Chapter 3

StudyIntonation: An Innovative CAPT System

This chapter introduces the *StudyIntonation* system, a computer-assisted teaching environment that focuses on improving the pronunciation skills of learners, specifically in the area of prosody, which includes rhythm, stress, and intonation. This system employs a range of digital signal processing methods, such as detection of speech activity, modeling of pitch visualization, and quality assessment of pronunciation and learner's progress. In addition, it is supported by an interactive interface customized for mobile devices, taking advantage of the availability and widespread use of modern technology. The development of the system is the result of a close collaboration between an international team involving researchers from the University of Aizu, Peter the Great St. Petersburg Polytechnic University, St. Petersburg Speech Technology Center, and Hachinoche College.

3.1 StudyIntonation: Design Goals and System Structure

StudyIntonation is a learning environment that provides feedback on pronunciation exercises to learners based on signal processing algorithms. These are used to construct interpolated pitch graphs displayed on a mobile screen, with the support of an audio-visual content repository and the extensible course developer toolkit incorporating a number of model courses for different L2 languages into the prosody training workflow. Figure 3.1 shows the main elements of the architecture of the *StudyIntonation* system and the major steps of pitch processing, visualization, and estimation.

The primary goal of *StudyIntonation* is to address and resolve four major global challenges confronting prosody teaching and learning:

- Addressing the Neglect of Prosody Training: Despite the proven importance of prosody in effective communication, it is often overlooked in language teaching curricula. Hence, *StudyIntonation* seeks to redress this imbalance by providing a robust and accessible platform explicitly dedicated to prosody training.
- Encouraging Learner Engagement: Learners often express restraint when asked to practice speaking a new language. Therefore, *StudyIntonation* aims to counter this inhibition by offering an engaging digital learning environment that encourages frequent and active participation.
- Empowering Teachers: Despite the critical role of prosody, teachers often lack the training to provide efficient instruction in suprasegmental aspects of language. To address this challenge, *StudyIntonation* provides an extensive courseware development kit, enabling teachers to design and implement task series and courses suited to diverse learner requirements.

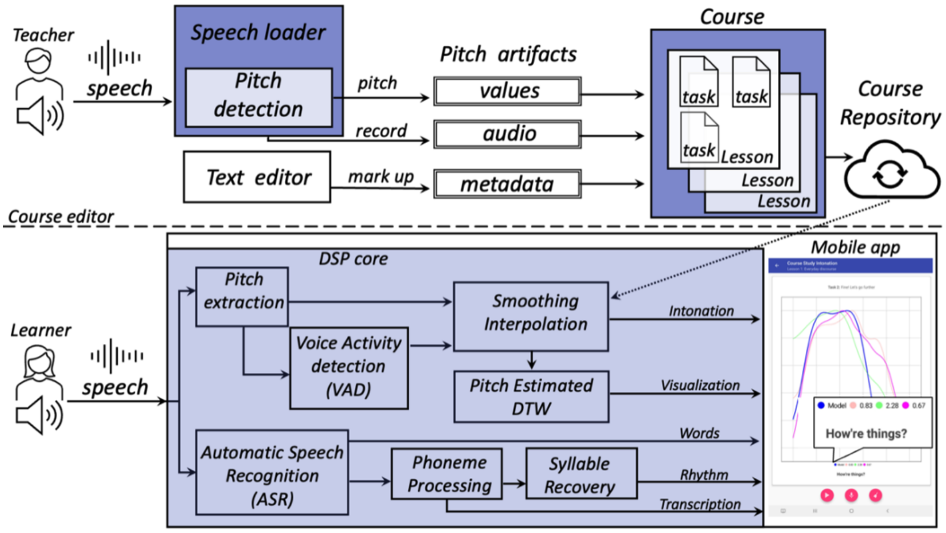


Figure 3.1: *StudyIntonation* workflow.

- Addressing CALL Software Deficiencies: Many existing CALL platforms lack well-designed linguistic and pedagogical content, a gap that *StudyIntonation* aims to fill by providing a research-driven, linguistically informed, and pedagogically sound platform.

StudyIntonation aims to address the following language learning aspects:

- The pedagogical soundness of its learning content.
- Influence of learning style and influence of prevalent learner modality (e.g. visual learners, auditory learners, etc.).
- The adequacy of the learners' attempts displayed in the mobile application.
- The consistency between visual feedback and the numeric evaluation metrics used.
- The developmental dynamics of learners.

3.2 System's Components and Underlying Technology

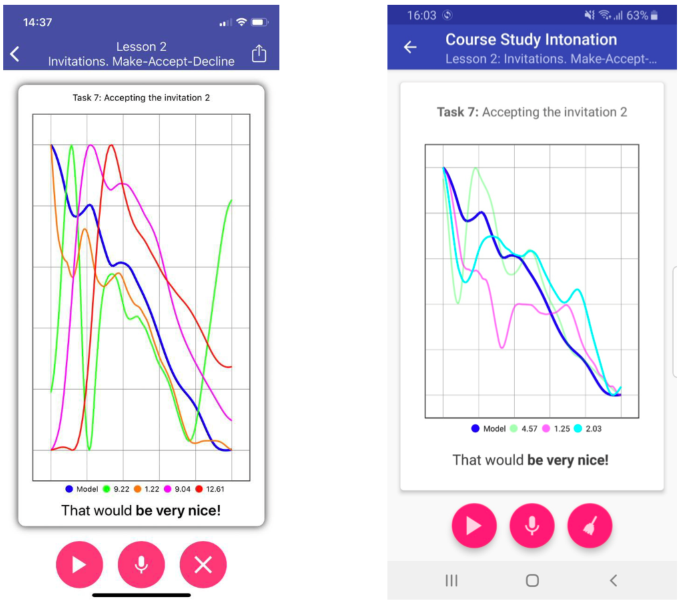
3.2.1 Course Editor Module

The *StudyIntonation* approach suggests the collaboration between teachers and native speakers in co-creation of the learning content and supposes various learning scenarios with respect to the learning style of a student. *StudyIntonation* consists of two modules: course editor for teachers and mobile application for learners. The editor or teacher can add pedagogically developed courses to the repository. These courses are available for learners to practice. Figure 3.1 shows the architecture of each module in the *StudyIntonation* system.

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

Creating courseware from scratch does not require any programming experience and can be completely implemented by a team responsible for content creation, which can include teachers, narrators, or actors [110]. The entire procedure consists of the following steps:

- Record intended pronounced phrases using an external sound editor (such as Audacity [111]) and store them as .mp3 files.
- Provide a text file with marked-up text subtitles in the form of t_{start}, t_{stop}, WORD (e.g., 2678 2687 "Hello").


14:37
Lesson 2
Invitations. Make-Accept-Decline
Task 7: Accepting the invitation 2
Model 9.22 1.22 9.04 12.61
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!

That would be very nice!

Figure 3.2: Task infterface of *StudyIntonation* for Apple iOS and Android.

- Input the reocording (as .mp3 file) to CourseInspector utility through the command line interface. This utility will produce a pitch file with the corresponding pitch sample set. A tuple of .mp3, .pitch, and .text files corresponds to one task in a lesson within a course. CourseInspector also performs the final validation of each task.

The course creation phase, which occurs before producing .mp3 recordings, is an inherent part of the process. This involves the planning, selection, and organization of course content. While this step might be time-consuming, it is an essential part where the teacher's input shapes the design of the courseware.

3.2.2 Mobile Application and DSPCore

The most important module of the system is a mobile application that can be used for university education and self-training. The initial version of the *StudyIntonation* prototype was developed for Android OS. Later, similar functionality was implemented for Apple iOS as an independent application. The efforts towards more similarity of both versions resulted in a cross-platform CAPT environment that works identically under Android OS and Apple iOS. The achievement of cross-platform functionality is primarily based on the selection of development tools and libraries available for both target operating systems. For instance, pitch graphs are processed and rendered using MP Android Chart and iOS Charts, and a similar approach was employed for other functional components, such as json parsers. Both Android and iOS applications use the same DSPCore library for pitch processing, resulting in identical pitch displays as shown in Figure 3.2. This library is integrated into Android and iOS applications via Java and Objective C wrappers, respectively. As *StudyIntonation* evolved into a full-fledged mobile CAPT, cross-platform, and multifunctional, it was validated with respect to formal evaluation methods of CAPT tools [112] to understand its soundness for linguistics and pedagogy [113].

The Digital Signal Processing (DSP) Software Core of *StudyIntonation* has been developed using C++ [113]. It is used to analyze and exhibit learner and model pitch curves and calculate metrics similar to correlation, mean square error, and dynamic time warping distance. However, this functionality caters only to the highest level of pronunciation instruction: phrasal intonation. In consideration of the necessity to display more intricate phonological aspects of pronunciation,

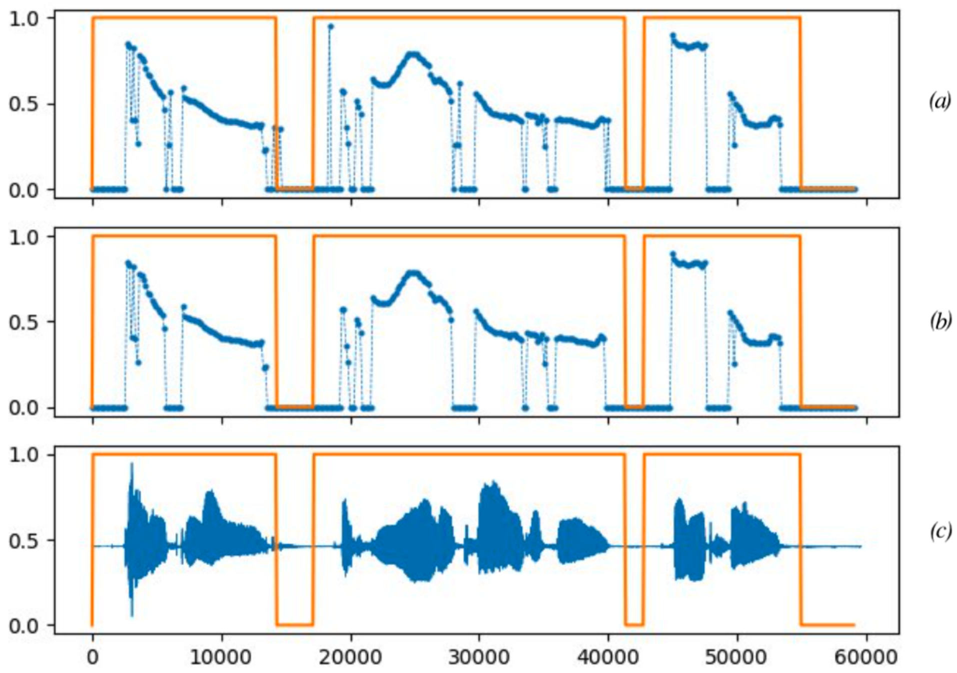


Figure 3.3: Raw pitch with thresholding (a), VAD applied (b), and speech signal (c) for the phrase "Yeah, not too bad, thanks!". Rectangle areas mark intervals with voice detected.

such as phonemes, stress, and rhythm, and in line with a comprehensive approach to Computer-Assisted Pronunciation Training, we incorporated the ASR system, Kaldi, equipped with the pre-trained ASR model, Librispeech, into the DSPCore.

As a result, in addition to the pitch processing thread equipped with Voice Activity Detection (VAD) functionality, three other output threads emerged from the ASR block: (1) for phonemes; (2) for syllables, stress, and rhythm; and (3) for recognized words. Currently produced offline for model speech, these features are anticipated to be integrated into the mobile CAPT to process learner speech in real time.

To test the DSPCore function, we used the prosodically labeled corpus IViE [114]. The intonation contours in the audio recordings of IViE speech stimuli were loaded into Praat [27] and DSPcore. The corresponding tone movements were observed for each IViE record that shows the tones indicated in the tone markup (H*, L+H*, L%, 0%, etc.).

Voice Activity Detection

The possibility of visual feedback for intonation has been realized through the extraction of the fundamental frequency, or pitch, a standard operation in acoustic signal processing. Pitch detection algorithms yield three vectors of data: timestamps, pitch values, and confidence. Yet, due to detection noise, pitch point clouds and discontinuities, and occasional prominences, the raw pitch series is not immediately discernible to the human eye and cannot be visualized "as is." Therefore, standard processing stages such as pitch detection, filtering, approximation, and smoothing are necessary to produce a pitch curve suitable for educational use.

In the context of live speech recording for pitch extraction, it is common practice to apply VAD to raw pitch readings before proceeding to other signal conditioning stages. This aids in eliminating possible recording imperfections. Consequently, VAD was integrated into the DSPCore of *StudyIntonation*. This VAD was realized using a three-step algorithm as per [115], with *logmmse* [116] utilized in the speech enhancement stage.

In Figure 3.3, the upper plot (a) shows the post-preliminary raw pitch thresholding within a range of 75 Hz to 500 Hz and at a confidence level of 0.5. Plot (b) illustrates the pitch curve after VAD, while plot (c) shows the speech signal over time (in milliseconds). The rectangular areas in all plots highlight the signal intervals where the voice has been detected. Pitch samples that fall outside of voiced segments are suppressed, effectively excluding background noise preceding and following the utterance, as well as hesitation and pauses.

Phoneme Processing and Transcription

The process of phoneme and rhythm analysis is performed by the ASR component of DSP-Core, which integrates the Kaldi ASR [117] tool with the pre-trained LibriSpeech ASR [118] model [1]. Kaldi [119] is a comprehensive open-source speech recognition tool, designed to allow the extraction of results at any intermediate stage of speech-to-text processing. There is an extensive selection of ASR models compatible with the Kaldi library. The ASpIRE Chain Model and LibriSpeech ASR are the most frequently used English speech recognition models. These models were evaluated based on vocabulary size and recognition accuracy. The ASpIRE Chain Model Dictionary comprises 42,154 words, whereas the LibriSpeech Dictionary includes only 20,006 words but showcases a reduced error rate (8.92% compared to 15.50% for the ASpIRE Chain Model). The LibriSpeech model was selected for mobile platform use with Kaldi ASR due to its compact size and decreased error rate.

The incoming audio signal is divided into equal frames ranging from 20 ms to 40 ms, since sound becomes less uniform over longer durations, leading to a corresponding reduction in the precision of the attributed characteristics. The frame length is set at Kaldi's optimal value of 25 ms, with a frame offset of 10 ms. Acoustic signal features such as Mel-frequency cepstral coefficients (MFCC), frame energy, and pitch readings are independently extracted for each frame. The decoding graph was constructed using the pre-trained LibriSpeech ASR model.

With a decoding graph and acoustic model, as well as the extracted features of the incoming audio frames, Kaldi performs lattice determinations. These form an array of phoneme sets, indicating the probabilities that the selected phoneme set corresponds to the speech signal. Therefore, once the lattice is obtained, the most probable phoneme is chosen from the sets of phonemes. Ultimately, the following information about each speech frame is generated and stored in a .ctm file:

- Unique audio recording identifier
- Channel number (if audio recordings are single-channel, it is 1)
- Timestamps of phoneme commencement in seconds
- Phoneme duration in seconds
- Unique phoneme identifier

The LibriSpeech phoneme list is then utilized to match the strings of the .ctm file with the corresponding unique phoneme identifier.

Rhythmic Pattern Retrieval

The phonemic-level text transcription produced by Kaldi was subsequently utilized to segment the source text into syllables, determine their lengths, highlight pauses, and construct the rhythm. The LibriSpeech model features a set of phonemes with four postfixes: 'B' (beginning), 'I' (internal), 'E' (ending), and 'S' (single). This distinctive attribute of LibriSpeech phonemes facilitates selecting and segmenting individual words from a phoneme sequence.

For most English words, the count of syllables aligns with the number of vowel phonemes. Syllabic segmentation is based on information on word boundaries and vowels within a word and also conforms to the rule of maximum consonant count at the start of a syllable [120]. This method has achieved a segmentation accuracy of 93%. The accuracy can be further increased by manually adding words that do not fit the general rule. At this point, it is also important to consider pauses between syllables and the duration of the syllables for subsequent extraction of rhythmic patterns. We store the data of syllables and pauses in the following data structure: data type; syllable or pause; duration; start time; maximum energy on a data interval; and an array of included phonemes. For pauses, the maximum energy is considered zero, and the phoneme array contains the phoneme that defines silence, marked *SIL* in LibriSpeech. The duration of a syllable or pause is computed as the total of all phoneme durations within a syllable. The relative start time is considered the start time of the first phoneme in the syllable.

The maximum energy of a syllable is determined using the MFCC acquired from all audio signal frames. This process involves identifying and combining frames related to a specific syllable and identifying the maximum frame energy within this area. By knowing the start and end timestamps of a syllable, it is possible to find the first frame that includes the syllable's start and the last frame encompassing the final phoneme, a process that is linear to the number of frames within the syllable. These derived syllable characteristics are then used to identify stressed and unstressed syllables.

In instances where transcription and rhythm are separately addressed, the former can be visualized as aligned orthographic and phonetic phrases. From a technical standpoint, all ASR-based features are incorporated into the following C++ software components: Transcription API, Transcription Recognizer, Transcription Analyzer, and Syllable Builder.

3.2.3 Pitch Estimation and Quality Metrics

In order to assess the proximity of the speaker's recorded pitch curves to the model, we formulated metrics that remain unaffected by the speaker's unique vocal characteristics and speaking rate.

Pearson's Correlation Coefficient and Mean Squared Error

Early versions of *StudyIntonation* used PCC and MSE. These are common measures that are used to quantify the similarity or dissimilarity between two sets of data points. When applied to pitch analysis in the context of language learning, these metrics can provide helpful information about how closely a learner's speech (the recorded pitch) matches a reference or ideal model.

PCC is a statistical measure that calculates the correlation, or linear relationship, between two variables. A PCC of +1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 indicates no correlation. In the context of pitch comparison, a high PCC (closer to +1) between the recorded pitch and the model pitch would indicate a strong similarity between the two pitches. It means that the learner's intonation follows the same pattern as the model, even if the exact pitch frequencies differ.

PCC is calculated as follows:

$$PCC = \frac{\sum_{i=1}^{n}(S^r(t_i) - \overline{S^r(t)})(S^m(t_i) - \overline{S^m(t)})}{\sqrt{\sum_{i=1}^{n}(S^r(t_i) - \overline{S^r(t)})^2 \sum_{i=1}^{n}(S^m(t_i) - \overline{S^m(t)})^2}} \tag{3.1}$$

where:
$S^r(t)$ and $S^m(t)$ are record and model pitch series, respectively,
$\overline{S^r(t)}$ and $\overline{S^m(t)}$ are the means of $S^r(t)$ and $S^m(t)$,
t – time parameter,
n is the total number of samples in the series.

MSE is a measure of the average of the squares of errors, that is, the difference between the model pitch sample and the recorded pitch of the learner. The smaller the MSE, the closer the learner's recorded pitch is to the model. However, unlike PCC, MSE does not tell us anything about the pattern of the speech, just how much the recorded pitch values deviate from the model on average.

MSE is calculated as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(S^r(t_i) - S^m(t_i))^2 \tag{3.2}$$

Dynamic Time Warping

DTW is a method that is often used in time series analysis. Its primary purpose is to find an optimal alignment between two sequences of data points, which may vary in speed or timing. In simpler terms, it is a way to measure how similar two sequences are, even if one of them is stretched or compressed in time compared to the other.

In the context of speech and language processing, DTW can be used to compare spoken words or sentences, including prosodic similarity [47, 48]. For instance, comparing a student's pronunciation with a reference one directly when the student speaks slower or faster, you will find many differences because the phonemes in the faster recording are played in a shorter amount of time compared to the slower recording. But if one uses DTW, the algorithm aligns the two sequences in a way that "warps" the time axis, essentially stretching the faster recording or compressing the slower one so that they match up better. Then, similarity between two recordings can be measured more accurately.

Figure 3.4 illustrates the DTW scores for the pitch quality estimation for the series of user attempts for the phrase *How is the conference going for you?* The first two screenshots illustrate that User 1, a native speaker, was able to adapt his intonation to the pattern of the model while gradually improving his scores after the series of 6 attempts. Having established that replication is possible, the next challenge was to test whether non-native speakers could also replicate the model utterances. The following two screenshots on the right-hand side of Figure 3.4 illustrate an interesting phenomenon, in which User 2, a non-native speaker, was able to outperform a native speaker.

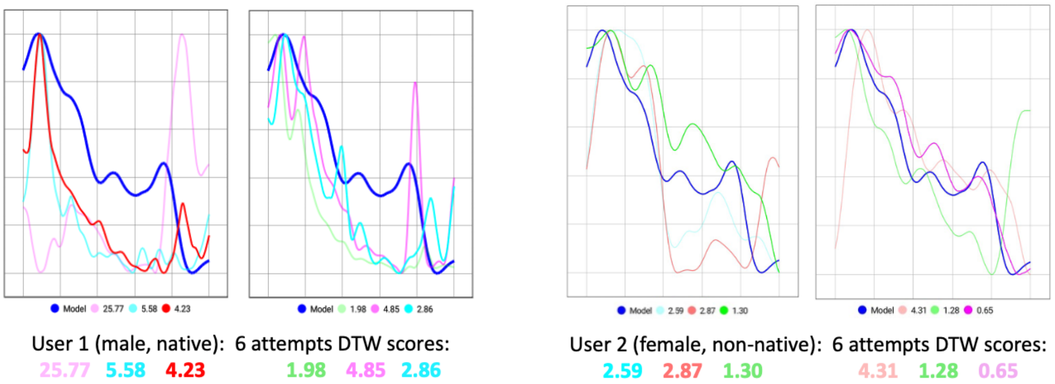


Figure 3.4: DTW scores for the pitch quality estimation achieved by both native and non-native speakers.

Though DTW provides an objective tempo-invariant primary estimation of the learner's ability to replicate the model, the pitch graph does not provide sufficient information to discover whether and to what extent improvements in prosody could be made. This means that the feedback combining pitch graphs and numerical scores is constructive and consistent, but not instructive enough.

Recurrence Quantification Analysis

Though pitch graphs with DTW estimation can provide an objective primary estimation, they still lack a corrective or instructive interpretation. Therefore, we need metrics that would enable the learners' progress and prosody production estimation. Larsen-Freeman pointed out that a systematic view is needed to promote the spontaneous use of intonation by L2 learners. Complexity Theory and Dynamic System Theory [121] act, at present, as our theoretical basis for second language development. Within a dynamic approach, the relationships of phonological features are understood as an interrelated dynamic system, giving the prospect of representation of individual language evolution dynamics.

Non-linear dynamics theory might be quantitatively incorporated in L2 pronunciation teaching using Recurrence quantification analysis (RQA) and CRQA [62], [6]. These metrics can contribute to further improvements in constructing an instructive and more personalized analysis of synchronization between the initial model, the learner, and the referential native speaker's attempts. Chapter 4 is dedicated to the use of RQA and CRQA for dynamic assessment during suprasegmental training.

3.3 Summary

In summary, this chapter introduces *StudyIntonation* software for prosody learning that uses algorithms and data models for prosody recognition, classification, approximation, visualization, and estimation. This makes *StudyIntonation* a good platform to test the ideas presented in this work. Effective online learning requires more than the digitization of language learning materials; it requires learners to interact with the study materials. Rapid technology transformation requires further digitalization, which is the use of digital technologies and data to transform processes and create an environment, where digital technology brings completely new possibilities. From the learner's perspective, the *StudyIntonation* environment and the corresponding mobile tools provide an authentic speech context with real-time pitch graphs of speakers, perform learners' speech recording, display model speakers and learners' pitch graphs to generate contrastive feedback and calculate speech quality measures aimed at progressively improving learners' speech. By offering different ways of perception for the activities, the system addresses the problem of supporting a diversity of user learning styles. According to the review [122], an approach used in *StudyIntonation* is promising in its goal of serving as a mobile-assisted pronunciation training tool for classroom and individual learning purposes, however, there are still open issues to be addressed in future application releases. We hope that this work will help with that.

Chapter 4

Dynamic Assessment during Suprasegmental Training with Mobile CAPT

This chapter attempts to explore suprasegmental teaching and learning as a complex dynamic process that features variability, self-organization, and emergence. A review of the relevant literature and definitions can be found in Section 2.3.2. We focus on the concepts Dynamic System Theory (DST) and Sociocultural Theory (SCT) L2 pedagogy and propose how Dynamic Assessment (DA) could be applied in the *StudyIntonation* CAPT system. Our research suggests that SCT-based L2 suprasegmental learning, equipped with a dynamic model and CRQA, can explain the processes involved in a learner's interaction with a suprasegmental-focused CAPT.

The following questions were used to guide our investigation:

1. How a dynamic model could be applied to L2 suprasegmental teaching and learning with a CAPT system?
2. How could the development processes of the learners in terms of their Zone of Proximal Development (ZPD) be observed in the course of suprasegmental training?
3. How could the DA approach based on the developmental trajectories of the learners add to more personalized CAPT feedback and instruction focus?

4.1 Methodology

Research informed by DST is defined and shaped by specific questions under consideration, such as the emergence and dynamics of specific language skills [56, 123], but an explicit form of applicability of dynamic models in empirical L2 research designs is still needed [124, 125]. We adopted a dynamic model for the Vygotskian developmental mechanism from [53, 61, 72] for a narrow context of the interaction of a CAPT system.

We used the learning content of *StudyIntonation*, i.e., a subset of English phrasal intonation patterns, as the array of external and internal contents [53]; metrics based on DTW [48] and CRQA descriptors as indicators of learning dynamics (in the form of phase shifts, developmental jumps, etc.) [62]. We searched for quantitative indicators of learners' entrance into their ZPD in the course of CAPT system interaction along with spotting the specific tasks, which could be of maximum usefulness because of sensitivity, responsiveness, and perception increase while being within one's ZPD. Therefore, we obtain individually tailored DA-based feedback by performing assessment and instruction.

4.1.1 Dynamic Model of Vygotskian Developmental Mechanism

Dynamic system's current state generates its successive state by a rule of change ("evolution rule", the driver of its change) and thus produces a trajectory in a state space [61]:

$$x_{t+1} = f(x_t) \rightarrow x_{t+2} = f(x_{t+1}) \rightarrow \dots \tag{4.1}$$

Van Geert [72] incorporates the Vygotsky dialectical mechanism of development into the dynamic model in the following way: performing an action, the system (e.g., a learner) activates a particular content c_n (a pattern, a skill, a rule, etc.), which is associated with a specific developmental level n that is responsible for that action. Any event which can be either an action or an experience is a confluence between an internal content $c_n \in I$, $I(c_1, c_2, ..., c_n, ...,)$ and an external content $c_n \in E$, $E(c_1, c_2, ..., c_n, ...,)$ which share the same developmental level index n. $I \subseteq E$, that implies the environment E is viewed as a potential source of learning and development.

The effect of an action on the further development of the system depends on two contents – the first is the activated content c_n, which represents its actual developmental level A_i, which Vygotsky calls the Actual Development Zone (ADZ); the second content c_k, $k > n$ is defined by the help or information or feedback resulting from performing the action. This second content defines or specifies the potential level of the system P_i, which is a developmental level corresponding to a set of contents (patterns, skills, rules, etc.) in the array I that is most sensitive to the effect of experience brought about by the activated content c_k. This sensitivity to instruction is what Vygotsky contended with his ZPD concept.

The existence of content that is more sensitive to experience than the others is, according to [53], based on two opposing tendencies that are likely to occur in learning and developing systems: preference for novelty and preference for familiarity. These tendencies can be expressed in the form of a pair of exponential function [53, 72]:

$$f_{familiar}(i) = ab^{ci} \tag{4.2}$$

$$f_{novel}(i) = dg^{fi} \tag{4.3}$$

where $a, d, f > 0, b, g \in (0, 1), c < 0$, and i is the distance between contents in the array I. Van Geert [53] names the most preferred content for both functions as the one at the cross-section point of $f_{familiar}(i)$ and $f_{novel}(i)$:

$$i = n + \Delta M = \frac{\log\left(\frac{a}{d}\right)}{f\log(g) - c\log(b)} \tag{4.4}$$

Each time the system has gone through an action/experience the content arrays are updated. The maximal gain occurs at two places: at the content corresponding with the actual output level A_t, and the balance point of maximum sensitivity to experience defined by equation 4.4.

The mathematical model for Vygotsky ZPD concept of learning includes two evolution rules for actual and potential developmental levels [72]:

$$A_{i+1} = A_i(1 + R_{A_i} - R_{A_i}\frac{A_i}{P_i}) \tag{4.5}$$

$$P_{i+1} = P_i(1 + R_{P_i} - R_{P_i}\frac{P_i}{P_k}), \tag{4.6}$$

where R_{A_i} – learning rate, R_{P_i} – teaching rate, P_k – goal state. The learning and teaching rates update rules are defined as:

$$R_{A_i} = r_A - |\frac{P_i}{A_i} - o|\alpha(P_k - P_i) \tag{4.7}$$

$$R_{P_i} = r_P - |\frac{P_i}{A_i} - o|\beta(P_k - P_i), \tag{4.8}$$

where r_A, r_P are constant growth factors, o sets the optimal $\frac{P}{A}$ ratio, α, β are damping parameters. This model describes the reciprocal interaction between the principal variables A_i and P_i with the support of a set of control variables. The effect of the dynamic model is that it allows for observing the learner's gradual transition from one developmental level to another.

4.1.2 CAPT Environment as a Dynamic Model with DTW and CRQA Metrics as Developmental Descriptors

DST was shown to be sound and beneficial in digital environments for language learning [75, 126], where the importance and contribution of multimodal input was supported by neuroimaging studies on the optimal operation of the human brain in multisensory environments [127].

Interaction with *StudyIntonation* is multimodal, involving listening to a recording, observing a pitch curve, and shadowing a model phrase. As a means of feedback, the system produces a pitch curve for recorded speech and calculates DTW-based similarity metrics. The learning content of the English course of *StudyIntonation* is produced by native speakers. It covers a set of 74 examples of phrasal intonation patterns and is structured into 4 groups with respect to various speech situations. This approach is in line with recent SLD research, which highlights the influence of social factors on language acquisition and argues a holistic, top-down approach as paramount for L2 pronunciation instruction [60].

Most CAPT resources still need accurate instructive feedback and learning strategies [58, 128]. As suprasegmental L2 teaching and learning may be understood as a dynamic process, we searched to improve the instruction and feedback mechanism by mapping CAPT context to a dynamic model (Table 4.1).

CAPT system	**Dynamic model**
Learner (narrow context)	Internal array of contents, skills, etc. $I(c_n)$
Courseware (source of experience)	External array of contents, events, patterns, etc. $E(c_n)$
A specific pattern (task)	Familiar content in ADZ c_n Novel content in ZPD c_k
Performance metrics (DTW, CRQA)	Developmental indicators

Table 4.1: Correspondence between DST concepts and components of CAPT System for suprasegmental learning

The selection of DTW and CRQA as developmental indicators is based on the fact that when monitored over time, they reflect the alterations or embody characteristics of intonation acquisition and model/learner prosodic synchronization. The capacity of CRQA metrics to differentiate and describe various L2 DST-based research, speech synchronization, and emotion recognition tasks has been demonstrated multiple times (see, for example, [62, 129]), while DTW is known to be a traditional measure of pitch curve similarity [48].

The interpretation of DTW-based performance metrics in terms of dynamic model of ZPD relies upon the admission that a specific DTW-based metric depends upon the familiarity/novelty of a specific content and may be understood as a familiarity/novelty measure of a specific experience, thus a DTW-based metric could be used as a visible indicator of a specific content activation. Therefore, the variability of the DTW metric should signal the transition of the learner from one developmental level to another [56] and produce unimodal and bimodal curves. The transition is detected by the emergence of a second peak yielding a bimodal pattern. After the transition, the first peak disappears and the original unimodal pattern at a higher level is restored [53].

4.2 *StudyIntonation* Dataset Collection

A dataset of 1050 records of 1 British English native speaker and 5 non-native speakers with L1 Russian doing *StudyIntonation* tasks was collected. A group of learners performed a shadowing task during one-hour learning sessions occurring intermittently over a 24-month period.

The whole observation time was split into task-wise and session-wise timescales:

1. Two successive attempts of one task (30 sec);
2. All attempts for a specific task (15 min);
3. All attempts for all tasks covered within a learning session (45-60 min);
4. All attempts for a specific task within the observation period (24 months); and
5. All attempts for all tasks within the observation period (24 months).

Since pronunciation dynamics demonstrates short-term and acute shifts [130], the most significant developmental effects were obtained for scales 1 and 2 (Fig. 4.1).

Each record in the dataset bears the following markup:

1. Subjective expert decision whether learner's attempt is good or not (binary, 0 or 1);
2. DTW metric of model and learner pitch similarity;
3. CRQA metrics: cross-recurrence rate (RR), percentage of determinism (DET), average diagonal length (AVG_DIAG), etc. [82].

The embedding dimension [82] for CRQA metrics $d \in [2, 4]$ for pitch contours was obtained by the False Nearest Neighbors algorithm (FNN) [131]. This value complies with DST-based speech f_0 extraction algorithms, (e.g. [80, 81]), where the number of embedding dimensions for f_0 was shown to be within an interval of 2 to 5. CRQA metrics were calculated for embedding dimension, $d = 3$ and point proximity radius, $\epsilon = 2$.

4.3 Results and Discussion

4.3.1 DTW and CRQA as Joint Developmental Descriptors

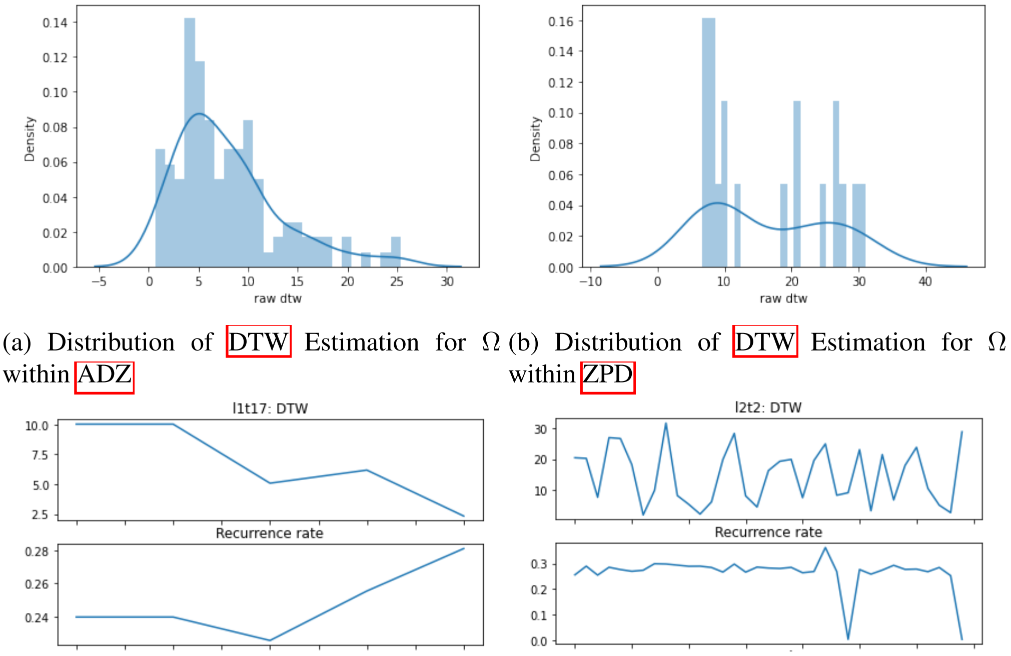
To explore whether using DTW and CRQA metrics together adds to the assessment of learner performance more than when used separately, we trained logistic binary classifiers to discriminate learner attempts from the *StudyIntonation* dataset and examined the importance of the features of DTW and CRQA (Table 4.2). Logistic regression classifiers were trained using DTW together with various CRQA feature subsets as a part of the input. Table 4.2 shows classifier accuracy for DTW; DTW, RR; and RR, DET feature sets, where a combination of DTW and RR shows the best discrimination ability.

- DTW-based classification had the lowest accuracy of 0.60.
- CRQA-based classifier had relatively poor accuracy of 0.65.
- DTW+RR-based classifier achieved better accuracy of 0.71.
- Other CRQA features, when included into the feature set, confused the logistic classifier.

Feature set	Classifier metrics (p – precision, r – recall)	Accuracy (F1)
DTW	p_0: 0.65 r_0: 0.56 p_1: 0.56 r_1: 0.65	0.60
RR, DET	p_0: 0.70 r_0: 0.61 p_1: 0.61 r_1: 0.69	0.65
DTW, RR	p_0: 0.75 r_0: 0.68 p_1: 0.67 r_1: 0.73	0.71

Table 4.2: DTW and CRQA feature sets

4.3.2 Transition to ZPD. DA within CAPT interaction



(a) Distribution of DTW Estimation for Ω within ADZ

(b) Distribution of DTW Estimation for Ω within ZPD

(c) DTW+RR Joint Dynamics in ADZ: Task 17 of Lesson 1
”I have to cancel the meeting.”

(d) DTW+RR Joint Dynamics in ZPD: Task 2 of Lesson 2
”Would you like to visit the museum?”

Figure 4.1: Microgenetic effects at 15 min timescale, which covers all attempts for a specific task within one training session of one learner X. Transition to ZPD is located by oscillation of DTW between two levels

In [56,57,65] it was shown that learners’ performance does not increase linearly, but passes through periods of progression and regression alternatively. These are not isolated jumps, but the stages of a continuous developmental process; and each individual learner demonstrated unique patterns of this developmental trajectory.

We located a transition and, thus, a ZPD of each learner by oscillations of DTW and observed an increased responsiveness to input audiovisual stimuli, manifesting itself by the occurrence of low DTW. The rate of cross-recurrence RR in ADZ tends to grow smoothly alongside with a decrease of DTW (Fig. 4.1a,4.1c). During the state of transition, DTW oscillates between two levels, while RR dynamics may be either indifferent or oscillate randomly (Fig. 4.1b, 4.1d). We fixed the specific types of tasks, where such an oscillation started to occur and where learner’s efforts should be directed to (Table 4.3). Although not the focus of this research, a notable finding was that a non-native speaker was, at times, able to replicate the model more accurately than the native speaker. It appears that possessing “a good ear” may be a more important determinant of success than the mother tongue.

ADZ
L1T17: *I had to cancel the meeting.*
L1T18: *I’d really appreciate that.*
ZPD
L1T19: *Are you going for lunch now?*
L2T1: *Would you like to join me for dinner?*
L2T2: *Would you like to visit the museum?*
L1T25: *I’m glad, that’s really kind of you, thank you!*

Table 4.3: Example of DA-driven task assignment for *StudyIntonation* in accordance with learner’s ADZ and ZPD location

4.4 Summary

Teaching individual segmental and suprasegmental features can positively influence the global construct of L2 pronunciation proficiency [132, 133]. Maximum sensitivity to particular contents of developmental levels means that experiences at these levels yield a maximal effect [53]. The main outcome of this research is how to perform DA during the teaching of phrasal intonation with a CAPT system and how to determine learners’ movement from one developmental level to another. While working through ADZ, DTW metrics are either immediately low or rapidly and steadily converge to small values. A good rising edge of RR is present, which indicates that two phonological systems are synchronizing with each other. When the transition to a new level (ZPD entrance) is approaching, there is a group of tasks where DTW is high, but immediately after instruction there is a short effect of prosodic memory, which results in a low DTW metric for one attempt. The student produces a good result, but cannot hold this effect longer. This variability signals a maximum sensitivity to instruction and experience. It is necessary to spot the type of tasks, where oscillations occur, specific for each learner, and direct the focus of efforts there. In the example in Figure 4.1, these are longer interrogative and exclamatory sentences (Table 4.3). A distant ZPD is formed by contents where learners cannot yet access the model and all indicators are unstable; but step by step, as a consequence of teaching in the ZPD, learners become more familiar with contents that are ahead of their current actual developmental level [56, 58, 61, 123].

Chapter 5

Classification of Accented English Using Deep Learning

In the context of CAPT feedback personalization, accent recognition can be used to assess the pronunciation of a foreign language in computer-assisted pronunciation training systems, especially together with automatic speech recognition (ASR). The core problem is that conventional acoustic language models adapted to fit standard language corpora cannot satisfy the recognition requirements for accented speech. Successful accent identification can be one of the characteristics that provides an opportunity for more informative and better personalized feedback and exercises to language learners according to their manner of speaking [89,94] as speakers with the same accent have been observed to have similar trends in mispronunciation [9]. Therefore, a more accurate accent classification can help improve the robustness of mispronunciation detection and diagnosis to mitigate the adverse effects of accent for the benefits of CAPT systems [95,134].

In Section 5.1, we describe the relevant research work that contributes to the solution of the accent recognition problem, along with positioning our work in the framework of existing models. Section 5.2 introduces the methodology, including CNN construction, accent detection, feature selection, data collection, model parameter classification, and the tools we used. In Section 5.3, the experimental results are presented across hyperparameter selection, regularization, and with respect to different sets of audio signal features used by the CNN classifier. Section 5.4 reports the evaluation approach using standard Information Retrieval (IR) metrics, including accuracy, precision, recall, and F1.

5.1 Scope of Research

In this section, we explain the accent groups we focus on and analyze the existing models for accent classifiers and the input speech signal parameters regularly used in related works. We also extract the relevant research questions such as inner model configuration, optimal feature set and rhythm impact. Table 5.1 lists the papers which are the closest to the scope of our study.

As a sub-task of speech and language recognition, accent detection algorithms are built using the standard classification models and machine learning architectures including CNN [85, 90, 136, 141], feedforward neural networks (FFNN) [89], hidden Markov model (HMM) [92], k-nearest neighbor (KNN) model [142], Gaussian mixture model (GMM) [143,144], long short-term memory (LSTM) and bidirectional LSTM (bLSTM) [145,146], Random forest, and Support vector machine (SVM) [92,142,144,147,148].

The accuracy of accent classification is significantly affected by the selection of input features. As of 2022, the best experimental results had been achieved using mel-frequency cepstral coefficients (MFCC), along with other types of input features including spectrogram (SG),

Paper, year	Feature set	Model	Classes	Accents	Dataset
[135], 2022	Mel-spectrogram	CNN	5	5 Kashmiri accents	Custom
[136], 2021	SG	CNN (LeNet)	5	DU, FR, JA, NS, PO	IViE, Cambridge English Corpus
[91], 2021	SG	CNN	5	AR, FR, GE, IN, NS	SAA
[85], 2020	MFCC, SG, CG, SC, SR	CNN	5	AR, FR, NS, SP, ZH	SAA
			3	AR, NS, ZH	
[137], 2020	MFCC	CNN with attention	2	IN, NS	Custom
			4	IN	
			9	IN, NS	
[138], 2020	MFCC	Logistic Regression	3	HA, IG, YO	Custom
[92], 2019	MFCC	LSTM, RF	4	NS, SP	Custom
[89], 2017	MFCC, LPCC	FFNN	6	GA, IN, IT, JA, KO, NS	Wildcat
[90], 2017	SG	CNN (AlexNet)	3	NS, SP	SAA
[139], 2017	MFCC	GMM	3	ML	Custom
[140], 2012	Mel-spectrogram statistics	FF-MLP	3	IN, MS, ZH	Custom
[88], 2005	2nd and 3rd formants	GMM	2	IN, NS	Custom (SAA subset)

Table 5.1: Retrospective summary of related works

chromagram (CG), spectral centroid (SC), spectral rolloff (SR), and mel-weighted single filtered frequency (SFF) spectrogram [85,148].

In particular, Singh, Pillay and Jembere [85] managed to achieve a maximum accuracy of 53.92% while classifying 5 accents and 70.38% for 3 accents using mel-cepstral coefficients, extracted from three-word audio segments. The authors used the Speech Accent Archive dataset [149].

Based on the AlexNet architecture, Ensslin et al. [90] trained a tailored classification on 3 accents using 227x227 spectrogram images as input features applied to the same Speech Accent Archive dataset. The authors reported CNN accuracy of 61% while recognizing among 3 English accents, namely British, American, and Spanish. In [137], the authors defined a list of requirements and collected a dataset that complies with these requirements to test a number of different classifiers including CNN, and CNN with an attention mechanism. Using MFCC as input features in CNN with the attention mechanism showed the best accuracy with up to 100% for 2 classes, 99.0% for 4 classes, and 99.5% for 9 classes.

The highest average recognition accuracy was achieved with the combination of MFCC and FFNN, thus, giving 91.43%, compared to 78.73% when combining LPCC and NN, and 87.55% – for the case when combining MFCC and GMM [89]. Yusnita et al. [140] used a feed-forward multilayer perceptron (FF-MLP) consisting of two layers as a classifier, and achieved a maximum recognition accuracy 99.01% while classifying among 3 accents using their own dataset of audio recordings. As input features, the statistical parameters of mel-spectrograms were used.

In ASR, human speech can be described by various phonetic and prosodic features that affect the perception of accent to varying degrees. Speech is a multi-layer structure which can be analyzed at different levels from sounds (phonemic sequences) up to melody and rhythm. Meanwhile, the physical and acoustic features of the speech signal can be considered as well. As we can see from the recent research work described earlier, using MFCC as input characteristics is one of the most common approaches in ASR solutions [85,89,92,137]. Specifically, in [85] the accuracy of accent classification was evaluated using MFCC, spectrogram, chromogram, spectral centroid, and spectral rolloff as input features. The authors of [85] also suggested that further experiments are required to check the promising case of combining MFCC with other types of available characteristics.

In this chapter, we test this hypothesis using MFCC in combination with other spectral characteristics. Specifically, we investigated whether using time-frequency and energy features (as recommended in [150]) could improve the automatic accent detection accuracy when used jointly with MFCC as input features. We describe the experimental environment and results, demonstrating that the greatest contribution to recognition is made by the presence of stable time-frequency patterns of energy distribution, represented by amplitude mel-spectrograms on a linear scale, which alone could be fed into the classification model, as the mel-spectrogram captures all the relevant pronunciation-specific details [151].

5.2 Materials, Methods and Tools

A standard approach used in ASR assumes that CNN works with inputs which are, in fact, two-dimensional images representing the audio signal features [85]; thus, the number of neurons in the CNN input layer is equal to the number of characteristics of each feature vector [89]. The output value of the accent detection classifier is the probability distribution vector which attributes the speech sample to a specific accent class (where the classes correspond to the languages).

5.2.1 Adopting CNN Model to Speech Signal Processing

Sound waves are complex non-stationary signals, which explains why direct classification of sound recordings is rarely used. Selecting and extracting the best representation of an acoustic signal is an important task in ASR design, since this decision significantly affects the recognition quality and efficiency.

Accent can be understood as a composition of the phonemic and prosodic components of pronunciation: sounds, linking, intonation, stress, rhythm, etc. Accent-sensitive information could be obtained from the signal directly, but the more conventional way is when a raw audio signal undergoes a specific time-frequency transformation to calculate more sophisticated speech parameters, e.g. MFCC [85]. After extracting features, machine learning methods perform accent classification [89, 92, 150]. Our methodology is largely defined by the decisions which should be made in the course of all stages of automatic accent detection and considers four main topics – dataset, feature selection, batch normalization, and machine learning model design.

5.2.2 Data collection

All experiments were performed with speech samples from the Speech Accent Archive [149] maintained by George Mason University. The Speech Accent Archive is a crowd-sourced collection of speech recordings of the following text passage.

> *"Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station."*

The archive contains meta-information about the demographic and linguistic background of speakers. In total, the archive contains 2,982 samples (as of last known update at `https://accent.gmu.edu/`) recorded in more than 200 different native languages. Accents are classified according to the native language of the speaker.

It is important to mention that the dataset from the Speech Accent Archive conforms to the major ASR suitability requirements, namely:

- **Speaker diversity** assuring an adequate representation of different varieties of pronunciation;
- **Uniformity of material** referring to the same content and context;
- **Phonetic balance**, when individual phonemes do not occur too often;

- **Presence of a semantic load of sentences** avoiding semantic factors that might affect pronunciation [137];
- **Working with speech segments** rather than independent words.

The latter aspect is extremely important, since pronunciation patterns for words spoken separately differ from phrasal patterns expressed in the context of related speech because of eventual assimilation (in which phonemes become similar to neighboring phonemes) or elision (where phonemes are omitted).

Data Classes (L1 Languages)

We used nine languages that were groupped and labeled into several groups: Germanic languages (English (EN), German (GE), Dutch (DU), Swedish (SW)), Romance languages (Spanish (SP), Italian (IT), French (FR)) and Slavic languages (Polish (PO), Russian (RU)). Uneven distribution of recordings belonging to the different classes might deteriorate the accuracy of recognition for classes represented with fewer examples, thus, worsening the quality of the classification in general. On the other hand, using all the available examples belonging to classes containing a much larger number of recordings might lead to heavier computations without significant improvements in recognition accuracy. Therefore, for larger groups we limited the number of samples used by 80 recordings.

The distribution of available recordings during the experimental period according to L1 classes was as shown in Figure 5.1.

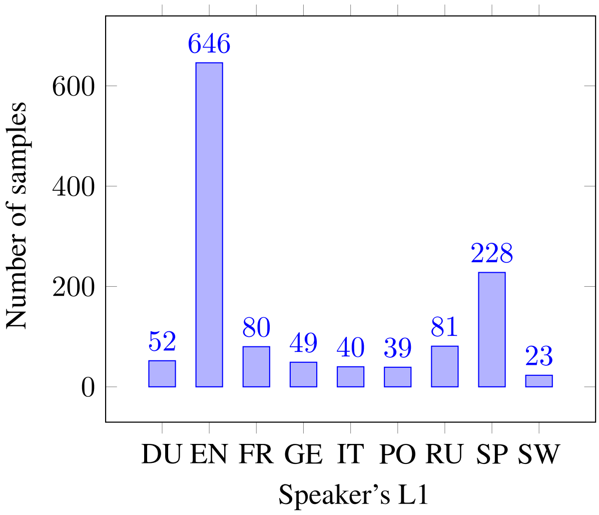


Figure 5.1: Distribution of audio recordings by classes (during experimental period)

Preparing audio files for recognition

The problem of speech signal recognition differs from the recognition of static images. In speech recognition, the object of analysis is the dynamic process and not a static image or pattern. Thus, a recognizable speech pattern is represented by feature vectors rather than a single vector. Since the presence of an accent is affected by many factors, people can have a hybridization of accents. Recognizing the accent at any point in time may be a better solution than over the entire audio signal, that is why signal segmentation is used. According to [85], classifying short segments of an audio file will more accurately classify the speaker's accent.

Thus, audio recordings with a sampling rate of 22050 Hz were split into multiple consecutive frames of 25 ms, each with an overlap of 10 ms based on the experimental investigations discussed in Section 5.3.2.

The downside of using crowd-sourced datasets is that neither the recording environment nor the recording equipment is consistent between speakers, resulting in significant sample noise and differences in recording volume [137]. Therefore, in order to reduce the differences between audio recordings in the form of linear distortions, before training the model, the obtained data need to be normalized within each audio recording, for example, using z-normalization (using z-score):

$$x' = \frac{(x - \mu)}{\sigma}, \tag{5.1}$$

where μ – mean value, σ – standard deviation.

Fragments of Silence

Table 5.2 reports our experiments on how the presence or absence of pauses in audio files affects the classification results. There is another important but contentious aspect, whether it is necessary to keep or remove fragments of silence (pauses) from the input to achieve the best recognition quality. To solve this, we arrange experiments for both approaches. For these preliminary investigations we used a restricted set of characteristics that included 13 MFCCs and fundamental frequency $F0$ only.

Table 5.2: Comparing classifiers with preserved or removed fragments of silence

	Fragments of silence			
	Preserved		Removed	
L1	Accuracy	Error	Accuracy	Error
EN RU SP SW	0.71	0.83	0.70	0.84
FR IT SP	0.71	0.73	0.68	0.82

As we can see in Table 5.2, the presence of pauses can be a strong indicator of a foreign accent. For Romance languages, the difference is more noticeable than for Germanic ones. Therefore, we decided to keep the fragments of silence in audio recordings for further accent classification experiments. Thus, the audio files processed in all the subsequent experiments are used in their original form, i.e., with all the pauses preserved.

5.2.3 Feature Selection

A common approach to speech signal processing is to use short-term analysis, assuming that the signal characteristics remain unchanged within a short time frame. Thus, speech utterances are compared to the feature vectors, presumably differing in their distribution with different speakers' L1s. For speech signal analysis, the frame length is close to 10-30 ms, with an overlap between the frames is equal to approximately half their length [89]. The signal is split into 25 second fragments overlapping by 10 seconds.

The characteristics of the audio signal are extracted from the frames using an applicable feature extraction method, such as constructing a compact representation of an audio signal using a set of mel-frequency cepstral coefficients resulting from a cosine transformation of the real logarithm of the short-term spectrum, represented on a mel-frequency scale [150], the latter being arguably based on studies of the ability of the human ear to perceive sounds at different frequencies [89].

MFCC uses a spectrum capable of reflecting a phoneme utterance representing a curve in the amplitude-frequency plane, which makes it applicable to speech recognition tasks [85]. To find MFCCs, the signal is divided into short frames, then a window function is used. The discrete Fourier transform is performed giving a periodogram of the original signal. Filters are applied to the periodogram, evenly spaced on the mel-axis, which yields the output in the form of a spectrogram. The spectrograms can be then represented on a linear or logarithmic scale. The last step in finding the MFCCs is to apply the discrete cosine transform to decorrelate the resulting coefficients. Since the human ear perceives a limited range of frequencies, ASR problems usually use the first few MFCCs as input features, often limited to 13 MFCCs [85,137]. In our work, we use the same number of MFCCs (13).

Our feature set was formed on the base of **amplitude mel-spectrograms on a linear scale**. The audio signal frequencies f were converted to mel-spectrograms $M(f)$ as follows:

$$M(f) = 2595 \log_{10}(1 + \frac{f}{700}) \tag{5.2}$$

Linear amplitude mel-spectrograms were chosen because they performed better at classifying accents than logarithmic amplitude mel-spectrograms, power mel-spectrograms, and SFF mel-spectrograms. By experimenting with mel-spectrograms with 32, 64 and 128 bands as input features, we found that the optimal balance between learning rate and recognition accuracy can be achieved using mel-spectrograms with 64 bands (see Section 5.3 for details).

As suggested in [85], combining MFCC with additional features can contribute to further improvements in recognition accuracy. We organized a number of representative experiments to verify this hypothesis. We experimented with six additional features used to extend the MFCC-based model:

- **Spectral centroid** (SC) indicates the frequency at which the energy of the spectrum is concentrated, or where the center of mass of the sound is located.
- **Spectral roll-off** (SR) is a measure of the asymmetry of the spectral shape of the signal. SR represents the frequency below which a given percentage (85%) of the total energy of the spectrum lies. This value is used to determine vocalized sounds in speech since unvoiced sounds have a large proportion of the energy contained in the high frequency range of the spectrum.
- **Chromagram** is usually a 12-dimensional feature vector representing the amount of energy for each of the signal's height classes (such as C, C#, D, D#, E, etc.).
- **Zero Crossing** (ZCR) represents the number of changes in the signal sign within a segment. ZCR can be helpful in describing the noisiness of the signal. For unvoiced speech, the ZCR characteristic takes on higher values.
- **Root mean square** (RMS) is a standard measure representing the average signal strength. Calculating RMS directly from the audio recordings is faster because it does not require calculating STFT. However, using a spectrogram can give a more accurate representation of signal energy over time because its frames can be split into windows. Since the characteristics of the signal can be stored in an external file in advance (before training the model), decreasing the extraction time is not critical. That is why, in our case, to improve the signal representation accuracy, RMS is calculated based on the signal spectrogram.
- **Fundamental frequency** (F_0) is the lowest frequency of the periodic signal. F_0 is the frequency at which the vocal cords of a person vibrate while producing the voiced sounds. The fundamental frequency F_0 carries a lot of information about the pitch of the voice at any given time and, therefore, about the overall intonation of the speech. It has been studied that F_0 makes a significant contribution to the perception of foreign accents [86], which is especially noticeable for Germanic and Romance languages [152]. Estimation of the fundamental frequency of the signal is carried out using the autocorrelation-based YIN algorithm [153]. According to this algorithm, a cumulative mean normalized difference function is computed for short overlapped audio fragments. Then, the smallest delay that gives the minimum of the normalized difference function below the threshold is chosen as the period estimate of the signal. Finally, the period estimate before converting to the corresponding frequency is refined using parabolic interpolation. Since there is no upper limit to the frequency search range for YIN, this algorithm is also suitable for higher

voices. In addition, YIN is a relatively simple algorithm that requires few parameters to be tuned.

To sum up, the first input feature set includes 30 audio descriptors, namely: 13 MFCCs, 12 chroma coefficients, SC, SR, ZCR, RMS and F0.

5.2.4 Batch Normalization

Training a deep neural network is a complex process involving the distribution of the input data for each layer; changes in the parameters of the current layer impact subsequent layers. Thus, small changes in network parameters are amplified as the network gets deeper. This, in turn, slows down training because it requires a lower learning rate and careful parameter initialization. This phenomenon is often called "intrinsic covariant shift" [154]. In this case, covariance refers to feature values and the issue of possible internal covariant shift can be resolved by batch normalization [155]. Batch normalization can improve the performance of artificial neural networks, even in the presence of correlation between input values, while being part of the model architecture and being performed in hidden layers for each mini-batch during the training stage. The use of mini-batch is preferred over separate input values at each training step.

Covariant shift poses a problem in machine learning because the learning function tries to fit the training data, and should the distribution of the test and training data differ, using the learning function may lead to erroneous results [156]. Commonly used machine learning methods work well under the assumption that the input parameters in the test and training samples belong to the same feature space and have the same distribution. In this case, when the distribution changes, the underlying statistical models need to be rebuilt from scratch using the new training data [157].

5.2.5 Classification Model

The classification model for accent detection is built on CNN used in [85]. The model consists of two convolution layers with a ReLU activation function. The first and second convolution layers contain 32 and 64 blocks, respectively. After each convolution layer, batch normalization and pooling are applied. The flatten layer is followed by two dense layers of direct propagation.

The first dense layer consists of 128 neurons and has the ReLU activation function. For the second layer, we set the number of neurons equal to the number of accents and use the *softmax* activation function. The input of the model is a feature matrix extracted from audio signals.

Following the approach in [85], for the basic implementation of the model, the convolution filters with size (3, 3) and pooling layers (2, 2) with a stride of 2 are selected. To avoid overfitting, we use the dropout method with a variable probability of any neuron turning to zero – depending on the type of input data, a value from 10% to 50% is used. We use categorical cross-entropy as a loss function during training.

The earning loss function is minimized using the adaptive moment estimation (Adam) algorithm [158], where the constant learning rate coefficient is 0.001, and the parameters β_1 and β_2 are 0.9 and 0.999, respectively.

The test data are about 25% in the case of using mel-spectrograms as input data and 15% in other cases. Figure 5.2 draws the workflow.

As tools for CNN modeling, training, implementation, and visualization, we use a number of standard Python libraries. In particular, the accent classifier is implemented and trained using the Keras library providing a high-level interface to the Tensorflow computing platform. Librosa digital signal processing library is used for audio signal processing and extraction of input characteristics. The classification quality metrics are calculated using the Scikit-learn package.

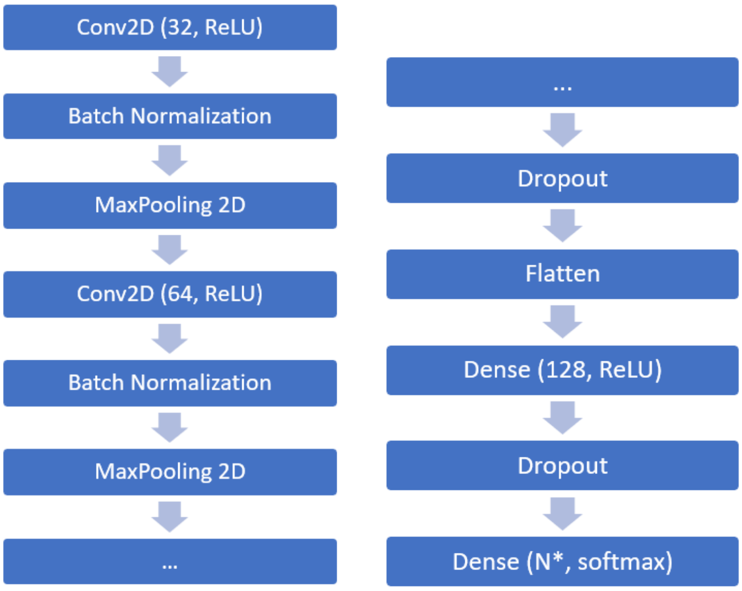


Figure 5.2: Classification process model
* N -– number of recognition classes

Matplotlib library is used to visualize the results of the experiments. The Comet.ml[1] platform is used to present the results of network training, to build error matrices (confusion matrices), as well as to save the statistics (the results obtained, the source code, the set of hyperparameters used, the graphs plotted, etc.) on a remote server.

5.3 Experiments and Results

The first part of our experiments considered the architecture of CNN across hyperparameter selection, regularization, and data augmentation. The second part was about bringing together various acoustic features fed to the input layer of the CNN model to improve accent recognition accuracy. All experiments were carried out for several classes of European accents.

5.3.1 CNN Model Tuning and Data Augmentation

The kernels of CNN convolution layers are convolution filters, where the cross- correlation operation takes place. The size of the kernel corresponds to the dimensions of the filter mask. The most common filter sizes for convolution layers in machine learning problems are (3, 3) and (5, 5). Better overall training results often stem from using smaller filters, which require less computational power and fewer backpropagation weights. However, it is important to note that no single value is suitable for all models: filter sizes need to be optimized based on the particular type of task.

Since neighboring pixels are highly correlated, pooling can be used to reduce the size of the output data. The farther the two pixels are from each other, the less correlated they are expected to be. Thus, a larger step in the pooling layer leads to more information loss. The standard pooling stride is (2, 2). Different filter size configurations are used for different types of input features. The basic model of the classifier uses the filters of size (3, 3) in convolution layers and (2, 2) in pooling layers. Following the recommendations from [85], for a set of 30 characteristics (described in 5.2.3, except for amplitude of mel-spectrograms), we used 2D filter configurations for convolutional layers.

[1] https://www.comet.ml/

Using linear amplitude mel-spectrograms as the input for classifying among {FR, IT, SP} accents, a number of filter configurations were tried. The length of the input feature matrices used to represent the input data is 100. The learning process was stopped as soon as the change in the recognition accuracy was less than 1% within 10 epochs. The highest recognition accuracy and a relatively short model training time were achieved when using the filters of size (3, 3), namely, 99.04%, both in convolution layers and in pooling layers, as Table 5.3 shows.

Table 5.3: Results of using different filter sizes with mel-spectrograms

French, Italian, Spanish (Romance Languages)				
Kernel size	**Pool size**	**Learning Time (mm:ss)**	**Accuracy**	**Error**
(3, 3)	(2, 2)	41:06	0.98	0.06
(3, 3)	(3, 3)	20:01	0.99	0.03
(5, 5)	(3, 3)	17:57	0.98	0.06
(7, 7)	(3, 3)	26:14	0.98	0.05

Thus, filters of size (3, 3) in hidden layers are the most universal and optimal for the CNN considered within the framework of ASR. Inclusion of additional features to MFCC improves the quality of recognition for many sets of languages, which confirms the hypothesis that improvements in ASR may result from combining MFCC with other types of available characteristics.

During the data augmentation phase, we tested the cases with a maximum horizontal shift of 5 and 10% for a subset of data, including the audio recordings for the foreign accent group {RU, SP, SW} as well as the audio files without a foreign accent (EN). MFCC was used as input data, as well as their alternate combinations with fundamental frequency and spectral centroid. The results are presented in Table 5.4.

Table 5.4: Classification results at different shift percentages for a set of languages of different language groups

English, Spanish, Swedish, Russian (Mixed group)				
	Maximum horizontal shift during augmentation			
	0.05		0.1	
Features	**Accuracy**	**Error**	**Accuracy**	**Error**
MFCC	0.65	0.94	0.65	0.86
MFCC + F0	0.68	0.84	0.72	0.78
MFCC + spectral centroid	0.65	0.93	0.66	0.86

Table 5.5: Classification results at different shift percentages for a set of Romance languages

French, Italian, Spanish (Romance Languages)		
Horizontal Shift	**Accuracy**	**Error**
0.05	0.75	0.62
0.1	0.75	0.59
0.15	0.74	0.62
0.2	0.77	0.55
0.25	0.75	0.58
0.3	0.76	0.58

Based on these results, we hypothesized that increasing the percentage of data shift may lead to higher recognition results. Given this assumption, we trained the classifier on the data set of accents {FR, IT, SP}, to which augmentation was applied. The result in Table 5.5 led us to conclude that the optimal accuracy/error value is reached when maximum percentage of horizontal shift during data augmentation is about 20%.

5.3.2 Input Acoustic Feature Sets

Acoustic feature sets fed into the CNN model were examined from three vantage points to obtain the feature set which yields the best recognition accuracy: input data dimensionality, possible MFCC combinations with other acoustic features and the impact of mel-spectrograms, which turned to be the most accent-dependent and thus the most eloquent input feature to improve classifier performance.

Dimension of input

While working with speech signals, it is necessary to consider the patterns of change in the characteristics describing these signals over time since speech is viewed as a time-dependent function. Thus, it is essential to consider the sequences of feature vectors rather than a single one-dimensional vector.

The division of the input features into larger or smaller chunks may introduce bias. Larger chunks can enable discovering longer speech patterns (more likely to be accent-dependent), but the training set becomes smaller, and training on high-dimensional data is naturally more computationally expensive (and, therefore, slower). Selecting shorter fragments allows using more input data, but can deteriorate the information captured about the accent from a fragment of feature vectors.

The experiments were performed using the sequences of vectors of mel-cepstral coefficients as input features. The training stops when the change in accuracy is at least 0.5% for an interval of 20 epochs or when 300 epochs were reached among five accents and 170 epochs in other cases. The probability of a neuron reaching zero when using the thinning method was 50%.

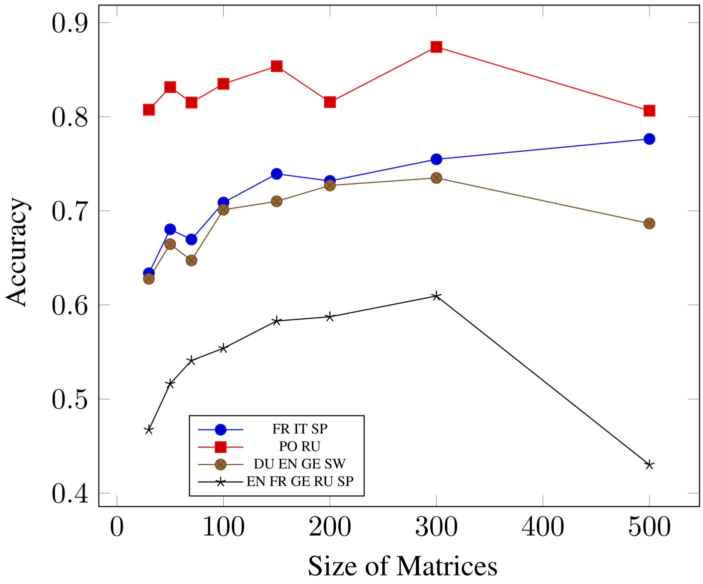


Figure 5.3: Accuracy variation when changing the size of the input matrices of features.

From our experiments, we can conclude that increasing the size of input data blocks to a certain value leads to an improvement in recognition accuracy, which can be seen in Figure 5.3-5.4. However, increasing the size of the matrices is inversely proportional to the number of input

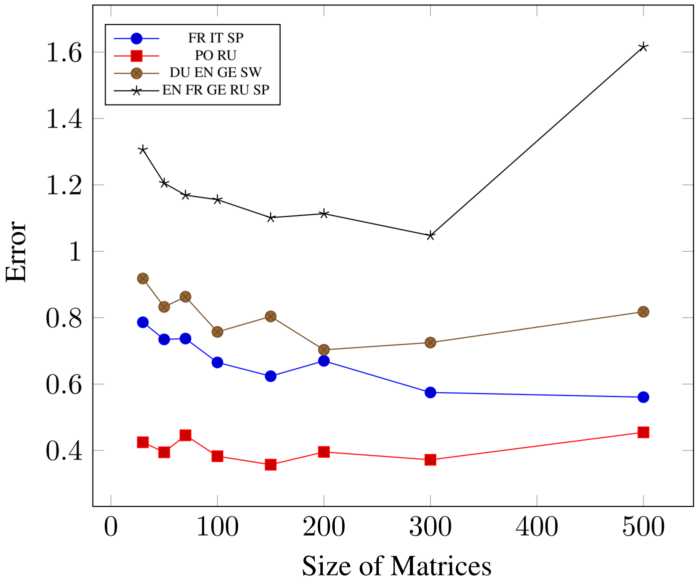


Figure 5.4: Error variation when changing the size of the input feature matrices.

instances, which leads to the inability of the model to fully capture accent-dependent patterns. For Romance languages, the recognition accuracy increased by increasing the number of input characteristics up to 300 per block. The noise, however, also increased during the training with the extended matrices. For training the classification of Slavic accents, the maximum accuracy was achieved with the length of input data blocks equal to 150 feature vectors, and 200 vectors for Germanic languages, and 300 vectors for the mixed language group.

Table 5.6 shows the results of an experiment performed on Romance accents using mel-spectrograms as input, varying the number of mel-bands used to represent the spectrograms.

During the experiments, we used a dropout of 0.25. The size of the filters in the convolution layers was (5, 5). The size in the pooling layers was (3, 3). Training was stopped when the recognition accuracy ceased to change by at least 1% for ten epochs.

As can be seen in Table 5.6, mel-spectrograms, consisting of 64 frequency bands, proved to be the most effective and were chosen as input characteristics for recognition. Although the use of 128-band mel-spectrograms can slightly increase the recognition accuracy, it substantially increases the training time. On the contrary, using mel spectrograms consisting of 32 mel-frequency bands is less computationally expensive but leads to a significant increase in error.

Based on the experimental results summarized in Table 5.6, we can conclude that the optimal size of the input feature matrices is 75 vectors when using amplitude mel-spectrograms on a linear scale.

MFCC combined with additional features

Now, let us consider the case of extending MFCC with a number of additional features, as suggested in [85]. MFCC speech characteristics are widely used in accent detection because they provide a compact but densely informational representation of an audio signal, resulting in high classification accuracy. In [85, 150], it was suggested that the accuracy can be further improved by adding additional information to the MFCC. However, adding an arbitrarily large number of input features would be detrimental, since excessive information would slow down the classifier's training process and increase the model overfit due to the noise. Therefore, it is important to select a limited number of suitable representative characteristics. Therefore, in

Table 5.6: Classification results for different sizes of input matrices for a set of Romance languages (mel-spectrograms)

French, Italian, Spanish (Romance Languages)				
Number of mel-bands	**Size of Input Matrices**	**Training Time (hh:mm:ss)**	**Accuracy**	**Error**
32	30	00:11:06	0.888	0.283
	50	00:13:10	0.924	0.201
	75	00:13:20	0.958	0.122
	100	00:14:53	0.948	0.154
	150	00:12:16	0.921	0.235
	200	00:15:03	0.867	0.319
64	30	00:28:55	0.974	0.086
	50	00:19:48	0.998	0.041
	75	00:19:16	0.991	0.033
	100	00:17:57	0.985	0.056
	150	00:25:15	0.988	0.039
	200	00:36:25	0.959	0.144
128	100	01:23:56	0.998	0.007
	150	01:31:14	0.983	0.087

this work, we strived to discover essential characteristics for the MFCC extension that would positively affect the classification accuracy while maintaining the basic filter sizes in the hidden layers of the classifier.

It should be noted that our selection of features does not contradict to Fisher criterion [159–161], although, in our work, we did not explicitly use it.

Training was stopped when the training accuracy of 90% or 120 epochs is reached for all accent sets except the case {EN, RU, SP, SW}, where the training process terminates as soon as 350 epochs are reached. The results obtained are shown in Table 5.7 and Table 5.8.

The obtained values of the testing error demonstrate that in half of the cases with filter sizes (3, 3) in convolutional layers and (2, 2) in the pooling layers, the accent-dependent patterns captured are worse compared to only using MFCC as input characteristics of audio signals.

In the case of the accent group {EN, GE, IT, PO}, adding the fundamental frequency to the mel-cepstral coefficients helps to increase the recognition accuracy by about 3%. For the set {EN, RU, SP, SW}, the most effective selection was to use all types of additional characteristics. The increase in classification accuracy is also about 3% compared to the usage of MFCC alone.

Intonation makes a significant contribution to the recognition of a foreign accent. Based on the fact that the F_0 contour in most experiments did not improve the classification results, we can conclude that a description of intonation is contained within MFCC. When extracting MFCC, information about F_0 is partially preserved due to the close distance between the low-frequency channels of the mel-filters [162].

Mel-spectograms

Linear scale mel-amplitude spectrograms extracted from the audio signals can also be tried as the inputs of the classifier. At the same time, according to the previously established optimal parameters of the classifier, the size filters (3, 3) are used in both the convolution and the pooling layers, while the size of the input feature matrices is 75 elements. The number of epochs is limited to 60, while the learning process was stopped when the change in recognition accuracy was less than 1% within ten epochs.

Table 5.7: Classification results using different types of input features for Slavic and Romance accents

Features	**Test Accuracy**	**Test Error**
Russian, Polish (Slavic Languages) Threshold Accuracy – 0.72		
MFCC	0.84	0.37
MFCC + F0	0.83	0.4
MFCC + spectral centroid	0.85	0.39
MFCC + spectral decay	0.84	0.4
MFCC + chromogram	0.79	0.44
MFCC + ZCR	0.84	0.38
MFCC + RMS	0.83	0.41
All	0.81	0.41
French, Italian, Spanish (Romance Languages) Threshold Accuracy – 0.43		
MFCC	0.75	0.6
MFCC + F0	0.69	0.71
MFCC + spectral centroid	0.67	0.73
MFCC + spectral decay	0.68	0.72
MFCC + chromogram	0.63	0.84
MFCC + ZCR	0.71	0.68
MFCC + RMS	0.7	0.7
All	0.66	0.8

Table 5.8: Classification results when using different types of input features for accents of mixed language groups

Features	**Test Accuracy**	**Test Error**
English, Italian, German, Polish (Mixed group) Threshold Accuracy – 0.29		
MFCC	0.62	1.00
MFCC + F0	0.65	0.88
MFCC + spectral centroid	0.61	0.94
MFCC + spectral decay	0.63	0.96
MFCC + chromogram	Threshold not passed	
MFCC + ZCR	0.64	0.9
MFCC + RMS	0.64	0.91
All	0.6	0.95
English, Spanish, Swedish, Russian (Mixed group) Threshold Accuracy – 0.33		
MFCC	0.72	0.81
MFCC + F0	0.71	0.83
MFCC + spectral centroid	0.68	0.88
MFCC + spectral decay	0.68	0.93
MFCC + chromogram	0.68	0.92
MFCC + ZCR	0.68	0.85
MFCC + RMS	0.67	0.95
All	0.75	0.7

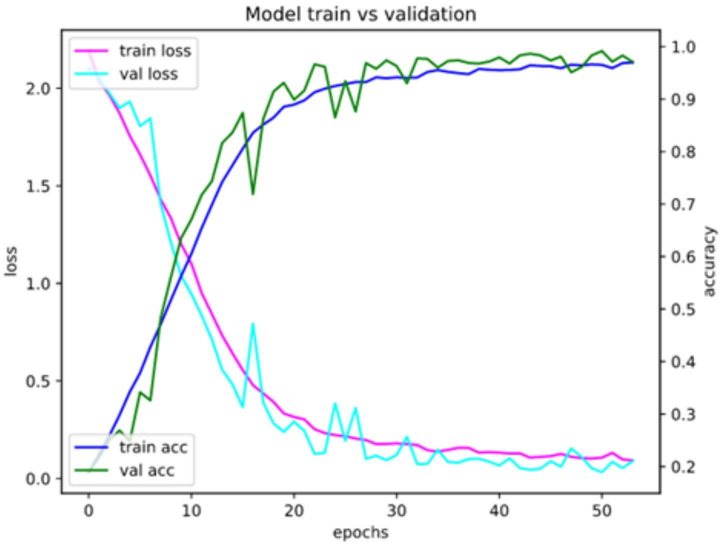


Figure 5.5: Accuracy and loss on training and test data during classifier training among the set of accents {DU, EN, FR, GE, IT, PO, RU, SP, SW}.

Table 5.9: Accuracy and loss for trained Accent Classification Models

Accents	Accuracy	Loss
PO RU	0.987	0.039
FR IT SP	0.986	0.052
DU EN GE SW	0.982	0.075
EN RU SP SW	0.988	0.042
EN GE IT PO	0.985	0.053
DU EN FR RU	0.984	0.039
EN FR GE RU SP	0.978	0.071
DU EN FR GE RU SP	0.964	0.097
DU EN FR GE IT PO RU SP SW	0.986	0.044
Average	0.982	0.056

We applied a dropout for different sets of accents to eliminate overfitting with values ranging between 10% and 25%. Regularization is applied intermittently either to the training or to the test set, to cope with model redundancy and inability to generalize the data.

Figure 5.5 shows accuracy and loss in the training model for the largest set of accents used in the experiments. Dark blue and green graphs show the accuracy of the training and test data, while the graphs in pink and light blue show the training and test data validation.

At the end of the training, the model achieves similar accuracy and loss values for the training and test data. For a smaller number of epochs compared to previous experiments, it was possible to achieve a much smaller error and greater accuracy, which means that using amplitude mel-spectrograms on a linear scale allows the model to place broad boundaries between classes. Accuracy and loss values achieved while testing the resulting models for classifying different sets of accents are presented in Table 5.9.

Table 5.10 presents the achieved results against the works reviewed in Section 5.1.

Amplitude mel-spectrograms on a linear scale showed high efficiency in recognizing foreign accents in English speech. But the results turned out to be slightly lower compared to [137]. This may be due to heterogeneous audio recordings contained in the Speech Accent Archive dataset.

Table 5.10: Comparison of existing solutions with the obtained results

Source	Classifier	Number of classes recognized	Precision	Accuracy of CNNs trained on mel-amplitude spectrograms on a linear scale
[137]	CNN (with attention mechanism)	2	1.0	0.987
		4	0.99	0.984
		9	0.995	0.986
[90]	CNN (AlexNet)	3	0.61	0.986
[88]	GMM	2	0.862	0.987
[89]	FFNN	6	0.914	0.964
[85]	CNN	3	0.703	0.986
		5	0.539	0.978
[140]	FF-MLP	3	0.99	0.986

In contrast to the homogeneous dataset in [137], in which all entries were made using the same equipment.

Compared to other solutions using the Speech Accent Archive dataset – [85,90] and with [88] where the dataset was used, based on text from the Speech Accent Archive, the implemented model achieved much better recognition accuracy by tuning hyperparameters, dimensionality of input features, and selecting amplitude mel-spectrograms on a linear scale as input features. The better recognition quality compared to [88] can be explained, among other things, by the fact that the authors of [88] removed silence fragments from audio recordings before extracting characteristics. During this research, we found that pauses in speech have a positive effect on the ability to determine accents.

5.4 Evaluation

The quality of the CNN-based classifier can be evaluated by creating a confusion matrix and connected standard information retrieval (IR) metrics that include overall precision, precision, recall, and F1.

The confusion matrix is a matrix C, where $C_{i,j}$ is equal to the number of observations that belong to class i and recognized as an object of class j. Such $C_{i,j}$, where $i = j$, is the number of observations in which the object class was recognized correctly. The size of the matrix C is $N \times N$, where N is the number of classes.

Figure 5.6 shows the error matrix for the largest set of accents used in the experiments.

Based on extracting the numbers of true positive TP, true negative TN, false positive FP and false negative FN cases from the confusion matrix, the overall accuracy is as follows:

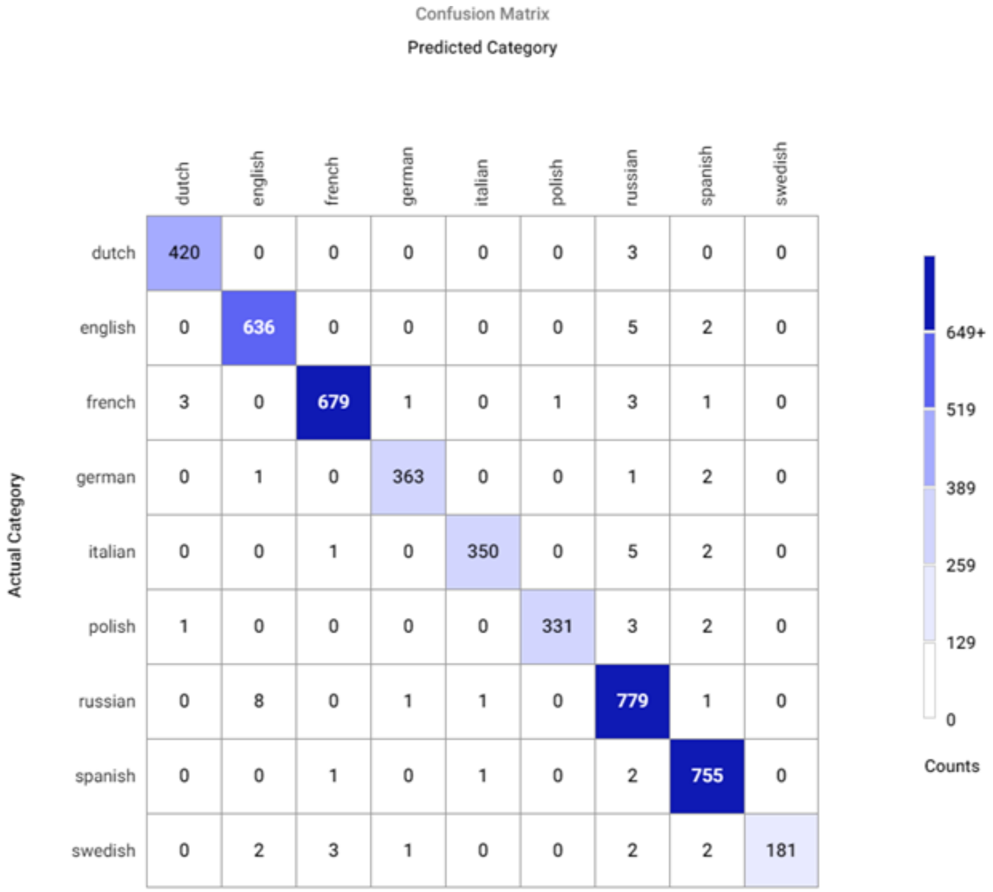
$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} = \frac{\sum_{i=0}^{N} C_{ii}}{\sum_{i=0}^{N}\sum_{j=0}^{N} C_{ij}} = 0.986 \tag{5.3}$$

The standard precision, recall and F1 metrics are as follows:

$$Precision = \frac{TP}{TP+FP} \tag{5.4}$$

$$Recall = \frac{TP}{TP+FN} \tag{5.5}$$

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{5.6}$$



	dutch	english	french	german	italian	polish	russian	spanish	swedish
dutch	420	0	0	0	0	0	3	0	0
english	0	636	0	0	0	0	5	2	0
french	3	0	679	1	0	1	3	1	0
german	0	1	0	363	0	0	1	2	0
italian	0	0	1	0	350	0	5	2	0
polish	1	0	0	0	0	331	3	2	0
russian	0	8	0	1	1	0	779	1	0
spanish	0	0	1	0	1	0	2	755	0
swedish	0	2	3	1	0	0	2	2	181

Figure 5.6: Classification error matrix among accents set {DU, EN, FR, GE, IT, PO, RU, SP, SW}.

Table 5.11 lists the computed IR metrics for each of the accent class, as well as the integral values across all the classes used in the experiments.

The average precision, recall, and F1 values for the considered accent classifier are 98.4%, 97.8%, and 98.1%, respectively. The resulting values show good classification quality for a classifier based on amplitude mel-spectrograms on a linear scale while distinguishing among the 9 classes.

5.5 Summary

Let us summarize the major results and findings of the current chapter.

1. Using additional audio signal information on time-frequency and energy features (such as spectrogram, chromogram, spectral centroid, spectral rolloff, and fundamental frequency) to the MFCC is proven to increase the accuracy of the accent classification compared to conventional feature set based on MFCC and raw spectrograms.
2. Amplitude mel-spectrograms on a linear scale (in contrast to logarithmic scale used in most studies) appear more powerful in accent classification task and make it possible to produce state-of-the-art accent recognition accuracy.
3. Based on our experiments, we demonstrated that the pauses in speech have a positive effect on the ability to determine accents. This is why they should not be eliminated from the input, at least with respect to the accent classification process.

Table 5.11: Average values of precision, recall and F1

Class	Precision	Recall	F1
DU	0.985	0.965	0.975
EN	0.984	0.983	0.983
FR	0.98	0.984	0.982
GE	0.978	0.98	0.98
IT	0.997	0.973	0.987
PO	0.993	0.977	0.987
RU	0.983	0.98	0.983
SP	0.968	0.992	0.978
SW	0.987	0.97	0.977
Average	0.984	0.978	0.981

4. The experiments conducted enhanced our understanding of how intonation may impact accent recognition. Based on the fact that the fundamental frequency contour in most experiments did not improve the classification results, we concluded that the intonation features are subsumed within MFCC. To our knowledge, considering the problem of accent recognition in connection with the analysis of prosody features of language makes an important and additional novel contribution.

The amplitude mel-spectrograms on a linear scale showed effectiveness in solving the problem of determining the speech accent in a foreign language using a CNN-based classifier even when applied to sparse speech data from a crowd-sourced dataset. We note that for the case of a crowd-sourced dataset, the accuracy is very close to the results of experiments with high-quality homogeneous data reported in many reviewed works, such as [135–137]. The better recognition quality compared to [88] can be explained by using the model that preserves silence fragments in the audio recording, which may correlate with the specificity of speech traits depending on the speaker's L1. Further studies may be helpful to expand the number of recognition classes, using an intermediate classifier to determine the L1 language group of the speaker before classifying a particular accent, and using a dataset with a variety of spoken content.

Chapter 6

Tailoring Feedback for *StudyIntonation* system

In this chapter, we look at ways to make *StudyIntonation*, a pronunciation training tool, better for learners. We focus on improving the way learners view the visualization of their pronunciation and hints on improvements. We also discuss how the system can be changed to better suit learners from different language backgrounds. This is covered in more detail in the upcoming sections on user interface enhancements (Section 6.1) and adding new features for language background adaptation by integrating automatic speech recognition and accent recognition components (Section 6.2). By updating the course editor and the mobile application, we are making important changes to help learners understand and practice the sounds of a new language more effectively. These updates are all about making the learning experience better: more straightforward, more tailored, and more in tune with what learners need. We believe that these improvements will make it easier for learners to improve their pronunciation.

6.1 Enhancing Visual Feedback in Pronunciation Training System

Unlike traditional 'listen and repeat' tools, the *StudyIntonation* application's use of visual feedback offers to use visual channel to learning intonation. This method is particularly effective because it engages multiple learning modalities, combining auditory and visual inputs to reinforce learning. It allows learners to understand and correct their intonation patterns more effectively, as they can literally 'see' their speech and compare it to the desired standard. In addition, visual feedback is instant and precise, which makes it easier for learners to identify and focus on specific areas that need improvement. This is especially beneficial for nuances in speech that are difficult to detect by ear alone. By integrating visual feedback, *StudyIntonation* helps bridge the gap between hearing and understanding the subtleties of intonation, fostering a deeper and more intuitive grasp of the language.

However, presenting a user with a pitch graph of their pronunciation compared to an example of native speech does not necessarily explain what needs to be improved to achieve the desired pronunciation. Some learners (especially those without a technical background) may find it difficult to understand the pitch graph notation and the numerical score produced by the system.

We proposed the following changes to the interface components aiming at improving and tailoring the CAPT feedback to language learners:

- Multiple attempts review;
- Pitch graph segmentation and segmented visualization;
- Portraying rhythm (including music notation for users with musical background).
- Demonstrating a context;

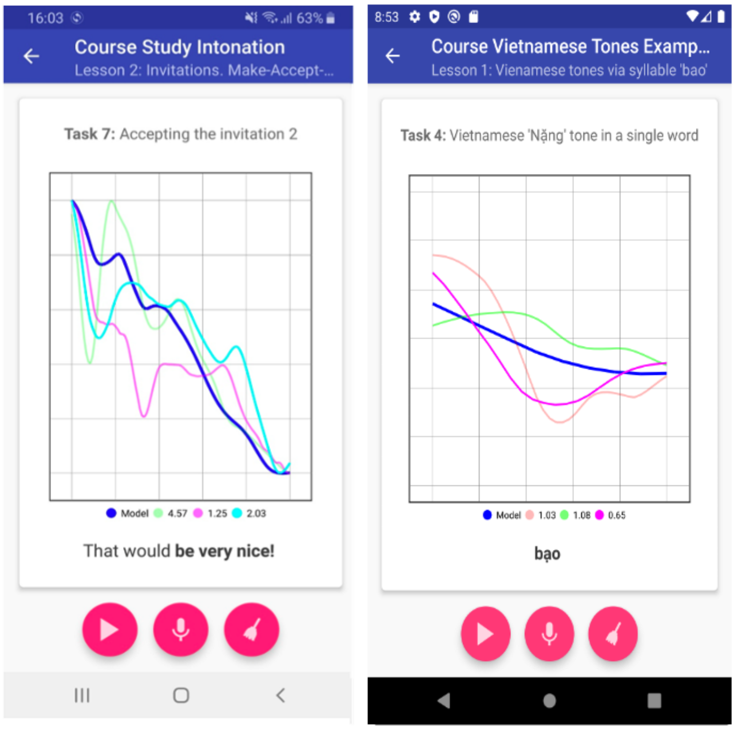
- Interactive scoring and gamification;

6.1.1 Pitch Graph Contours with Multiple Attempts Review

The *StudyIntonation* application presents the pitch of a native speaker using plotted pitch contours that demonstrate intonation patterns. For some learners, this graphical representation can be more intuitive and comprehensible than auditory feedback alone. It provides a visual benchmark against which they can measure their pronunciation efforts. The ability to practice and compare their speech patterns directly with the model facilitates a continuous improvement and adjustment process that is self-guided and reflective.

While testing the application and collecting feedback from learners, we noticed that each attempt to repeat the phrase brings slightly different curves, even though the user thinks that they repeated the phrase exactly the same. It underscores the importance of a personalized learning experience where learners can visualize and assess their multiple attempts overlaid against the reference pitch graph. Such side-by-side comparisons are instrumental in helping users discern how each attempt varies and what specific changes bring them closer to achieving the correct intonation.

Figures 6.1 (a) and (b) illustrate the visualization of multiple attempts of the learner together with the teacher's (reference) pronunciation for English and Vietnamese courses, respectively. The interface, designed with user-friendly icons at the bottom of the screen, represents the following actions: listen to the reference pronunciation, record the attempt, discard the most recent attempt, and preview the recorded attempts as seen in Figure 6.1 (b).



Course Study Intonation
Lesson 2: Invitations. Make-Accept-...

Task 7: Accepting the invitation 2

That would **be very nice!**

Course Vietnamese Tones Examp...
Lesson 1: Vienamese tones via syllable 'bao'

Task 4: Vietnamese 'Nặng' tone in a single word

bạo

(a) L2: English **(b) L2: Vietnamese**

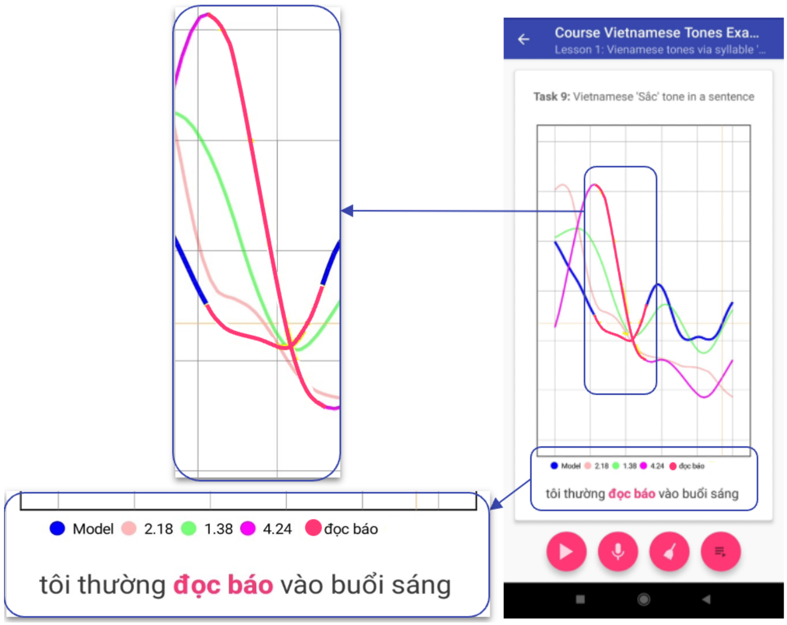
Figure 6.1: Multiple attempts pitch graphs and scores on the same screen.

For those aiming to improve the intonation of the studied language, instant feedback that shows the difference not only between the reference pronunciation but also between the previous attempts can make it clearer for the user to understand how to use the application and improve the effectiveness of the CAPT system.

6.1.2 Pitch Graph Segmentation and Segmented Visualization

Segmentation and highlighting of pitch parts corresponding to relatively independent segments (e.g., particular tones in syllables, stressed words within the longer phrase, elements with a higher degree of intonation variability, etc.) can be beneficial to allow users to focus on particular aspects. For tonal languages, such as Chinese and Vietnamese, the correct intonation is important at phrasal, lexical, and morphemic levels, since conveying the correct meaning is tightly connected to appropriate and accurate tone articulation. Even for non-tonal languages, such as English or Japanese, adequate modeling of tone movements within an utterance helps in teaching a more expressive and nuanced communication.

StudyIntonation application can be improved to analyze the learner's speech and highlight specific areas where their pitch deviates from the reference pronunciation. This precise feedback allows learners to focus on correcting specific intonation patterns, making their learning process more efficient. Figure 6.2 presents an exercise from the Vietnamese course [163] that highlights in red a part of the phrase for the reference pronunciation and the most recent attempt of the learner.



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

tôi thường **đọc báo** vào buổi sáng

Figure 6.2: Pitch graph segments: an example from a course in Vietnamese.

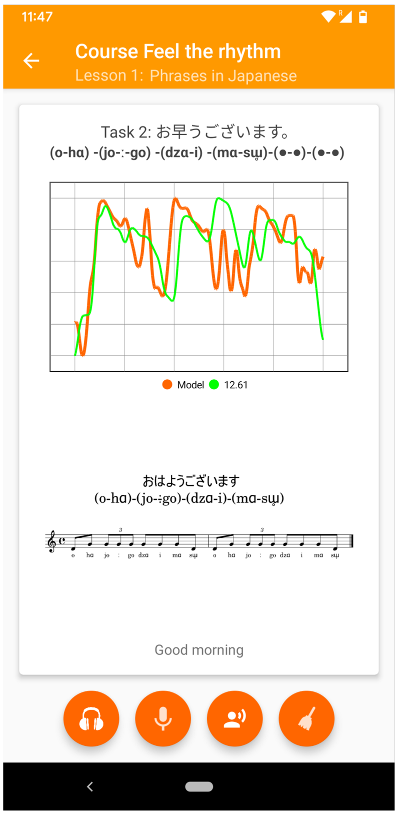
6.1.3 Portraying Rhythm

Rhythm serves as a fundamental aspect of speech and guides listeners through the nuances of language. Native speakers naturally use rhythmic cues to emphasize key points within their speech [1], which non-native speakers may inadvertently alter, leading to potential misunderstandings [164]. Recognizing this, we suggest improving the *StudyIntonation* system to provide learners with clear visual cues to master the rhythm of the target language, improving not only their pronunciation, but also overall communicative effectiveness.

For languages like Japanese, where the timing of syllables or moras is evenly distributed, understanding the rhythm is as crucial as the pronunciation of the words themselves. The concept of isochrony, or equal timing (as explained in Section 2.1.3), is a distinct feature of such languages. To help learners internalize this rhythmic pattern, *StudyIntonation* could introduce exercises that encourage the repetition of phrases with the correct timing and pauses. This method mirrors musical training, where repetition with a consistent rhythm solidifies skill acquisition.

However, conventional phonetic transcriptions, such as those used in the International Phonetic Alphabet (IPA), typically do not convey rhythmic or pitch information, which is critical for languages with nuanced tonal or rhythmic elements. To bridge this gap, we propose to use an extension of the IPA to include symbols that represent rhythmic patterns and pauses in the *StudyIntonation* system. This would allow learners to visualize the rhythm of speech in a structured and accessible manner, similar to reading a musical score.

Moreover, to further facilitate the understanding of rhythm, the system can incorporate elements of music notation to represent the speech melody. The use of time signatures can guide learners on the pace of speech, while the notation of high and low pitches can help with the acquisition of the correct tonal patterns. In particular, for languages such as Japanese, where pitch accent is significant, music notation can provide a clear representation of pitch movements within a phrase. Figure 6.3 shows an example of how all this can be implemented for the Japanese course, as described in [165].



Course Feel the rhythm
Lesson 1: Phrases in Japanese

Task 2: お早うございます。
(o-hɑ) -(jo-ː-go) -(dzɑ-i) -(mɑ-sɯ̥)-(●-●)-(●-●)

おはようございます
(o-hɑ)-(jo-ːgo)-(dzɑ-i)-(mɑ-sɯ̥)

Good morning

Figure 6.3: Music notation and extended IPA for L2 Japanese.

Integrating these musical elements into the system not only aids in the comprehension of speech rhythm and tone, but also caters to the multimodal learning preferences of users. Using visual representations of rhythm and pitch, the *StudyIntonation* system can enhance the learning experience, allowing learners to practice and perfect their pronunciation with a deeper understanding of the rhythmical and tonal aspects of the language. Such visual aids are particularly useful for learners who may have a musical background or for those who find visual learning tools more effective than auditory ones alone. The inclusion of these visual feedback mechanisms into the *StudyIntonation* system reflects an approach to language teaching that recognizes the complexity of speech and aims to make it more tangible to learners by providing them with extensive and subtle ways to improve their speech.

6.1.4 Providing a Context

In the area of language learning, context is not just a background detail — it is a critical element that shapes the way we understand and produce language. Pronunciation is no exception; it is deeply intertwined with the context in which words and phrases are used. Without a grasp of context, learners may find themselves capable of producing sounds accurately in isolation but struggling to apply those sounds appropriately in real-world communication. This disconnect can lead to misunderstandings or the inability to engage in fluent conversation, despite having good foundational knowledge of the language's phonetics.

The importance of context in pronunciation cannot be overstated. It guides the emotional tone, emphasis, and rhythm of speech, which are essential to convey meaning. As mentioned in [166], speakers often choose between tones (for example, between referring or promoting tones) depending on whether there is a known context or the referred information is completely new [167]. For example, the stress and intonation patterns in a question differ from those in a statement, and these nuances are often lost without the right contextual cues. A phrase said in surprise should sound different from the same phrase stated as a matter of fact. Contextual understanding helps learners not just say the words right, but also express them fittingly, aligning with the speaker's intent and the listener's expectations.

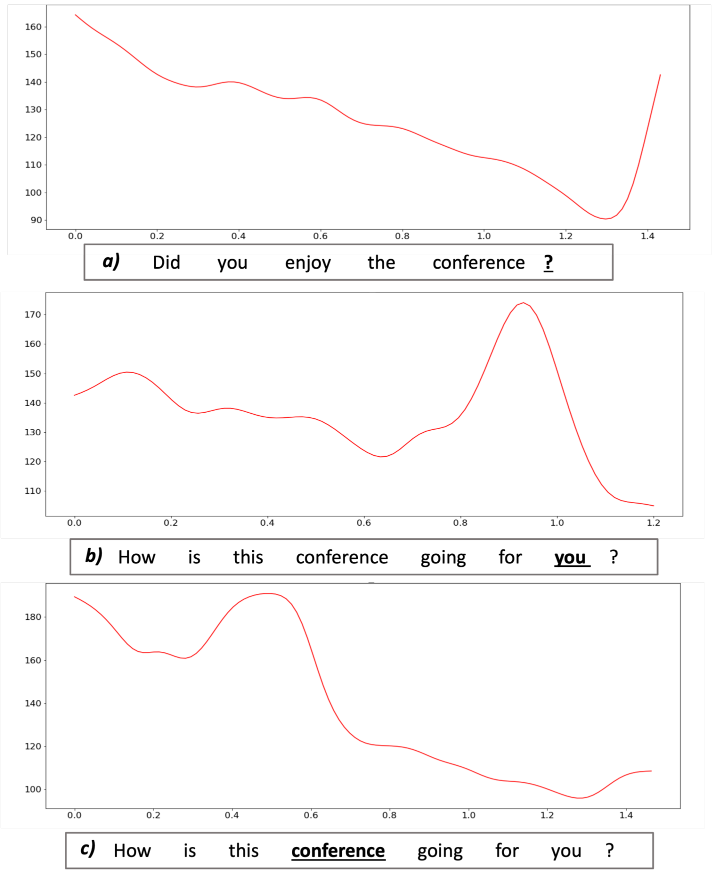


Figure 6.4: Example of the intonation in different contexts.

In Figure 6.4, modified from [166], we observe the variations in the intonation for the questions *'Did you enjoy the conference?'* and *'How is the conference going for you?'*. Scenario (a) shows the default intonation, where no particular word is being focused. It is a closed-ended question characterized by a primarily descending tone that concludes with a slight upward inflection. Scenario (b) is an 'open-ended' type of question that starts with 'Why?', 'How?', and 'What?'. It encourages a full answer, rather than a simple 'yes' or 'no' response that is usually given to a closed-ended question. It highlights the word 'you', likely following the speaker's sharing of their own experience. Scenario (c) underscores the 'conference', possibly implying a focus within a broader conversation topic. This selective emphasis, known as pitch prominence by phonologists, can be applied to any part of speech and is usually tied to the existing conversational context.

Including context-rich exercises in a pronunciation application does more than just improve the accuracy of sound production; it immerses learners in the culture and communicative essence of the language. It helps them understand the appropriate use of language in social situations. This understanding is pivotal for learners to become effective communicators, not merely proficient speakers. For language learners, especially those who may not have the opportunity to practice in a native environment, such contextualized learning is invaluable. It provides a simulated immersion experience that can greatly accelerate their proficiency and confidence in using the language in various scenarios.

One suggestion we have for enhancing the exercises in *StudyIntonation* is to include a question or sentence that provides context, such as asking a person you have just met at a conference. To illustrate this, we can use the example *'What is your area of expertise?'*. By introducing a context before the target phrase, which is highlighted in gray in Figure 6.5, the user can understand why a specific variation of the phrase requires a different intonation.

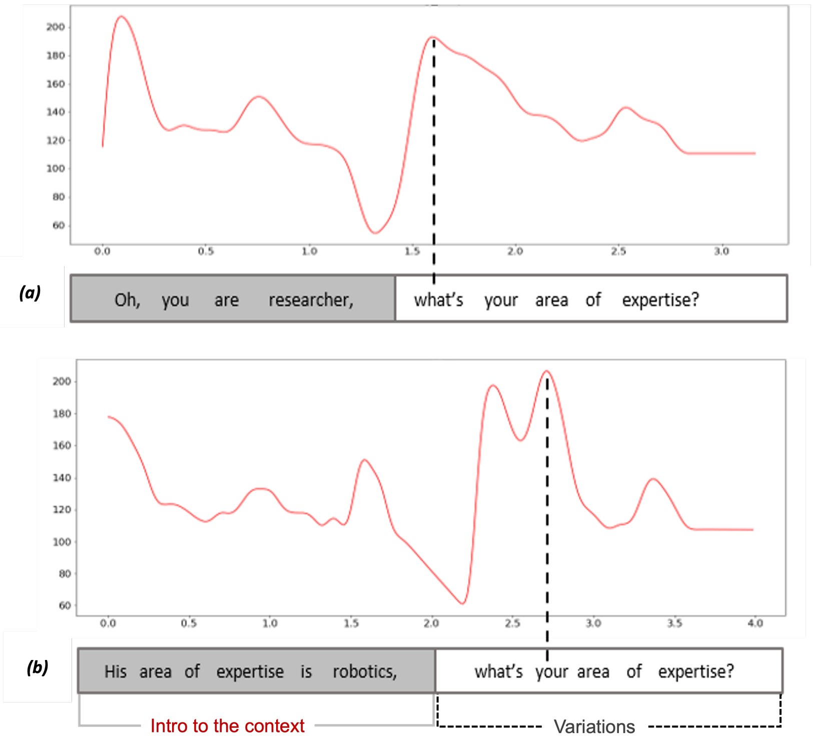
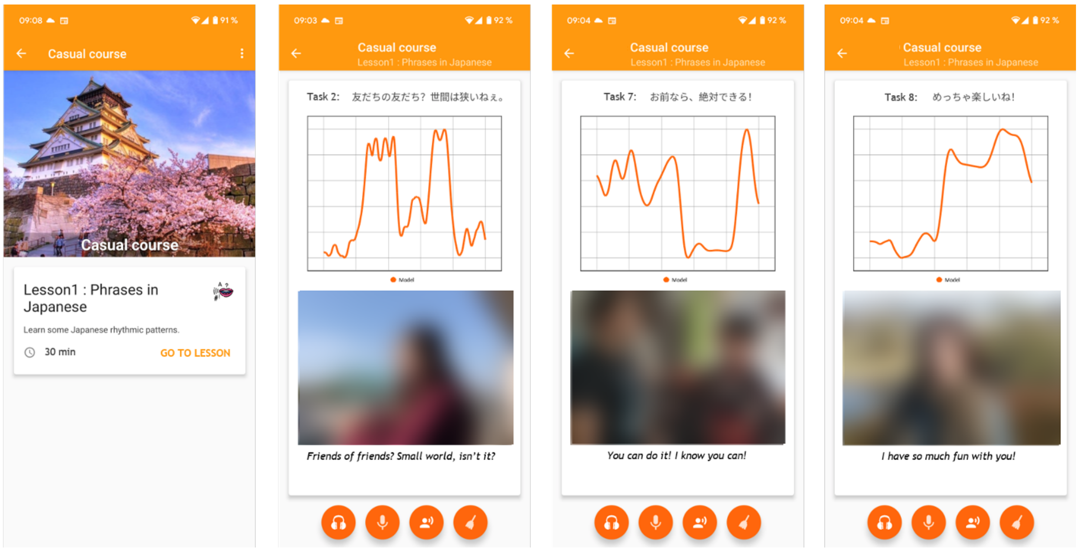


Figure 6.5: Example of the same phrase in different context.

Another way of providing context is to add a short video of the situation. Figure 6.6 shows examples of application screens from Japanese course with a demonstration of a short contextualized video. By providing phrases or short videos that describe the situational use of language, learners can practice pronunciation within the context of real conversational usage. This approach helps bridge the gap between isolated word pronunciation and the dynamic nature of spoken language, where the context dictates the tone, pace, and pitch of speech. Such visual and situational context can significantly enhance a learner's ability not only to repeat the sounds but also to grasp and reproduce the natural flow of the language, making language learning more holistic, intuitive, and ultimately more effective.



09:08 91 %

Casual course

Casual course

Lesson1 : Phrases in Japanese

Learn some Japanese rhythmic patterns.

30 min

GO TO LESSON

09:03 92 %

Casual course

Lesson1 : Phrases in Japanese

Task 2: 友だちの友だち？ 世間は狭いねぇ。

Friends of friends? Small world, isn't it?

09:04 92 %

Casual course

Lesson1 : Phrases in Japanese

Task 7: お前なら、絶対できる！

You can do it! I know you can!

09:04 92 %

Casual course

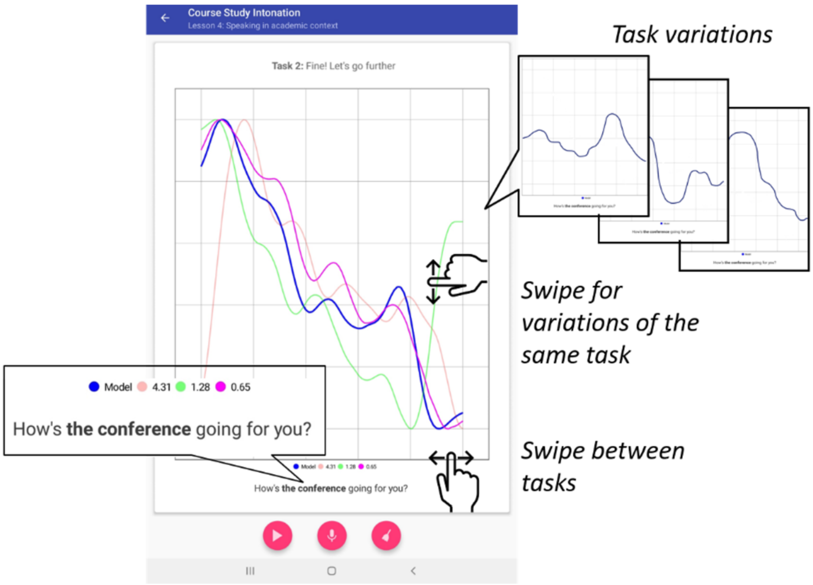
Lesson1 : Phrases in Japanese

Task 8: めっちゃ楽しいね！

I have so much fun with you!

Figure 6.6: Demonstrating a short contextualized video.

Modeling contrastive and attitudinal pronunciations within context is a powerful tool for teaching phrasal intonation. To focus attention on the distinctions between various intonations dependent on context, exercises featuring phrase variations can be grouped together. This allows learners to seamlessly navigate between different intonation options, enhancing their understanding of how context shapes pronunciation. We propose an interface enhancement that offers a sequence of related tasks, all linked to the same core exercise but each demonstrating a different potential intonation pattern. This could be accessed via intuitive swipe gestures, commonly found in modern mobile applications. Figure 6.7 conceptualizes how users could navigate through a series of intonation options with a simple swipe, thereby reinforcing the learning of context-sensitive speech patterns.



Course Study Intonation

Lesson 4: Speaking in academic context

Task 2: Fine! Let's go further

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

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

Task variations

Swipe for variations of the same task

Swipe between tasks

Figure 6.7: Tasks with the stacks of alternative intonation exercises.

6.1.5 Interactive Scoring and Gamification

Presenting pitch graphs and numerical scores based on DTW algorithm offers a detailed insight into a learner's pronunciation. However, this information may overwhelm or confuse some users, especially those less familiar with the technical aspects. To avoid this problem, we propose a more user-friendly approach to visual feedback that communicates the effectiveness of pronunciation attempts in a concise and understandable way.

This user-centric approach transforms abstract numerical scores into encouraging, personalized feedback messages. This transformation requires a thorough analysis to define the correlation between score ranges and the corresponding feedback messages. Educators and course creators play a central role in this process, determining the thresholds that differentiate between the feedback categories.

Consider, for example, a scenario in which a student received a numerical score lower than a predetermined threshold *X*. It results in an 'Excellent!' message, because for DTW a lower score is better. This positive reinforcement implies that the learner's pronunciation closely matches the reference, and they are ready to progress to the next exercise. On the contrary, if the DTW score exceeds the higher threshold *Y*, the learner is greeted with a 'Please try again' prompt, suggesting that further practice is necessary to improve. Scores that fall within the intermediate range [X, Y] would produce a 'Good!' message, leaving the decision to either retry or advance at the learner's discretion.

These tailored messages could be accompanied by more granular feedback. For example, highlighting specific segments of the pitch graph, where the learner's intonation diverged from the reference one, could offer actionable insights. It can be achieved using the approach described in Section 6.1.2.

Furthermore, integrating gamification into this interactive scoring system can significantly boost learner engagement and motivation. By incorporating game design elements such as points, levels, and badges, learners can track their progress in a visually compelling and gratifying way. A leaderboard could foster a friendly competitive environment, encouraging learners to improve their scores. For instance, achieving a streak of 'Excellent!' ratings might unlock a digital trophy, while consistent improvement over multiple attempts could be rewarded with virtual prize. These game-like features not only make learning more enjoyable, but also provide clear milestones and rewards that mirror the progression in a learner's pronunciation capabilities.

By applying interactive scoring with gamification, we can transform the potentially arduous process of pronunciation practice into an engaging and motivating journey, turning practice into play, and errors into opportunities for learning and growth. This approach, rooted in the principles of edutainment [168], ensures that learners are not just passive recipients of feedback but active participants in a dynamic learning experience that encourages their progress and fosters a positive and persistent approach to language mastery.

6.2 Personalizing Pronunciation Training for Learners with Different L1 Background

We propose to integrate advances in ASR and accent recognition technology to create a more tailored and effective system by addressing the lower accuracy of traditional ASR models for accented speech and providing personalized exercises based on the L1 background. Specifically, we outline the incorporation of an accent recognition model into the *StudyIntonation* mobile application, allowing to identify learners' first language (L1) backgrounds. By doing so, we enable course content creators to design linguistically context-aware exercises and employ ASR technology to enhance speech detection accuracy and accelerate transcription generation during the content creation phase. Furthermore, we propose using neural style transfer techniques for accent neutralization to adapt learners' accents before comparing them to reference pronunciations. The proposed updates to the *StudyIntonation* system, in the form of ASR-based and accent-targeted solutions, are shown in Figure 6.8. This integrated approach offers learners a more precise and personalized learning experience, optimizing pronunciation training.

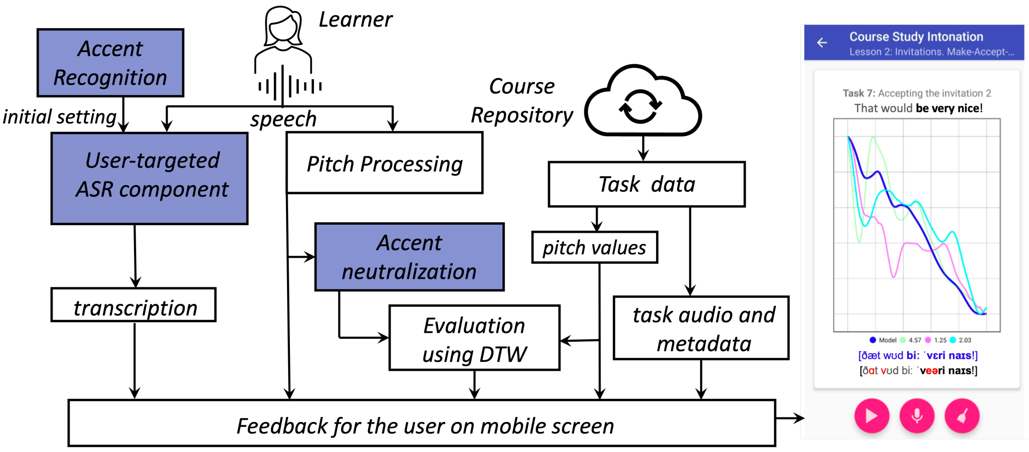


Figure 6.8: Integration of user-targeted ASR and accent solutions.

Non-native speakers often exhibit unique pronunciation traits influenced by their L1, that can be better captured by user-targeted ASR models. 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. ASR technology can be deployed directly on the user's mobile device, generating transcriptions of the user's spoken utterances. These transcriptions are then compared to the reference pronunciation. Phonemes that differ from the reference are highlighted in red, which offers a clear visual indication to the learner of areas where improvement is needed. Figure 6.8 illustrates the visual feedback of the pitch graphs and the evaluation score extended with the reference and the actual phonetic transcription of user speech.

The inclusion of an accent neutralization model (using neural style transfer [103]) modifies the learner's accent to facilitate a more accurate comparison with reference pronunciation [104] using the DTW algorithm [48].

6.2.1 Integrating Accent Recognition

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.

Building on the foundations laid in Chapter 5, we propose an accent recognition module that plays an important role in personalizing feedback. During the initial setup of the application, the learner is offered to read a phrase from the Speech Accent Archive [149] (as explained in Section 5.2.2), so the application can discern the user's L1 background. The evaluation results are 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 (more about it in Section 6.2.2 below). 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. Implicit accent recognition may allow CAPT systems to provide feedback and practice activities that are both discrete and targeted, avoiding the explicit acknowledgement by the user which could lead to privacy, identity, and self-esteem concerns due to stereotypes and prejudices associated with accents.

This approach to incorporating accent recognition into the initial setup process not only enhances the personalization of the CAPT system, but also 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.

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. We expand the capabilities of *StudyIntonation* by introducing several key features to empower content creators in developing more personalized and effective language learning courses. This teaching approach aims to tackle the unique pronunciation barriers that each learner may face due to their L1 influence, paving the way for more effective learning outcomes. We update the structure of the pronunciation task metadata for the Course Editor module (detailed in Chapter 3, Section 3.2.1), by allowing the tasks to be tagged for specific L1 backgrounds, as shown in Figure 6.9. If such case, the *StudyIntonation* system will select the next task based on the learner's background.

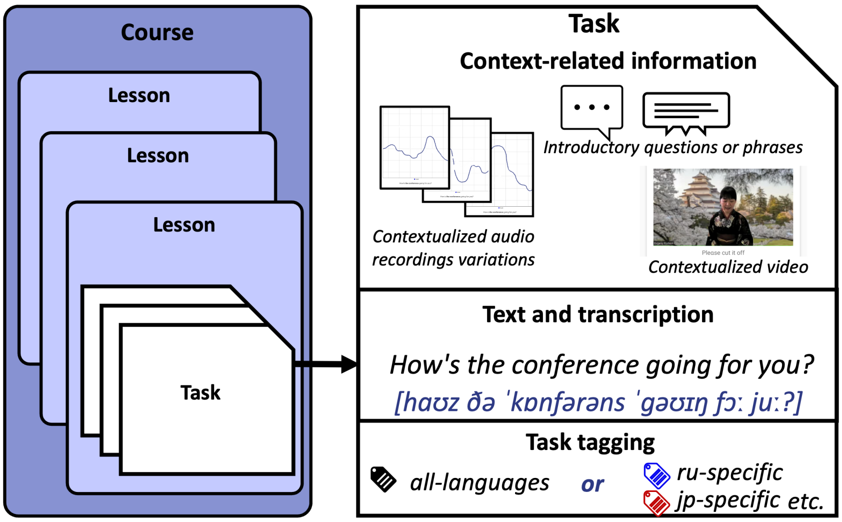

Course
Lesson
Lesson
Lesson
Task
Task
Context-related information
Introductory questions or phrases
Contextualized audio recordings variations
Please cut it off
Contextualized video
Text and transcription
How's the conference going for you?
[haʊz ðə ˈkɒnfərəns ˈɡəʊɪŋ fɔː juː?]
Task tagging
all-languages
or
ru-specific
jp-specific etc.

Figure 6.9: Expanding metadata for pronunciation tasks.

6.2.2 Tailoring Automatic Speech Recognition

Prosody teaching systems, by definition, focus on suprasegmental features, such as intonation and rhythm patterns, and ignore segmental features, such as the pronunciation of individual phonemes. One of the challenges in teaching suprasegmental pronunciation is to ensure that the learner not only repeats the intonation correctly but also pronounces the phonemes appropriately. Some CAPT environments integrate ASR models to capture segmental features to "understand" the words and 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 [169]. 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 characteristics of different L1 backgrounds, we ensure more accurate and personalized feedback for our users.

Data

We used the L2-ARCTIC [100] dataset for L2 fine-tuning. This corpus is focused on non-native English and contains English speech recordings from 24 non-native speakers of six different L1 backgrounds: Arabic, Hindi, Korean, Mandarin, Spanish, and Vietnamese. Each of the L1s consists of two female speakers and two male speakers. Each speaker in the datasest has contributed approximately one hour of phonetically-balanced read speech. L2-ARCTIC is based on the original L1 English corpus, CMU ARCTIC [170] and consists of 1,132 carefully selected sentences from Project Gutenberg [170]. Since we have a number of students from Vietnam and China studying English pronunciation using the *StudyIntonation* system at the University of Aizu, for preparing user-targeted ASR models we selected the subset of recordings made by speakers with Vietnamese and Chinese (Mandarin) L1 background.

Model

We applied transfer learning techniques to a multilingual Wav2Vec2 model [171] for Vietnamese and Mandarin L1 backgrounds. We selected the Wav2Vec2-Base-960h model trained on 960 hours of speech data from LibriSpeech [118]. It has fewer parameters compared to the wav2vec 2.0 Large (LV-60) model (which was additionally self-trained on 53.2k hours of Libri-Light [172]) or a newer cross-lingual model XLS-R [173]. We prioritize the ability to run the model on mobile devices smoothly to slightly better accuracy achieved by using models with more parameters.

Self-supervised learning enables Wav2Vec2.0 model to learn from audio data where transcriptions (labels) are not available. This approach involves training the model to predict parts of the audio it has not heard based on the parts it has, thereby learning meaningful representations of speech. In addition, the model can learn rich representations from raw audio. Training in an extensive and varied set of languages enhances its ability to generalize in various linguistic contexts. It learns contextualized speech representations by randomly masking feature vectors before passing them to a transformer network during self-supervised pre-training. The model is trained to predict the correct speech unit for masked audio parts while also learning what the speech units should be. This allows us to capture nuanced variations in pronunciation in different accents.

Training the Model

To tailor the model to specific accents, we fine-tune it using Connectionist Temporal Classification (CTC) [174] on speech recordings and corresponding transcriptions from the L2-ARCTIC dataset. We resampled the audio from 44.1 kHz to the required sampling rate of 16 kHz by the Wav2Vec model. We use PyTorch [175], an open-source machine learning framework, in tandem with Hugging Face's Transformers library [176], 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. The HuggingFace's Transformers library allows us to efficiently execute our language models on mobile devices [177], thus ensuring the wide accessibility and seamless operation of our system for users anywhere and anytime.

For the training stage, we implement a data collator that dynamically pads training batches to the longest sample in the batch, and use a WER metric. WER is a common metric used in ASR to measure the performance of a model. It is based on the Levenshtein distance, but at the word level instead of the phoneme level. It represents the percentage of errors in the transcribed text compared to the reference text. A lower WER indicates better performance of the speech recognition model.

We load a pre-trained checkpoint of Wav2Vec2-Base-960h 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 training configuration and hyperparameter tuning, we follow the recommendations from the Wav2Vec2 paper [171] by employing the following learning rate schedule:

- *warmup2constant*: This stage involves a warm-up phase transitioning to a constant learning rate. During warm-up, the learning rate gradually increases, helping the model to start learning effectively without drastic changes that might harm the learning process. Then it shifts to a constant rate for stability.
- *constant*: This stage uses a constant learning rate, providing a stable environment for the model to learn.
- *decay1* and *decay2*: These stages involve a decay in learning rate, decreasing it over time. It allows for finer adjustments as the model's training progresses. Two decay stages use different settings.
- *constant_after_decay*: This stage returns to a constant learning rate after a period of decay. It is a strategy to fine-tune the model after the initial major learning phases have been completed.

Results

Table 6.1 presents the performance evaluation in terms of WER of a speech recognition model for Vietnamese and Chinese L1 speakers. The evaluation is shown for two different model conditions – 'baseline' and 'fine-tuned' – and across two sets of data, 'dev' (development, also known as validation) and 'test'. The baseline row represents the performance of the original unmodified model. It is the starting point against which improvements are measured. The fine-tuned row shows the performance after the model has been fine-tuned. Fine-tuning involves adjusting a pre-trained model to better fit specific data or characteristics, in our case, English speech from Vietnamese and Mandarin L1 speakers. 'Dev' column shows the model's performance on a development (validation) dataset. 'Test' column represents the model's performance on a test dataset, which is used for the final evaluation and not part of the training process. For the baseline model, the WER values are 0.3530 (dev) and 0.3486 (test) for Vietnamese L1 and 0.2838 (dev) and 0.2795 (test) for Mandarin L1, indicating the performance of the model before any fine-tuning. For the fine-tuned model, the values improve significantly to 0.2152 (-39.04%) on 'dev' and 0.2164 (-37.92%) on 'test' for Vietnamese L1 and 0.1598 (-43.69%) on 'dev' and 0.1687 (-39.64%) on 'test' for Mandarin L1, showing that fine-tuning has positively impacted the model's ability to recognize the Vietnamese and Chinese speaker's English speech.

Table 6.1: WER Evaluation for Vietnamese and Mandarin L1

	Vietnamese L1		**Mandarin L1**	
	dev	**test**	**dev**	**test**
baseline	0.3530	0.3486	0.2838	0.2795
fine-tuned	0.2152	0.2164	0.1598	0.1687
change (in %)	-39.04%	-37.92%	-43.69%	-39.64%

Training ASR models on datasets of non-native speakers moves away from the generic approach of treating all language learners the same, regardless of their native language. This shift toward personalized pronunciation training is evident in the focus on specific language families or mother tongues. However, the use of the L2-Arctic dataset, which consists of phrases from out-of-copyright texts, might reduce the performance of the ASR model for everyday conversations. We are working on incorporating spoken language datasets like ICNALE [178] for improvement. When considering accent recognition, it is crucial to acknowledge the complex linguistic environment that learners navigate, often influenced by multiple factors, including their first language, which shapes their ability to learn a new language.

6.3 Summary

StudyIntonation, initially oriented toward learning English pronunciation, 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).

Enhancing the interactive features in *StudyIntonation* system can lead to a more user-friendly learning environment, better suited to the individual pace and proficiency of learners. This adaptation equips teachers and course creators with tools to facilitate more effective and learner-centric language acquisition. But, most of all, it benefits learners by providing a tailored learning experience. The proposed enhancements include:

- Multiple attempts review allows learners to visually compare several attempts of their pronunciation side-by-side with the reference to track progression and make incremental adjustments.

- Pitch graph segmentation and segmented visualization spotlight individual tones or stress patterns within a phrase, aiding learners in concentrating on the modification of specific intonation segments.

- Rhythmic representation illustrates rhythm and pitch variation in language, and music notation, as an additional learning tool, is beneficial for those with a musical background.

- Demonstrating context that involves the use of contextual elements, such as phrases or short videos, to create a situational setting for pronunciation exercises improves comprehension and conceptualization.

- The variation in intonation patterns offers a selection of exercises within different contexts, allowing learners to explore different variations of the same phrase.

- Replacing abstract numerical scores with interactive categorical feedback (e.g., 'Excellent!', 'Good!', 'Try again') can help learners to understand performance evaluations more intuitively.

- Gamification may transform the monotonous routine of practice into an exciting language learning adventure. By introducing elements such as points, levels, badges, and challenges, learners could be motivated to engage more deeply and persistently with their pronunciation exercises.

These proposed updates aim to make the learning process not only more interactive and engaging but also more adaptive to the individual learning journey, catering to the unique needs and preferences of each learner.

In an attempt to offer tailored feedback to learners from diverse L1 backgrounds, we propose integrating accent recognition and user-targeted ASR models into *StudyIntonation* system. These additions are designed to significantly enhance the personalization of the CAPT platform, ensuring a more targeted and effective learning experience for each individual. The idea is to "teach" the system to identify a specific learner's first language (L1) background, thus creating grounds for personalization of pronunciation exercises. Knowledge of the user's L1 background and the ability to include L1-tailored exercises into the course empowers pronunciation course content creators to deliver more personalized teaching material. It allows the CAPT system to suggest targeted exercises and provide custom feedback based on common pronunciation challenges associated with specific L1 accents and additional content to practice the corresponding tasks.

In addition, we are improving the content creation process by applying state-of-the-art ASR models. Such automated generation of transcriptions for recorded tasks significantly reduces the manual workload of content creators. In case of inaccuracies in the automatically generated transcriptions, the creators have the flexibility to manually review, edit, and save the corrected version, ensuring the quality of the learning materials.

Furthermore, accent recognition is a prerequisite for determining the most suitable ASR model based on the user's L1 background. Fine-tuned ASR models for such backgrounds lead to a significant improvement in phoneme detection accuracy compared to models trained on native speech datasets. The approach we employ helps us refine and optimize the learning process, making it more attuned to the individual traits of different spoken accents, ultimately enhancing its precision and usability for our diverse range of learners.

There are additional improvements that we are considering or working on. Although they do not belong to either visual enhancements or personalization, they are still important for the overall user experience. Such improvements include controlling the speed of the playback and having the same phrase recorded by different speakers. For the first point, users may find useful an option to adjust the speed of playback. Slowing down the audio can make it easier to dissect and understand the intricate phonetic components of the language. As for the second one, having phrases recorded by several speakers (possibly of different gender and/or age) can allow user to select the variation that better matches their liking.

We are currently developing a dynamic adaptation module for the *StudyIntonation* system that provides learners with personalized tasks based on their individual performance. This performance is quantified through the use of the cross-recurrence quantification analysis (as detailed in Chapter 4) and DTW algorithm to provide a tempo-invariant evaluation of the learner's performance. In the cases where a significant discrepancy is detected between the student's pronunciation and the reference standard (indicated by a high DTW score), the system may "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 dynamic assessment. 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.

Chapter 7

Conclusion

This dissertation presents possible solutions to improve Computer-Assisted Pronunciation Training (CAPT) systems, focusing on suprasegmental training, dynamic assessment, and personalized tools and interfaces for learners with different L1 background. Research integrates available algorithms and models of digital signal processing, machine learning, accent recognition, and automatic speech recognition (ASR) within an intelligent CAPT system, but also studies how sociocultural language pedagogy theories can be applied to develop technology-enhanced innovative methodologies in language learning. Specifically, in this dissertation, we address the *StudyIntonation* system, an iCAPT environment designed to improve prosody skills of language learners, including rhythm, stress, and intonation, as a testbed for the proposed solutions.

The *StudyIntonation* application diverges from the conventional 'listen and repeat' pedagogy by integrating visual cues to facilitate intonation learning. This method is effective for engaging multiple learning modalities, combining auditory and visual inputs to reinforce learning. It allows learners to understand and correct their intonation patterns more effectively, as they can visually compare their speech to the desired standard. Visual feedback, with its immediate and explicit nature, enables learners to discern and amend their intonation contours more effectively. It is particularly beneficial for those particularities that could be elusive to the ear alone.

However, presenting a user with a pitch graph of their pronunciation compared to an example of native speech does not necessarily explain what needs to be improved to achieve the desired pronunciation. Some learners, especially those without a technical background, might struggle to decipher pitch graph notation and numerical scores. In particular, the following interface components can support better tailored CAPT feedback to language learners:

- Reviewing multiple attempts allows learners to visually compare several attempts of their pronunciation side-by-side with the reference to track progression and make incremental adjustments [179].

- Segmented pitch graph and segmented visualization spotlight individual tones or stress patterns within a phrase, helping students concentrate on the modification of specific intonation segments [180].

- Visual modeling of language rhythm along with pitch variations presented using a simplified music notation that can be beneficial for learners with a musical background [181].

- Context demonstration by means of contextual elements, such as phrases or short videos that create situational settings for pronunciation exercises, thus contributing to the comprehension and conceptualization of the exercises of the learners [165].

- Intonation pattern variations that can be presented across the context, thus, enabling learners to train variant intonations of the same phrase [179].

Advances in accent recognition and ASR technology can contribute to creating a more personalized and effective system based on the L1 background of the learner. In [182], we detail how to implement accent recognition using deep learning techniques and demonstrate that a Convolutional Neural Network (CNN) model, trained on a diverse set of accents, is effective in classifying a range of accents. The model was applied to the sparse data from the Speech Accent Archive, which is a crowd-sourced collection of speech recordings. The sparsity of this dataset makes the implementation showcasing with respect to the practical implications in the real world. Although the techniques and features used in our work are known in the speech processing domain, extensive experiments involving their combination and the selection of optimal parameters for CNN filters have not been reported so far in their application to the specific problem of accent recognition, especially in context of the possible application to iCAPT systems.

Tailored ASR models can improve the recognition of accented speech and provide the learner with textual feedback on the pronounced phrase as we described in [183]. 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. To achieve efficient processing and robust understanding of diverse accents, we applied transfer learning techniques to an advanced multilingual model that can be used on mobile devices. We explain how to fine-tune the model using a dataset of accented English speech and show the improvements compared to the general ASR models. We describe how to integrate these models into the mobile application and the course editor module of the *StudyIntonation* system.

Enhancing the interactive features of a CAPT system, we adapt the learning environment in a way that makes 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.

It is worth mentioning that the *StudyIntonation* system mainly aims 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. 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.

References

[1] N. Bogach, E. Boitsova, S. Chernonog, A. Lamtev, M. Lesnichaya, I. Lezhenin, A. Novopashenny, R. Svechnikov, D. Tsikach, K. Vasiliev *et al.*, “Speech processing for language learning: A practical approach to computer-assisted pronunciation teaching,” *Electronics*, vol. 10, no. 3, p. 235, 2021.

[2] V. M. Research. Online language learning market size by type (individual learners and institutional learners), by language (english, spanish, chinese, french, german, japanese, and others), by geographic scope and forecast. [Online]. Available: https://www.verifiedmarketresearch.com/product/global-online-language-learning-market-size-and-forecast/

[3] J. B. Gilbert, *Teaching pronunciation: Using the prosody pyramid.* Cambridge University Press, 2008.

[4] M. J. Munro and T. M. Derwing, “Foreign accent, comprehensibility, and intelligibility in the speech of second language learners,” *Language learning*, vol. 45, no. 1, pp. 73–97, 1995.

[5] V. A. Murphy, *Second language learning in the early school years: Trends and contexts*. Oxford University Press, 2014.

[6] D. Liu and M. Reed, “Exploring the complexity of the l2 intonation system: An acoustic and eye-tracking study,” *Frontiers in Communication*, vol. 6, p. 51, 2021.

[7] D. Velázquez-López and G. Lord, *5 Things to Know About Teaching Pronunciation with Technology*. CALICO Infobytes, 2021. [Online]. Available: http://calico.org/infobytes

[8] P. Boula de Mareüil and B. Vieru, “The contribution of prosody to the perception of foreign accent,” *Phonetica*, vol. 63, pp. 247–267, 02 2006.

[9] C. Huang, T. Chen, and E. Chang, “Accent issues in large vocabulary continuous speech recognition,” *International Journal of Speech Technology*, vol. 7, pp. 141–153, 04 2004.

[10] I. R. MacKay, J. E. Flege, T. Piske, and C. Schirru, “Category restructuring during second-language speech acquisition,” *The Journal of the Acoustical Society of America*, vol. 110, no. 1, pp. 516–528, 2001.

[11] S. Watanabe, “How to teach the Japanese pitch pattern visually,” *Akita International University Global Review*, vol. 7, pp. 47–58, 2015.

[12] P. Roach, *English phonetics and phonology: A practical course*. Cambridge: Cambridge university press, 2009.

[13] A. Fox, *Prosodic features and prosodic structure: The phonology of suprasegmentals*. Oxford University Press, 2000.

[14] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper, "Enriching speech recognition with automatic detection of sentence boundaries and disfluencies," *IEEE Transactions on audio, speech, and language processing*, vol. 14, no. 5, pp. 1526–1540, 2006.

[15] I. Lehiste, *Suprasegmentals*. The MIT Press, 1970.

[16] M. Ploquin, "Prosodic transfer: From Chinese lexical tone to English pitch accent." *Advances in Language and Literary Studies*, vol. 4, no. 1, pp. 68–77, 2013.

[17] K. M. Yu, "The experimental state of mind in elicitation: illustrations from tonal fieldwork," *Language Documentation & Conservation*, vol. 8, p. 738–777, 2014.

[18] D. Hirst and C. de Looze, "Fundamental frequency and pitch," in *The Cambridge handbook of phonetics*, R.-A. Knight and J. Setter, Eds. Cambridge University Press, 2021, p. 336–361.

[19] K. L. Pike, *The Intonation of American English*. Ann Arbor: University of Michigan Press, 1945, vol. 1.

[20] D. Abercrombie, *Elements of General Phonetics*. Edinburgh: Edinburgh University Press, 1967.

[21] H. Kim and J. Cole, "The stress foot as a unit of planned timing: Evidence from shortening in the prosodic phrase," in *Proceedings of Interspeech*, 2005, pp. 2365–2368.

[22] A. Arvaniti, "Rhythm, timing and the timing of rhythm," *Phonetica*, vol. 66, no. 1-2, pp. 46–63, 2009.

[23] H. Kubozono, "The mora and syllable structure in japanese: Evidence from speech errors," *Language and Speech*, vol. 32, no. 3, pp. 249–278, 1989.

[24] Y. Kureta, T. Fushimi, and I. F. Tatsumi, "The functional unit in phonological encoding: evidence for moraic representation in native japanese speakers." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 32, no. 5, p. 1102, 2006.

[25] M. Levy, *Computer-assisted language learning: Context and conceptualization*. Oxford University Press, 1997.

[26] A. Neri, O. Mich, M. Gerosa, and D. Giuliani, "The effectiveness of computer assisted pronunciation training for foreign language learning by children," *Computer Assisted Language Learning*, vol. 21, no. 5, pp. 393–408, 2008.

[27] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot. Int.*, vol. 5, no. 9, pp. 341–345, 2001.

[28] D. Knuth. History of rosetta stone. [Online]. Available: https://www.rosettastone.eu/history

[29] C. S. Company. History of rosetta stone. [Online]. Available: http://www.carnegiespeech.com/products/nativeaccent.php

[30] M. C. Pennington and P. Rogerson-Revell, "English pronunciation teaching and research," *Londres: Palgrave Macmillan*, vol. 10, pp. 978–988, 2019.

[31] B. Lobanov, V. Zhitko, and V. Zahariev, "A prototype of the software system for study, training and analysis of speech intonation," in *Speech and Computer: 20th International Conference, SPECOM 2018, Leipzig, Germany, September 18–22, 2018, Proceedings 20*. Springer, 2018, pp. 337–346.

[32] D. Sztahó, G. Kiss, and K. Vicsi, "Computer based speech prosody teaching system," *Computer Speech & Language*, vol. 50, pp. 126–140, 2018.

[33] J. Kommissarchik and E. Komissarchik, "Better accent tutor–analysis and visualization of speech prosody," *Proceedings of InSTILL 2000*, pp. 86–89, 2000.

[34] M. Eskenazi, "The fluency pronunciation trainer," in *Proc. STiLL Workshop on Speech Technology in language learning, Marhallmen, 1998*, 1998.

[35] H. Franco, H. Bratt, R. Rossier, V. Rao Gadde, E. Shriberg, V. Abrash, and K. Precoda, "Eduspeak®: A speech recognition and pronunciation scoring toolkit for computer-aided language learning applications," *Language Testing*, vol. 27, no. 3, pp. 401–418, 2010.

[36] P. Martin, "Learning the prosodic structure of a foreign language with a pitch visualizer," in *Speech Prosody 2010-Fifth International Conference*, 2010.

[37] F. TALLEVI, "Teaching english prosody and pronunciation to italian speakers: the kaspar approach," Master's thesis, Politecnico di Milano, 2017.

[38] P. Boersma and V. Van Heuven, "Speak and unspeak with praat," *Glot International*, vol. 5, no. 9/10, pp. 341–347, 2001.

[39] E. Estebas-Vilaplana, "The teaching and learning of l2 english intonation in a distance education environment: Tl_tobi vs. the traditional models," *Linguistica*, vol. 57, no. 1, pp. 73–91, 2017.

[40] M. C. Pennington, P. Rogerson-Revell, M. C. Pennington, and P. Rogerson-Revell, "Using technology for pronunciation teaching, learning, and assessment," *English Pronunciation Teaching and Research: Contemporary Perspectives*, pp. 235–286, 2019.

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

[42] C. Cucchiarini and H. Strik, "Second language learners' spoken discourse: Practice and corrective feedback through automatic speech recognition," in *Smart technologies: Breakthroughs in research and practice*. IGI Global, 2018, pp. 367–389.

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

[44] J. Blake, N. Bogach, A. Zhuikov, I. Lezhenin, M. Maltsev, and E. Pyshkin, "Capt tool audio-visual feedback assessment across a variety of learning styles," in *2019 IEEE International Conferences on Ubiquitous Computing & Communications (IUCC) and Data Science and Computational Intelligence (DSCI) and Smart Computing, Networking and Services (SmartCNS)*. IEEE, 2019, pp. 565–569. [Online]. Available: https://doi.org/10.1109/IUCC/DSCI/SmartCNS.2019.00119

[45] D. Chun, "Signal analysis software for teaching discourse intonation," *Language Learning and Technology*, vol. 2, pp. 74–93, 1998. [Online]. Available: https://scholarspace.manoa.hawaii.edu/bitstreams/1404eb39-4d4e-417c-8788-0a7480e3b2f7/download

[46] A. Klapuri, "A method for visualizing the pitch content of polyphonic music signals." in *International Society for Music Retrieval*, 2009, pp. 615–620. [Online]. Available: https://archives.ismir.net/ismir2009/paper/000120.pdf

[47] Y. Permanasari, E. H. Harahap, and E. P. Ali, “Speech recognition using dynamic time warping (dtw),” in *Journal of Physics: Conference series*, vol. 1366. IOP Publishing, 2019, p. 012091. [Online]. Available: https://doi.org/10.1088/1742-6596/1366/1/012091

[48] A. Rilliard, A. Allauzen, and P. Boula_de_Mareüil, “Using dynamic time warping to compute prosodic similarity measures,” in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[49] F. Orsucci, R. Petrosino, G. Paoloni, L. Canestri, E. Conte, M. A. Reda, and M. Fulcheri, “Prosody and synchronization in cognitive neuroscience,” *EPJ Nonlinear Biomedical Physics*, vol. 1, no. 1, pp. 1–11, 2013.

[50] C. L. Webber and N. Marwan, *Recurrence quantification analysis: Theory and Best Practices*. Cham: Springer, 2015.

[51] J. Vásquez-Correa, J. Orozco-Arroyave, J. Arias-Londoño, J. Vargas-Bonilla, and E. Nöth, “Non-linear dynamics characterization from wavelet packet transform for automatic recognition of emotional speech,” in *Recent advances in nonlinear speech processing*. Cham: Springer, 2016, pp. 199–207.

[52] R. Fusaroli and K. Tylén, “Investigating conversational dynamics: Interactive alignment, interpersonal synergy, and collective task performance,” *Cognitive science*, vol. 40, no. 1, pp. 145–171, 2016.

[53] P. Van Geert, “A dynamic systems model of basic developmental mechanisms: Piaget, vygotsky, and beyond.” *Psychological review*, vol. 105, no. 4, p. 634, 1998.

[54] D. Larsen-Freeman, “Chaos/complexity science and second language acquisition,” *Applied linguistics*, vol. 18, no. 2, pp. 141–165, 1997.

[55] P. Hiver, A. H. Al-Hoorie, and R. Evans, “Complex dynamic systems theory in language learning: A scoping review of 25 years of research,” *Studies in Second Language Acquisition*, pp. 1–29, 2021.

[56] M. Verspoor and K. de Bot, “Measures of variability in transitional phases in second language development,” *International Review of Applied Linguistics in Language Teaching*, 2021.

[57] P. Chang and L. J. Zhang, “A cdst perspective on variability in foreign language learners’ listening development,” *Frontiers in Psychology*, vol. 12, p. 21, 2021. [Online]. Available: https://www.frontiersin.org/article/10.3389/fpsyg.2021.601962

[58] D. Larsen-Freeman, “On language learner agency: A complex dynamic systems theory perspective,” *The Modern Language Journal*, vol. 103, pp. 61–79, 2019.

[59] ——, “Saying what we mean: Making a case for ‘language acquisition’to become ‘language development’,” *Language Teaching*, vol. 48, no. 4, pp. 491–505, 2015.

[60] D. LaScotte, C. Meyers, and E. Tarone, “Voice and mirroring in sla: Top-down pedagogy for l2 pronunciation instruction,” *RELC Journal*, vol. 52, no. 1, pp. 144–154, 2021.

[61] P. L. van Geert, “Dynamic systems, process and development,” *Human development*, vol. 63, no. 3-4, pp. 153–179, 2019.

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

[63] L. S. Vygotsky, *Thought and language*. MIT press, 2012.

[64] M. Koopmans, "Education is a complex dynamical system: Challenges for research," *The Journal of Experimental Education*, vol. 88, no. 3, pp. 358–374, 2020.

[65] J. Hodgetts, *Pronunciation Instruction in English for Academic Purposes: An Investigation of Attitudes, Beliefs and Practices*. Springer, 2020.

[66] J. P. Lantolf, L. Kurtz, and O. Kisselev, "Understanding the revolutionary character of l2 development in the zpd: Why levels of mediation matter," *Language and Sociocultural Theory*, vol. 3, no. 2, pp. 153–171, 2016.

[67] K. M. Bragg, "Conversational movement dynamics and nonverbal indicators of second language development: A microgenetic approach," Ph.D. dissertation, University of Nevada, 2018.

[68] P. van Geert and H. Steenbeek, "The dynamics of scaffolding," *New ideas in Psychology*, vol. 23, no. 3, pp. 115–128, 2005.

[69] M. E. Poehner, J. Zhang, and X. Lu, "Computerized dynamic assessment (c-da): Diagnosing l2 development according to learner responsiveness to mediation," *Language testing*, vol. 32, no. 3, pp. 337–357, 2015.

[70] B. Tomlinson, "Assisting learners in orchestrating their inner voice for l2 learning." *Language Teaching Research Quarterly*, vol. 19, pp. 32–47, 2020.

[71] R. Ableeva, "Dynamic assessment of listening comprehension in second language learning," Ph.D. dissertation, The Pennsylvania State University, College of the Liberal Arts, 2010.

[72] P. van Geert, "Vygotskian dynamics of development," *Human development*, vol. 37, no. 6, pp. 346–365, 1994.

[73] M. Guevara, R. F. Cox, M. van Dijk, and P. van Geert, "Attractor dynamics of dyadic interaction: A recurrence based analysis," *Nonlinear Dynamics Psychology and Life Sciences*, vol. 21, no. 3, pp. 289–317, 2017.

[74] R. M. Lima Jr and U. K. Alves, "A dynamic perspective on l2 pronunciation development: bridging research and communicative teaching practice," *Revista do GEL*, vol. 16, no. 2, pp. 27–56, 2019.

[75] A. Bahari, X. Zhang, and Y. Ardasheva, "Establishing a nonlinear dynamic individual-centered language assessment model: a dynamic systems theory approach," *Interactive Learning Environments*, pp. 1–23, 2021.

[76] C. L. Nagle, "Assessing the state of the art in longitudinal l2 pronunciation research: Trends and future directions," *Journal of Second Language Pronunciation*, 2021.

[77] N. Esteve-Gibert and B. Guellaï, "Prosody in the auditory and visual domains: A developmental perspective," *Frontiers in Psychology*, vol. 9, p. 338, 2018.

[78] D. Liu, "Prosody transfer failure despite cross-language similarities: Evidence in favor of a complex dynamic system approach in pronunciation teaching," *Journal of Second Language Pronunciation*, vol. 7, no. 1, pp. 38–61, 2021.

[79] A. Shahzadi, A. Ahmadyfard, A. Harimi, and K. Yaghmaie, "Speech emotion recognition using nonlinear dynamics features." *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 23, 2015.

[80] S. Sood and A. Krishnamurthy, “A robust on-the-fly pitch (otfp) estimation algorithm,” in *Proceedings of the 12th annual ACM international conference on Multimedia*, 2004, pp. 280–283.

[81] D. E. Terez, “Robust pitch determination using nonlinear state-space embedding,” in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 2002, pp. I–345.

[82] N. Marwan, M. Carmen Romano, M. Thiel, and J. Kurths, “Recurrence plots for the analysis of complex systems,” *Physics Reports*, vol. 438, no. 5, pp. 237–329, 2007.

[83] E. Tzinis, G. Paraskevopoulos, C. Baziotis, and A. Potamianos, “Integrating recurrence dynamics for speech emotion recognition,” *Interspeech 2018*, Sep 2018.

[84] R. K. Thandil and K. M. Basheer, “Accent based speech recognition: A critical overview,” *Malaya Journal of Matematik (MJM)*, vol. 8, no. 4, 2020, pp. 1743–1750, 2020.

[85] Y. Singh, A. Pillay, and E. Jembere, “Features of speech audio for accent recognition,” in *2020 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)*, 2020, pp. 1–6.

[86] J. H. L. Hansen and L. M. Arslan, “Foreign accent classification using source generator based prosodic features,” in *1995 International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 1995, pp. 836–839 vol.1.

[87] H. Huang, X. Xiang, Y. Yang, R. Ma, and Y. Qian, “Aispeech-sjtu accent identification system for the accented english speech recognition challenge,” in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6254–6258.

[88] S. Deshpande, S. Chikkerur, and V. Govindaraju, “Accent classification in speech,” in *Fourth IEEE Workshop on Automatic Identification Advanced Technologies (AutoID’05)*, 2005, pp. 139–143.

[89] E. Tverdokhleb, H. Dobrovolskyi, N. Keberle, and N. Myronova, “Implementation of accent recognition methods subsystem for elearning systems,” in *2017 9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, vol. 2, 2017, pp. 1037–1041.

[90] A. Ensslin, T. Goorimoorthee, S. Carleton, V. Bulitko, and S. Poo Hernandez, “Deep learning for speech accent detection in video games,” *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, vol. 13, no. 1, Sep. 2017.

[91] P. Berjon, A. Nag, and S. Dev, “Analysis of French phonetic idiosyncrasies for accent recognition,” *Soft Computing Letters*, vol. 3, p. 100018, 12 2021.

[92] J. Bird, E. Wanner, A. Ekárt, and D. Faria, “Accent classification in human speech biometrics for native and non-native english speakers,” in *PErvasive Technologies Related to Assistive Environments (PETRA)*, 06 2019, pp. 554–560.

[93] Z. Zhang, Y. Wang, and J. Yang, “Accent recognition with hybrid phonetic features,” *Sensors*, vol. 21, no. 18, p. 6258, 2021.

[94] P. M. Rogerson-Revell, “Computer-assisted pronunciation training (capt): Current issues and future directions,” *RELC Journal*, vol. 52, no. 1, pp. 189–205, 2021.

[95] S.-W. F. Jiang, B.-C. Yan, T.-H. Lo, F.-A. Chao, and B. Chen, “Towards robust mispronunciation detection and diagnosis for l2 english learners with accent-modulating methods,” in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 1065–1070.

[96] M. Algabri, H. Mathkour, M. Alsulaiman, and M. A. Bencherif, “Mispronunciation detection and diagnosis with articulatory-level feedback generation for non-native arabic speech,” *Mathematics*, vol. 10, no. 15, 2022.

[97] S. Karpagavalli and E. Chandra, “A review on automatic speech recognition architecture and approaches,” *International Journal of Signal Processing, Image Processing and Pattern Recognition*, vol. 9, no. 4, pp. 393–404, 2016.

[98] P. Sullivan, T. Shibano, and M. Abdul-Mageed, “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*. Springer, 2022, pp. 21–44.

[99] T. Viglino, P. Motlicek, and M. Cernak, “End-to-end accented speech recognition.” in *Interspeech*, 2019, pp. 2140–2144.

[100] G. Zhao, S. Sonsaat, A. Silpachai, I. Lucic, E. Chukharev-Hudilainen, J. Levis, and R. Gutierrez-Osuna, “L2-arctic: A non-native english speech corpus.” in *Interspeech*, 2018, pp. 2783–2787.

[101] S. Aryal and R. Gutierrez-Osuna, “Can voice conversion be used to reduce non-native accents?” in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 7879–7883.

[102] S. H. Mohammadi and A. Kain, “An overview of voice conversion systems,” *Speech Communication*, vol. 88, pp. 65–82, 2017.

[103] K. Radzikowski, L. Wang, O. Yoshie, and R. Nowak, “Accent modification for speech recognition of non-native speakers using neural style transfer. eurasip j. audio speech music proc. 2021 (1), 1 (2021).”

[104] D. Felps, H. Bortfeld, and R. Gutierrez-Osuna, “Foreign accent conversion in computer assisted pronunciation training,” *Speech communication*, vol. 51, no. 10, pp. 920–932, 2009.

[105] A. De Meo, M. Vitale, M. Pettorino, F. Cutugno, and A. Origlia, “Imitation/self-imitation in computer-assisted prosody training for chinese learners of l2 italian,” *Pronunciation in Second Language Learning and Teaching Proceedings*, vol. 4, no. 1, 2012.

[106] D. Vigliano, E. Pellegrino *et al.*, “Self-imitation in prosody training: a study on japanese learners of italian,” in *Proceedings SLaTE 2015. Sixth Workshop on Speech and Language Technology in Education.* ISCA Special Interest Group SLaTE, 2015, 2015, pp. 53–57.

[107] S. Ding, C. Liberatore, S. Sonsaat, I. Lučić, A. Silpachai, G. Zhao, E. Chukharev-Hudilainen, J. Levis, and R. Gutierrez-Osuna, “Golden speaker builder–an interactive tool for pronunciation training,” *Speech Communication*, vol. 115, pp. 51–66, 2019.

[108] S. Ding, G. Zhao, and R. Gutierrez-Osuna, “Accentron: Foreign accent conversion to arbitrary non-native speakers using zero-shot learning,” *Computer Speech & Language*, vol. 72, p. 101302, 2022.

[109] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. Weiss, R. Clark, and R. A. Saurous, “Towards end-to-end prosody transfer for expressive speech synthesis with tacotron,” in *international conference on machine learning*. PMLR, 2018, pp. 4693–4702.

[110] E. Boitsova, E. Pyshkin, Y. Takako, N. Bogach, I. Lezhenin, A. Lamtev, and V. Diachkov, “Studyintonation courseware kit for efl prosody teaching,” in *Proceedings of the 9th International Conference on Speech Prosody*, 2018, pp. 413–417.

[111] T. Audacity, “Audacity,” *The Name Audacity (R) Is a Registered Trademark of Dominic Mazzoni Retrieved from http://audacity. sourceforge. net*, 2017.

[112] L. Henrichsen, “A system for analyzing and evaluating computer-assisted second-language pronunciation-teaching websites and mobile apps,” in *Society for Information Technology & Teacher Education International Conference*. Association for the Advancement of Computing in Education (AACE), 2019, pp. 963–968.

[113] A. Kuznetsov, A. Lamtev, I. Lezhenin, A. Zhuikov, M. Maltsev, E. Boitsova, N. Bogach, and E. Pyshkin, “Cross-platform mobile call environment for pronunciation teaching and learning,” in *SHS Web of Conferences*, vol. 77. EDP Sciences, 2020, p. 01005.

[114] E. Grabe, F. Nolan, and K. J. Farrar, “Ivie-a comparative transcription system for intonational variation in english,” in *Fifth International Conference on Spoken Language Processing*, 1998.

[115] Z.-H. Tan, N. Dehak *et al.*, “rvad: An unsupervised segment-based robust voice activity detection method,” *Computer speech & language*, vol. 59, pp. 1–21, 2020.

[116] Y. Hu and P. C. Loizou, “Subjective comparison and evaluation of speech enhancement algorithms,” *Speech communication*, vol. 49, no. 7-8, pp. 588–601, 2007.

[117] D. Povey. Kaldi asr. [Online]. Available: http://www.kaldi-asr.org/index.html

[118] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: an asr corpus based on public domain audio books,” in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[119] D. Povey, A. Ghoshal, G. Boulianne, N. Goel, M. Hannemann, Y. Qian, P. Schwarz, and G. Stemmer, “The kaldi speech recognition toolkit, workshop on automatic speech recognition and understanding,” *US IEEE Signal Processing Society, Hilton Waikoloa Village, Big Island, Hawaii*, 2011.

[120] R. Delmonte, “Prosodic tools for language learning,” *International Journal of Speech Technology*, vol. 12, pp. 161–184, 2009.

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

[122] A. Tan, “Study intonation: A mobile-assisted pronunciation training application,” *Teaching English as a Second Language Electronic Journal (TESL-EJ)*, vol. 25, no. 3, 2021.

[123] W. M. Lowie and M. H. Verspoor, “Individual differences and the ergodicity problem,” *Language Learning*, vol. 69, pp. 184–206, 2019.

[124] P. Hiver and A. H. Al-Hoorie, “A dynamic ensemble for second language research: Putting complexity theory into practice,” *The Modern Language Journal*, vol. 100, no. 4, pp. 741–756, 2016.

[125] ——, *Research Methods for Complexity Theory in Applied Linguistics*. Multilingual Matters, 2019. [Online]. Available: https://doi.org/10.21832/9781788925754

[126] R. Godwin-Jones, “Chasing the butterfly effect: Informal language learning online as a complex system,” *Language Learning & Technology*, vol. 22, no. 2, pp. 8–27, 2018.

[127] D. M. Hardison, “Multimodal input in second-language speech processing,” *Language Teaching*, vol. 54, no. 2, pp. 206–220, 2021.

[128] M. C. Pennington, “Teaching pronunciation: The state of the art 2021,” *RELC Journal*, vol. 52, no. 1, pp. 3–21, 2021.

[129] E. Tzinis, G. Paraskevopoulos, C. Baziotis, and A. Potamianos, “Integrating recurrence dynamics for speech emotion recognition,” *arXiv preprint arXiv:1811.04133*, 2018.

[130] K. De Bot, W. Lowie, and M. Verspoor, “A dynamic systems theory approach to second language acquisition,” *Bilingualism: Language and cognition*, vol. 10, no. 1, pp. 7–21, 2007.

[131] M. B. Kennel, R. Brown, and H. D. I. Abarbanel, “Determining embedding dimension for phase-space reconstruction using a geometrical construction,” *Phys. Rev. A*, vol. 45, pp. 3403–3411, Mar 1992.

[132] K. Saito and L. Plonsky, “Effects of second language pronunciation teaching revisited: A proposed measurement framework and meta-analysis,” *Language Learning*, vol. 69, no. 3, pp. 652–708, 2019.

[133] M. G. O’Brien, T. M. Derwing, C. Cucchiarini, D. M. Hardison, H. Mixdorff, R. I. Thomson, H. Strik, J. M. Levis, M. J. Munro, J. A. Foote *et al.*, “Directions for the future of technology in pronunciation research and teaching,” *Journal of Second Language Pronunciation*, vol. 4, no. 2, pp. 182–207, 2018.

[134] Y. Feng, G. Fu, Q. Chen, and K. Chen, “Sed-mdd: Towards sentence dependent end-to-end mispronunciation detection and diagnosis,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 3492–3496.

[135] S. S. Malla, “Acoustic features based accent classification of Kashmiri language using deep learning,” *Global Journal of Computer Science and Technology*, 2022.

[136] C. Graham, “L1 identification from l2 speech using neural spectrogram analysis,” *Interspeech*, 2021. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2021-1545

[137] A. Ahamad, A. Anand, and P. Bhargava, “Accentdb: A database of non-native english accents to assist neural speech recognition,” 2020.

[138] F. Oladipo, R. A. Habeeb, and A. E. Musa, “Accent identification of ethnically diverse nigerian english speakers,” *SSRN Electronic Journal*, 9 2020.

[139] M. Aswathi Sanal, “Accent recognition for malayalam speech signals,” *International Journal of Innovative Research in Computer and Communication Engineering*, 2017.

[140] Y. Ma, M. Paulraj, S. Yaacob, A. Shahriman, and S. K. Nataraj, “Speaker accent recognition through statistical descriptors of mel-bands spectral energy and neural network model,” in *2012 IEEE Conference on Sustainable Utilization and Development in Engineering and Technology (STUDENT)*, 2012, pp. 262–267.

[141] Q. T. Duong *et al.*, “Development of accent recognition systems for Vietnamese speech,” in *2021 24th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA).* IEEE, 2021, pp. 174–179.

[142] G. R. Krishna, R. Krishnan, and V. K. Mittal, “A system for automatic regional accent classification,” in *2020 IEEE 17th India Council International Conference (INDICON).* IEEE, 2020, pp. 1–5.

[143] J. Cheng, N. Bojja, and X. Chen, “Automatic accent quantification of Indian speakers of English.” in *Interspeech*, 2013, pp. 2574–2578.

[144] A. Lazaridis, E. el Khoury, J.-P. Goldman, M. Avanzi, S. Marcel, and P. N. Garner, “Swiss French regional accent identification.” in *Odyssey*, 2014.

[145] Y. Jiao, M. Tu, V. Berisha, and J. M. Liss, “Accent identification by combining deep neural networks and recurrent neural networks trained on long and short term features.” in *Interspeech*, 2016, pp. 2388–2392.

[146] F. Weninger, Y. Sun, J. Park, D. Willett, and P. Zhan, “Deep learning based Mandarin accent identification for accent robust asr.” in *INTERSPEECH*, 2019, pp. 510–514.

[147] G. Işik and H. Artuner, “Turkish dialect recognition using acoustic and phonotactic features in deep learning architectures,” *Journal of Information Technologies*, vol. 13, no. 3, pp. 207–216, 2020.

[148] R. Kethireddy, S. R. Kadiri, P. Alku, and S. V. Gangashetty, “Mel-weighted single frequency filtering spectrogram for dialect identification,” *IEEE Access*, vol. 8, pp. 174 871–174 879, 2020.

[149] George Mason University, “Speech accent archive,” 2021. [Online]. Available: https://accent.gmu.edu/

[150] F. Zheng, G. Zhang, and Z. Song, “Comparison of different implementations of mfcc,” *Journal of Computer Science and Technology*, vol. 16, pp. 582–589, 11 2001.

[151] S. Liu, D. Wang, Y. Cao, L. Sun, X. Wu, S. Kang, Z. Wu, X. Liu, D. Su, D. Yu *et al.*, “End-to-end accent conversion without using native utterances,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2020, pp. 6289–6293.

[152] L. Rasier and P. Hiligsmann, “Prosodic transfer from l1 to l2. theoretical and methodological issues,” *Nouveaux Cahiers de Linguistique Française*, vol. 28, 01 2007.

[153] H. K. Alain de Cheveigné, “Yin, a fundamental frequency estimator for speech and music,” *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[154] H. Shimodaira, “Improving predictive inference under covariate shift by weighting the log-likelihood function,” *Journal of Statistical Planning and Inference*, vol. 90, no. 2, pp. 227–244, 2000.

[155] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *Proceedings of the 32nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, F. Bach and D. Blei, Eds., vol. 37. Lille, France: PMLR, 07–09 Jul 2015, pp. 448–456.

[156] J. G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, and F. Herrera, “A unifying view on dataset shift in classification,” *Pattern Recognition*, vol. 45, no. 1, pp. 521–530, 2012.

[157] G. D. Y, N. G. Nair, P. Satpathy, and J. Christopher, “Covariate shift: A review and analysis on classifiers,” in *2019 Global Conference for Advancement in Technology (GCAT)*, 2019, pp. 1–6.

[158] S. Bock and M. Weiß, “A proof of local convergence for the adam optimizer,” in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–8.

[159] N. T. Longford, “A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects,” *Biometrika*, vol. 74, no. 4, pp. 817–827, 1987.

[160] T. Wu, J. Duchateau, J.-P. Martens, and D. Van Compernolle, “Feature subset selection for improved native accent identification,” *Speech Communication*, vol. 52, no. 2, pp. 83–98, 2010.

[161] L. Sun, T. Wang, W. Ding, J. Xu, and Y. Lin, “Feature selection using fisher score and multilabel neighborhood rough sets for multilabel classification,” *Information Sciences*, vol. 578, pp. 887–912, 2021.

[162] B. Milner and X. Shao, “Prediction of fundamental frequency and voicing from mel-frequency cepstral coefficients for unconstrained speech reconstruction,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 24–33, 2007.

[163] N. N. Van, S. L. Xuan, I. Lezhenin, N. Bogach, and E. Pyshkin, “Adopting studyintonation capt tools to tonal languages through the example of vietnamese,” in *SHS Web of Conferences*, vol. 102. EDP Sciences, 2021, p. 01007.

[164] D. Büring, *Intonation and meaning*. Oxford University Press, 2016.

[165] E. Pyshkin, A. Kusakari, J. Blake, N. B. Pham, and N. Bogach, “Multimodal modeling of the mora-timed rhythm of japanese and its application to computer-assisted pronunciation training,” in *2023 14th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*. IEEE, 2023, pp. 174–179.

[166] E. Pyshkin, J. Blake, A. Lamtev, I. Lezhenin, A. Zhuikov, and N. Bogach, “Prosody training mobile application: Early design assessment and lessons learned,” in *2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, vol. 2. IEEE, 2019, pp. 735–740.

[167] R. J. Vilches, “Who is in charge? an l2 discourse intonation study on four prosodic parameters to exert the pragmatic function of dominance and control in the context of l2 non-specialist public speaking,” *Complutense Journal of English Studies*, vol. 23, p. 33, 2015.

[168] D. Buckingham and M. Scanlon, “That is edutainment: media, pedagogy and the market place,” in *International forum of researchers on young people and the media, Sydney*, 2000.

[169] J. Chakraborty, R. Sinha, and P. Sarmah, “Influence of accented speech in automatic speech recognition: A case study on assamese l1 speakers speaking code switched hindi-english,” in *International Conference on Speech and Computer*. Springer, 2022, pp. 87–98.

[170] J. Kominek, “Cmu arctic databases for speech synthesis,” *CMU-LTI*, 2003.

[171] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[172] J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P.-E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen *et al.*, “Libri-light: A benchmark for asr with limited or no supervision,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7669–7673.

[173] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino *et al.*, “Xls-r: Self-supervised cross-lingual speech representation learning at scale,” *arXiv preprint arXiv:2111.09296*, 2021.

[174] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[175] A. e. a. Paszke, “Pytorch: An imperative style, high-performance deep learning library,” in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035.

[176] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, “Huggingface’s transformers: State-of-the-art natural language processing,” *arXiv preprint arXiv:1910.03771*, 2019.

[177] S. Gondi, “Wav2vec2.0 on the edge: Performance evaluation,” *arXiv preprint arXiv:2202.05993*, 2022.

[178] S. Ishikawa, *The ICNALE Guide: An Introduction to a Learner Corpus Study on Asian Learners’ L2 English*. Taylor & Francis, 2023.

[179] V. Mikhailava, E. Pyshkin, J. Blake, S. Chernonog, I. Lezhenin, R. Svechnikov, and N. Bogach, “Tailoring computer-assisted pronunciation teaching: Mixing and matching the mode and manner of feedback to learners,” in *INTED2022 Proceedings*. IATED, 2022, pp. 767–773.

[180] J. Blake, N. Bogach, A. Kusakari, I. Lezhenin, V. Khaustova, S. L. Xuan, V. N. Nguyen, N. B. Pham, R. Svechnikov, A. Ostapchuk *et al.*, “An open capt system for prosody practice: Practical steps towards multilingual setup,” *Languages*, vol. 9, no. 1, p. 27, 2024.

[181] E. Pyshkin and J. Blake, “Music and choreography metaphors in spoken language rhythm modelling and their application to computer-assisted pronunciation training for mora-timed japanese,” 2023.

[182] V. Mikhailava, M. Lesnichaia, N. Bogach, I. Lezhenin, J. Blake, and E. Pyshkin, “Language accent detection with cnn using sparse data from a crowd-sourced speech archive,” *Mathematics*, vol. 10, no. 16, p. 2913, 2022.

[183] V. Khaustova, E. Pyshkin, V. Khaustov, J. Blake, and N. Bogach, “Capturing accents: An approach to personalize pronunciation training for learners with different l1 backgrounds,” in *International Conference on Speech and Computer*. Springer, 2023, pp. 59–70.