5-2025

Phoneme Recognition for Pronunciation Improvement

Matthew Heywood
mheywood@unomaha.edu

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

Honors Capstone Report

in partial fulfillment for the degree
Bachelor of Science in Computer Science (Honors Distinction)
in Fall/Spring 2025

Phoneme Recognition for Pronunciation Improvement

Submitted by:

Matthew Heywood
E-Mail: mheywood@unomaha.edu
B.S. Computer Engineering

Abstract

This project aims to improve English pronunciation learning by investigating the origins of speech errors and developing a prototype tool to provide precise feedback to users. Drawing on foundational knowledge of speech errors, the study explores the creation of a new pronunciation tool meant to offer localized feedback, pinpointing specific errors and suggesting corrective measures. By addressing the limitations of existing approaches, the research endeavors to offer a more effective method for individuals seeking to refine their English pronunciation skills.

Incorporating cutting-edge technology, the developed tool harnesses the power of speech-to-phoneme AI models to streamline the process of English pronunciation refinement. By leveraging AI in tandem with modified lazy string matching algorithms, the tool compares the user's spoken input with the intended pronunciation, enabling a granular analysis of discrepancies. This innovative approach not only identifies specific speech errors but also provides users with actionable insights into the phonetic nuances of their pronunciation. The utilization of speech-to-phoneme AI models represents a significant leap forward in pronunciation instruction, offering users a personalized learning experience that automates some of the traditional methods used by speech-language pathologists. Through the seamless integration of artificial intelligence, the tool facilitates real-time feedback, allowing users to pinpoint areas of improvement with unparalleled accuracy. By harnessing the capabilities of AI-driven technology, this tool not only enhances the efficacy of pronunciation learning but also paves the way for future advancements in language education.

Contents

List of Figures

1 Introduction

Individuals who are deaf or hard of hearing (DHH) typically require specialized support to enhance their communication skills. Speech-language pathologists (SLPs) frequently provide therapy sessions to improve the pronunciation and comprehension of spoken language for these groups. This paper introduces an extension to "Pronunciation Pal," a capstone project dedicated to assisting DHH users in refining their English pronunciation through a range of innovative tools, including pictures and speech-sound diagrams. The extension focuses on incorporating a speech-to-phoneme AI model that provides localized, targeted feedback to help users improve their pronunciation more efficiently.

The newly added speech-to-phoneme capability is designed to offer precise, real-time feedback by analyzing users' spoken words and identifying specific areas where pronunciation errors occur. It then provides practical tips for correction, thereby enabling users to make targeted improvements. The need for such targeted feedback arises from the complexity of English phonemes and the frequent difficulty DHH individuals face in articulating certain sounds. Generalized feedback often fails to address the nuanced needs of these users, making specific, localized guidance imperative for effective learning.

The University of Nebraska Medical Center Munroe-Meyer Institute (MMI) has sponsored this project, leveraging its extensive experience in providing services to individuals with disabilities, including speech and language therapy. A key contributor to the project is Korey Stading, MS, CCC-SLP, a certified speech-language pathologist with over 20 years of experience in speech and language evaluation and therapy. Her expertise has been instrumental in ensuring that Pronunciation Pal is grounded in sound therapeutic practices and is feasible for real-world application.

Pronunciation Pal is designed to be accessible from anywhere, being deployed as a web-based application. Users can create profiles that store lists of words they are currently focusing on, making it easy to track progress and revisit challenging pronunciations. The addition of the speech-to-phoneme AI model significantly enhances the platform by providing

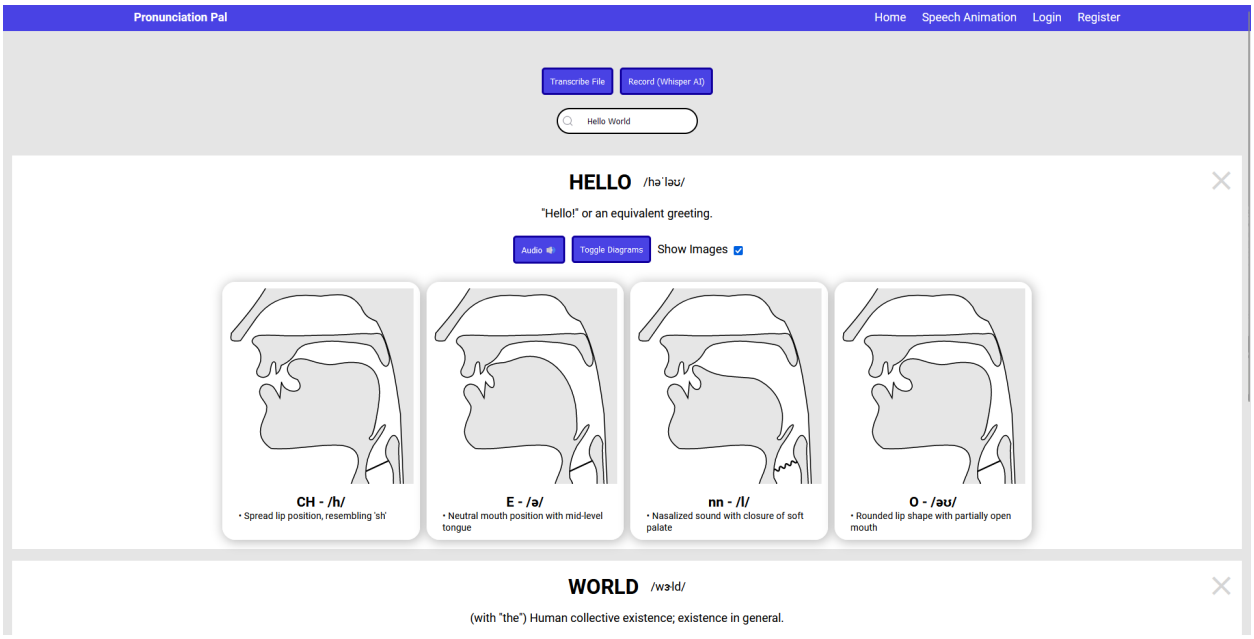

Figure 1: Pronunciation Pal Web Application

more precise and actionable feedback to users, thereby making their learning process more efficient and targeted.

Despite the necessity and potential of live testing to validate and fine-tune the application, our development process was limited to feedback from the developers and consultations with expert SLPs. Future work will aim to include comprehensive user testing to gather real-world efficacy data and further improve the application.

In summary, this paper focuses on the extension of Pronunciation Pal with a speech-to-phoneme capability, aspiring to be a valuable supplemental tool for individuals looking to improve their pronunciation skills, particularly those who are DHH. By combining expert consultation, cutting-edge technology, and accessible design, this project aims to offer an effective resource for speech therapy.

2 Background

The English language presents a unique challenge due to the complex and often irregular relationship between written letters and spoken sounds. Words such as "through,"

"island," "receipt," and "knock" exemplify the disconnection between orthography and pronunciation. This discrepancy poses significant obstacles for language learners and individuals with speech impairments. Unlike many languages with more predictable spelling rules, English lacks a one-to-one correspondence between graphemes and phonemes, a feature shared with other languages such as French and Danish. However, languages like Spanish and Finnish have much more regular orthographic systems, where letters more consistently map to sounds. Understanding the foundational concepts of phonemes, graphemes, and phonetic spelling is crucial for addressing these challenges and will be explored in this section.

2.1 Phonemes

Phonemes are the smallest units of sound in a language, functioning as the building blocks for spoken words. For example, the sounds /t/, /k/, and /n/ are distinct phonemes in English. The language contains 44 phonemes that encompass all the sounds heard within its vocabulary. Phonemes are not universal; they vary between languages and even among dialects of the same language. For instance, English distinguishes between the phonemes /v/ and /w/, while some other languages might not.

Different languages employ varying numbers and types of phonemes, making understanding phonemes very helpful for accurate pronunciation. The application of phonemes is critical in speech therapy, language instruction, and language processing technologies. Accurate identification and articulation of phonemes are fundamental to mastering spoken English, representing a core challenge that Pronunciation Pal aims to address.

2.2 International Phonetic Alphabet

Given the irregularities and inconsistencies in English orthography, dictionaries often provide phonetic spellings to indicate how words should be pronounced. Phonetic spelling employs characters or symbols for phonemes, ensuring a one-to-one correspondence between

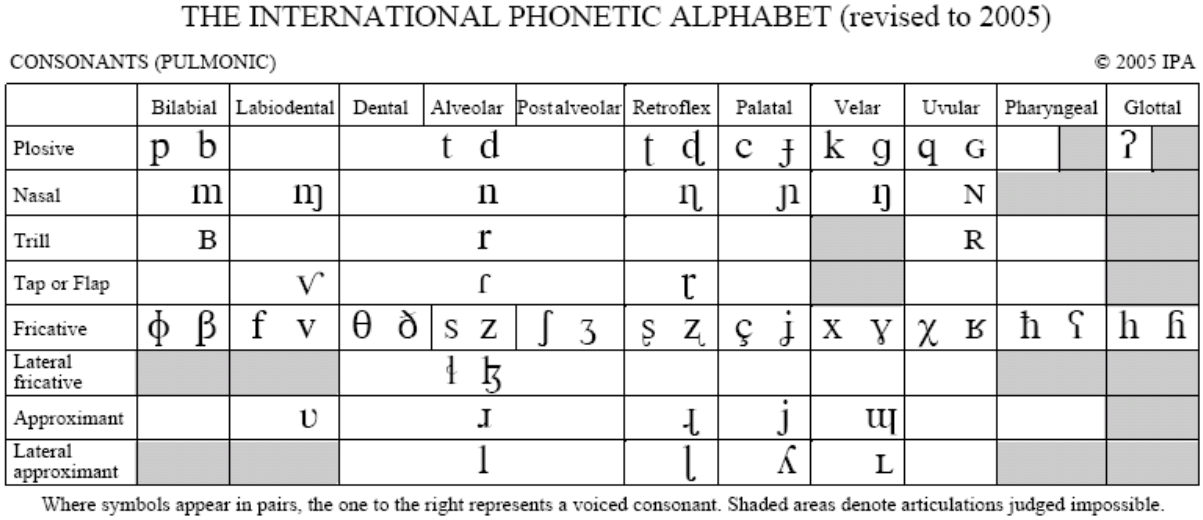

Figure 2: IPA Pulmonic Consonants (International Phonetic Association, 2024)

letters and sounds. A common dictionary representation might depict "thought" as "thawt," providing a clear guide to pronunciation using relatively intuitive graphemes.

Phonetic alphabets take this a step further by defining specific characters that each map directly to a sound. This standardized approach reduces ambiguity and aids in accurate pronunciation. The most widely adopted phonetic alphabet is the International Phonetic Alphabet (IPA). The IPA offers a standardized set of symbols representing the phonemes of all languages. For instance, the IPA notation for "thought" is /θɔːt/, and for "phone" it is /fəʊn/. The IPA provides a reliable method for accurate phonetic representation, which is also utilized in Pronunciation Pal to ensure precise feedback.

2.3 Graphemes

Graphemes are the written representations of phonemes using the alphabet of a given language; they compose the orthographic spelling of words according to the conventions and rules of that specific alphabet. In English, graphemes can take several forms:

- Single letters, like 'e' or 'g'.
- Digraphs, where two letters represent one sound, such as 'wh,' 'ck,' 'ea,' or 'ng.'

- Trigraphs, where three letters represent a single sound, such as 'tch' in "match."

Despite their primary function of representing phonemes, graphemes don't always correspond directly to phonetic sounds. For instance, the letter 'g' in "gate" represents the /g/ sound, while in "giant" it represents the /ʤ/ sound. Understanding grapheme-phoneme mappings is critical for learning to read and write in any language.

Even with an understanding of graphemes and phonemes, the relationship between spelling and pronunciation in English often remains inconsistent. For example, the letter 'o' sounds different in "pot," "move," and "goat." This complexity calls for advanced tools to aid correct pronunciation, tools such as Pronunciation Pal, which integrates AI models to assist users by offering explicit feedback.

2.4 Levenshtein Distance

Levenshtein Distance, also known as an edit distance, is a metric for measuring the difference between two sequences. It quantifies how many single-character edits (insertions, deletions, or substitutions) are required to transform one sequence into another. Named after the Soviet mathematician Vladimir Levenshtein who devised it in 1965, this metric has significant applications in various fields such as computational linguistics, bioinformatics, and spell-checking algorithms.

The Levenshtein Distance between two words "kitten" and "sitting," for instance, is calculated as follows:

- Change 'k' to 's': kitten → sitten (Substitution)
- Replace 'e' with 'i': sitten → sittin (Substitution)
- Add 'g' at the end: sittin → sitting (Insertion)

The Levenshtein Distance here is 3, as three edits are needed to transform "kitten" into "sitting."

In language learning and speech pathology, the Levenshtein Distance can be used to measure the accuracy of pronunciation. By comparing the phonetic transcription of a user's spoken word to the correct phonetic transcription, the tool can quantify how many phonemic edits are necessary to match the target pronunciation. This metric allows for precise, actionable feedback, guiding users ever closer to accurate pronunciation.

3 Summaries

The paper "Simple and Effective Zero-Shot Cross-Lingual Phoneme Recognition" by Qiantong Xu, Alexei Baevski, and Michael Auli from Facebook AI Research innovatively leverages labeled data from related languages to enhance zero-shot cross-lingual transfer learning in phoneme recognition (Xu et al., 2021). The authors utilize a multilingually pretrained wav2vec2 model fine-tuned to transcribe unseen languages, mapping phonemes between training and target languages using articulatory features. The experiment's results show this simple method significantly outperforms prior approaches that incorporated task-specific architectures and only partially used monolingual models, effectively providing a more comprehensive and efficient solution.

The research addresses the critical gap in speech technology's accessibility for the vast number of languages lacking large amounts of transcribed speech audio. Traditional approaches requiring separate unsupervised models for each language often overlook labeled data from related languages. The authors' zero-shot transfer learning method circumvents this by training a singular multilingual model on available labeled data from various languages. This not only facilitates the transcription of new, unseen languages but also offers significant performance benefits by consolidating phonological units into a global phoneme recognizer. The use of articulatory features enhances the mapping between phonemes, creating a robust lexicon for unseen languages without assuming direct relationships among the training and testing languages.

Furthermore, the method demonstrates impressive efficacy on extensive datasets, including CommonVoice, BABEL, and MLS. The study's comparison between pretraining methods reveals the superior performance of cross-lingual pretrained representations over monolingual ones. Additionally, it highlights the advantage of utilizing the full pretrained model. The findings underscore the approach's capability to handle multiple unseen languages efficiently through a single model, marking a notable step forward in unsupervised and semi-supervised learning applications in speech recognition technology.

This research was particularly useful for Pronunciation Pal's targeted feedback extension, as the researchers published their trained model. In particular for this project, the wav2vec2-ljspeech-gruut variant was used, as it was the most recently updated. The insights from utilizing articulatory features to map phonemes between languages were also helpful, enabling a more comprehensive and efficient solution for providing precise pronunciation feedback.

The next paper, "English Mispronunciation Detection Module Using a Transformer Network Integrated into a Chatbot," addresses a critical challenge in business intelligence and language learning: the accurate recognition and correction of mispronunciations in spoken English (Martinez-Quezada et al., 2022). The study aids non-native speakers in improving their pronunciation by employing an advanced Automatic Speech Recognizer (ASR) based on the Transformer network, integrated into a chatbot interface. This integration allows for seamless, conversational interaction, essential for language learners. The approach hinges on converting audio input into textual representation and applying the Levenshtein distance algorithm to highlight discrepancies between user-provided transcription and ASR output, combining the robustness of deep learning with natural language processing techniques.

The core innovation lies in utilizing the Transformer network, originally designed for machine translation tasks, to enhance ASR systems significantly. Unlike traditional models such as Hidden Markov Models (HMMs) and Recurrent Neural Networks (RNNs), which face limitations in state updates and computational efficiency, the Transformer processes

sequences in parallel, ensuring faster and more reliable outputs. The ASR module was trained on diverse datasets, including LibriSpeech and L2-ARCTIC, encompassing both native and non-native English accents. This broad training ensures the model's ability to generalize across different speech patterns, making it highly applicable in real-world scenarios.

This study was useful to the project as it provides insight into methods of combining AI with other methods in language processing in order to give better feedback. The approach of using Levenshtein distance was one of the inspirations behind the final algorithm designed for providing feedback in this extension.

4 Implementation

The implementation of the targeted feedback extension into Pronunciation Pal was a multi-faceted endeavor, segmented into three primary tasks. Each task addressed a critical component of the system, ensuring that the final product could provide precise and actionable feedback to users focusing on pronunciation improvement.

4.1 API Creation for Phoneme Transcription

The first task involved transforming the wav2vec2-ljspeech-gruut model into an accessible API. Given the extensive availability of transformer and machine learning libraries in Python, compared to their limited presence in JavaScript, it was decided to implement this functionality as an external service. Pronunciation Pal would then make queries to this service for transcription purposes.

To achieve this, the Flask web framework was employed to develop a straightforward web server capable of handling POST requests. These requests would include an audio file directed to a specific endpoint. Upon receiving the audio file, the server would process it through the transformer model and return the phonetic transcription to the main application. During initial testing, a minor issue with duplicate consonant phonemes was

identified, prompting the incorporation of a filter to eliminate such duplicates before sending the response back to the client.

4.2 Generating Meaningful Feedback

The second and most complex task was to convert the transcribed recording's phoneme string and the target phoneme string into a format conducive to delivering user feedback. Several key steps were involved:

- Insertion Feedback: Users needed to be informed about any extraneous phonemes they were introducing, providing them with the opportunity to refine their attempts.
- Deletion Feedback: Users had to be alerted if they omitted essential phonemes, enabling them to focus on including these in future pronunciations.
- Substitution Feedback: Mispronounced phonemes required specific and actionable advice. For instance, transitioning from the "o" in "cost" to "most" necessitates slightly altering mouth positioning. Exact details on feedback generation are discussed in the subsequent subsection; however, a clear understanding of these requirements guided this phase.

After thorough research, a decision was made to modify the Levenshtein distance algorithm for this context. Levenshtein distance calculates the minimal number of single-character edits (substitutions, insertions, and deletions) needed to transform one string into another. The modification involved adapting the algorithm to output a detailed list of the required edits, which would directly inform the feedback mechanism.

4.3 Delivering Pronunciation Feedback

The third and final task was to interpret the edit list generated by the modified Levenshtein algorithm and convert it into user-friendly feedback to correct pronunciation.

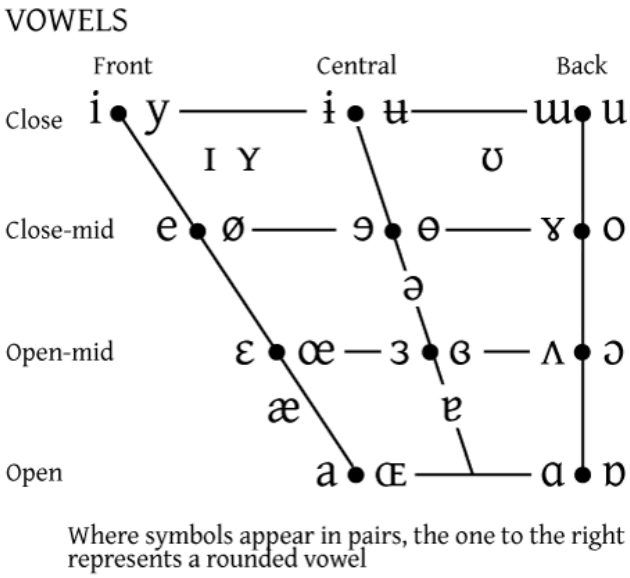

Figure 3: IPA Vowels Chart (International Phonetic Association, 2024)

- Insertion and Deletion Feedback: This was relatively straightforward, providing clear instructions such as "This phoneme is not necessary" or "Add a /b/ sound here."
- Substitution Feedback: This aspect was more intricate. Due to time constraints, the prototype focused primarily on vowels. Future extensions may include consonant feedback.

To address vowel substitutions, a mapping to a coordinate system based on the standard IPA vowel phoneme chart was implemented (see Figure 3). The horizontal axis represented tongue position, while the vertical axis indicated jaw placement. By calculating the differences between the coordinates of substituted vowels, modifiers like "a lot" or "a little" were indexed. These were then integrated into feedback phrases such as "open your mouth a little more" or "place your tongue significantly farther back," providing users with specific, actionable guidance on how to improve their pronunciation.

The source code for this extension can be found at `https://github.com/cttomcak/Team-12-SWAG-Pronunciation-Pal/tree/phoneme-analysis`

5 Results

Despite the limitations imposed by confidentiality agreements, HIPAA regulations, and other legal concerns that precluded the use of actual patients for testing, the project still underwent evaluation through alternative means. The primary test subject for this project was myself, supplemented by periodic volunteer contributions from friends and family who also engaged in testing the application. As the primary test subject, I was able to conduct extensive, iterative trials to refine the system. I recorded numerous audio samples, processed them through the API, and closely examined the feedback provided. Each iteration informed subsequent adjustments to both the transcription process and the feedback algorithm. This hands-on approach allowed for immediate problem identification and resolution, ensuring that the product met its intended objectives with high accuracy.

Friends and family members volunteered to test the system sporadically, providing additional data points for analysis. Their varied dialects, accents, and speech patterns offered a broader spectrum of testing conditions. The feedback garnered from these sessions highlighted the system's flexibility and effectiveness across different user profiles, thereby enhancing its robustness and adaptability.

The feedback engine identified and corrected mispronunciations, phoneme insertions, and deletions with decent accuraccy, although this could probably be fine tuned by training the model more. Vowels were not always as accurate as I had hoped, though whether this was due to the model or my inadequacies I do not know. Users received comprehensive advice that was easily understood and actionable, as illustrated in Figure 4. This figure shows an early mockup of a user-friendly interface, where specific phonetic corrections are indicated, providing clear and concise feedback. Unfortunately, due to the way the base Pronunciation Pal groups phonemes into visemes, highlighting improper characters was not feasible, although future iterations of this project could rework Pronunciation Pal to fix this issue. A demo of the final prototype can be found here: `https://youtu.be/9l1V0DaUF0s`.

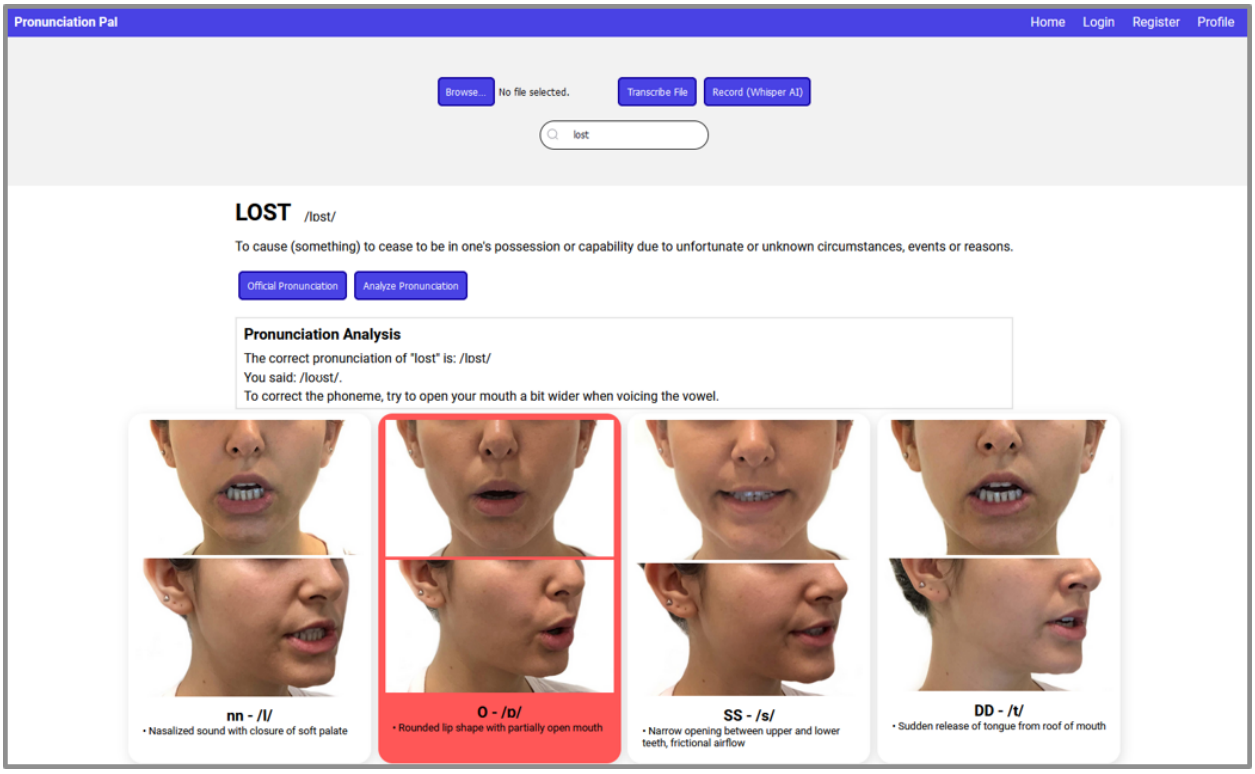

Figure 4: Demonstration of Feedback Engine

While the results indicate a promising start, several areas for future development and enhancement were identified:

- Substitution Feedback for Consonants: Currently, the prototype focuses on vowel substitutions. Expanding this capability to include consonant substitution would provide a more comprehensive pronunciation tool.

- Enhanced Feedback for Insertions and Deletions: Although the system provides basic feedback for phoneme insertions and deletions, refining these responses would offer more detailed guidance for users. Enhanced feedback could prevent repetitive mistakes and further improve pronunciation accuracy.

- Addressing Diphthongs: Diphthongs, which involve complex vowel sounds, pose an additional challenge. Developing specialized feedback for diphthongs would enable the system to address a broader range of speech nuances.

6 Challenges

The development of the targeted feedback extension for Pronunciation Pal was filled with a series of complex challenges, each necessitating specialized approaches and considerable problem-solving. The two main difficulties encountered during the project were related to managing duplicate consonants in the model output and modifying the Levenshtein distance algorithm for tailored feedback generation.

6.1 Handling Duplicate Consonants

One of the significant obstacles was the issue of the model outputting duplicate consonants in certain phonemic transcriptions. Several strategies were employed to address this, including:

- Changing Sample Rates: Initial attempts involved altering the audio sample rates fed into the model, hypothesizing that discrepancies in sampling accuracy might be contributing to the duplications. However, these attempts yielded limited success.

- Browser Audio Sample Rate Matching: Another strategy involved supplying the model with the audio sample rate directly obtained from the browser. This approach aimed to ensure consistency between input and processing stages but also proved inadequate in resolving the issue.

- Post-Processing Algorithm: Ultimately, the solution materialized through an algorithmic post-processing step that eliminated the duplicate consonant phonemes from the output. This pragmatic approach, although effective, underscored a limitation in my understanding of the internal workings of the AI model. It became evident that deeper expertise in AI might enable a more intrinsic fix within the model itself, yet without a clear starting point for such an endeavor, the implemented workaround was deemed satisfactory for the scope of this project.

6.2 Modifying the Levenshtein Algorithm

The necessity to adapt the Levenshtein distance algorithm to output a list of phonetic edits posed another formidable challenge. Initially, the complexity arose from a lack of in-depth understanding of the algorithm. Detailed exploration and extensive study were undertaken to address this knowledge gap, which involved:

- Algorithmic Comprehension: Extensive research was conducted to fully grasp the foundational principles and mechanics of the Levenshtein distance algorithm. Understanding the method's standard use in calculating the number of edits (substitutions, insertions, deletions) required to convert one string into another was a crucial first step.

- Non-Recursive Implementation: To enhance the algorithm's functionality for real-time feedback, a non-recursive implementation was developed. This adaptation was critical, as it facilitated the addition of the new capability for generating a list of specific edits needed for pronunciation correction.

7 Conclusion

In summary, the development of the targeted feedback extension for Pronunciation Pal has been a multifaceted and intellectually stimulating journey, marked by significant achievements and insightful learnings. The project was meticulously divided into three main phases: API creation for phoneme transcription, generating meaningful feedback using a modified Levenshtein algorithm, and delivering actionable user feedback. Each phase tackled unique technical challenges, ranging from managing duplicate consonants in the transcription model to adapting the Levenshtein distance algorithm for real-time, phoneme-specific advice. Despite the constraints imposed by confidentiality and HIPAA regulations, rigorous testing through self-assessment and voluntary contributions from friends and family validated the effectiveness and robustness of the system.

The results demonstrated the potential of the feedback engine to provide precise and practical pronunciation guidance, transforming how users interact with and learn from the system. Personal testing, supplemented by diverse volunteer contributions, underscored the system's adaptability and accuracy across different speech patterns. However, several areas for future enhancement were also identified, including the inclusion of consonant substitution feedback, more detailed responses for phoneme insertions and deletions, and the complex challenge of addressing diphthongs. These improvements, along with potential clinical trials and broader dataset inclusion, present a promising roadmap for the continued evolution of the tool.

The insights and achievements garnered from this project form a solid foundation for further advancements. By leveraging interdisciplinary collaboration, algorithmic innovation, and comprehensive testing, Pronunciation Pal can evolve into a more encompassing and effective resource for language learners and individuals with speech impairments. The project's accomplishments thus far highlight the transformative potential of integrating machine learning with linguistic feedback, setting the stage for future developments that could significantly enhance pronunciation training and speech therapy methodologies.

References

Bear, H. L. and Harvey, R. W. (2018). Phoneme-to-viseme mappings: the good, the bad, and the ugly. *CoRR*, abs/1805.02934.

International Phonetic Association (2024). Pulmonic consonants chart. `https://www.internationalphoneticassociation.org/content/ipa-pulmonic-consonants`.

Martinez-Quezada, M. E., Patricia Sánchez-Solís, J., Rivera, G., Florencia, R., and López-Orozco, F. (2022). English mispronunciation detection module using a transformer network integrated into a chatbot. *International Journal of Combinatorial Optimization Problems & Informatics*, 13(2):65 – 75.

meetDeveloper (2024). Free Dictionary API. `https://github.com/meetDeveloper/freeDictionaryAPI`.

Nasim, S. M., AlTameemy, F., Ali, J. M. A., and Sultana, R. (2022). Effectiveness of digital technology tools in teaching pronunciation to saudi efl learners. *FWU Journal of Social Sciences*, 16(3):68 – 82.

Xu, Q., Baevski, A., and Auli, M. (2021). Simple and effective zero-shot cross-lingual phoneme recognition. *CoRR*, abs/2109.11680.