IADIS International Journal on WWW/Internet
Vol. 17, No. 1, pp. 66-78
ISSN: 1645-7641

AI-ENABLED LANGUAGE SPEAKING COACHING FOR DUAL LANGUAGE LEARNERS

Ashutosh Shivakumar, Saurabh Shukla, Miteshkumar Vasoya, Imen M. Kasrani and Yong Pei
SMART Lab, Wright State University, Dayton, Ohio, USA

ABSTRACT

In this research article, we propose a human-AI teaming based mobile language learning solution that provides: 1.) automatic and accurate intelligibility analysis for multiple languages at various levels: sentence, phrase, word and phoneme; 2.) immediate feedback and multimodal coaching on how to correct pronunciation; and, 3.) evidence-based dynamic training curriculum tailored to each individual's learning patterns and needs, e.g., typical pronunciation errors and retention of corrected pronunciation. The use of visible and interactive AI-expert technology capable of intuitive emoji-based interactions will greatly increase student's acceptance and retention of learning with a virtual coach. In school or at home, it will readily resemble an expert reading specialist to effectively guide and assist a student in practicing reading and speaking by him/herself independently, which is particularly important for dual language learners (DLL) whose first language (L1) is not English as many of their parents don't speak English fluently and cannot offer the necessary support. Our human-AI teaming-based solution overcomes the shortfall of conventional computer-based language learning tools and serves as a supportive learning platform that is critical for optimizing the language-learning outcomes.

KEYWORDS

Dual Language Learners, Mobile Learning, Human-AI Teaming, Language Intelligibility Assessment, Mobile Cloud Computing

1. INTRODUCTION

Learning English just like any other language can be equally challenging to dual language learners, both young and adults (Krasnova and Bulgakova, 2014). Dual language learners (DLL) whose first language (L1) is not English need many opportunities to speak and read English (L2) to achieve the English language proficiency needed for academic success, social and emotional competencies. Many schools offer programs during school time that assist such children in developing language proficiency. But those programs may not be enough due to restriction of time and staffing.

In this research, we have proposed a mobile solution – iLeap, enabled by the latest artificial intelligence technologies, such as Machine Learning and Automatic Speech Recognition, that will support DLLs of young age. The iLeap learning tool offers them the option to practice accurate pronunciation with a virtual reading specialist and receive immediate feedback and instruction on how to correct pronunciation even when a native speaker is not available to assist. It will serve as a virtual assistant at school for the reading specialist since these students may require personalized attention which instructor cannot ensure due to limitation of staffing and practice time. Moreover, it helps address their biggest challenge in language learning **-** to extend the language practice and learning in school to home, as many of their parents don't speak English fluently and cannot offer the necessary help at home.

1.1 Survey of Existing Language Learning Applications

High quality apps continuously arrive for both iOS and Android that help DLLs to learn a new language effectively and efficiently. These apps cover almost all languages at little to no cost, and provide the private, virtual, and all-inclusive environment necessary for learning and perfecting a new language through reading, writing and speaking. They can be classified into 5 categories:

1. **Language courses**: Babbel, Duolingo and Busuu are among the most popular applications of language courses. These applications use translation and dictation to emulate traditional language classes. Learners read text and listen to videos, then interpret and answer questions. These apps are also used to help memorize vocabulary. For speaking training, they use the pronunciation of a native speaker for every word and phrase. Unfortunately, this is an unorganized way of learning information because these apps may start with complex words and tricky phrases, providing only a way for improving vocabulary rather than effective methods for enhancing conversational skills.
2. **FlashCards and SRS**: Memrise, Tinycards, and AnkiApp are popular examples of this category. They provide a way of practicing vocabulary using memorization of words and phrases, structured as a competitive game in which users are rewarded with points for every correct answer. It is worth noting that Memrise also has a unique feature that associates a new word with similar words from the user's native language to help make a link between words for better memorization.
3. **Educational games**: MindSnacks is such an educational game that helps users learn grammar and vocabulary and practice listening. In addition, this application teaches words and phrases by limiting the time in which to guess the correct answer. It is more applicable to children than to adults because it uses cartoon image.
4. **Q&A, chat and social**: The most popular chat and social applications used in learning new languages are HelloTalk, HiNative, and TripLingo. They use real-time conversation with unknown native speakers and a text-to-voice option to help pronounce received messages. TripLingo is different from HelloTalk and HiNative in that it provides the learner with information related to the place that he/she wants to travel. HiNative is a chat application that uses question and answer features so that the learner can ask the native about their language and culture. Hence, it is a place for one to introduce themselves more than it is a place to correctly practice a new language.

5. **Contextual reference**: Leaf is one of the contextual reference applications that explains the necessary words that the learner needs to know when encountering new situations.

However, current language training applications are limited in the following aspects:

1. Improve writing more than speaking: Most of the language training applications do not provide an efficient listening or speaking experience. Users can learn some new vocabulary and constructions, but unfortunately cannot carry on a deep conversation with a native speaker of the foreign language. Learning a new language is not only about learning new words and formulating new phrases with appropriate syntax; it is also about being understandable when pronouncing words (Heil, et al, 2016).
2. Lack of performance assessment and feedback: Current applications rarely evaluate speaking skills and language pronunciation quality. To make the learning of foreign language more efficient, applications need to deliver meaningful feedback that evaluates the quality of the user's speech. A successful application for learning new language needs to be able to make a real evaluation of mispronounced speech and recognize an incorrect accent.
3. It is mostly about gaming: Applications that depend more on gaming than the actual fundaments of a language can be problematic in the long-term, as passing levels and scoring becomes more important than learning and practicing the language.

Thus, there is a need of application that could assess the pronunciation of new learners, provide instant feedback on mispronounced words, pinpointing the mistake at the corresponding phonemes, and then be able to provide both audio and visual instructions on how to correct the pronunciation.

1.2 System Features of the Proposed iLEAP Solution

Our primary goal of this research is to support dual language learners for independent language learning with instant feedback and coaching. To achieve this goal, we have identified the following key capabilities and features necessary for supporting effective pronunciation training/learning.

1.2.1 Emphasis on Reading and Pronunciation Skills

The iLEAP system insists on developing the reading and pronunciation skills of the learners in multiple languages. The learners work on various books reading sessions through the app and the system assess their performance in real time. Books are suggested to the learner intelligently based on the profile data. The application leverages speech recognition API provided by Google Cloud Speech services as the Google Cloud Speech services support speech to text transcription of over 120 languages.

1.2.2 Intelligibility Assessment, Feedback and Phoneme Level Correction

The assessment of the performance is done in real-time through mobile-cloud computing. Learner gets to know immediately if he/she mispronounced any word through intuitive user interfaces. For instance, we make use of Android usability features Text highlighting, clickable spans to make the application easy to use. The mispronounced word is compared with original word further at phoneme level. For this work, we have used the set of 39 distinct phonemes from CMU (CMU-Sphinx project). The review at the end of each reading session breakdowns mispronunciations at the word and phoneme levels. When coaching is requested, only the phonemes that diverged on the recognized word from original word are uttered, with help of visual animation that show lip movements required to accurately pronounce that specific phoneme. For instance, if learner pronounce "LIFT" for original word "LEFT",

- Both words will be compared at phoneme level as:

  L EH F T → *L IH F T*
- The server returns mismatching phoneme "EH"
- The app will playback sound for "EH" with corresponding animation followed by utterance of original word "LEFT"

Moreover, our solution extends the phoneme – level intelligibility assessment to multiple languages including, e.g., French and Deutsche (German). This is brought to fruition because of the availability of International Phonetic Alphabet (IPA) based dictionary. The IPA, consisting of Latin alphabet, provides a representation of those qualities of speech that forms a part of the spoken language. Since the IPA is encoded in Unicode it is possible to use them in computers to discern the pronunciation of the text converted from speech.

The accurate analysis of learner speech makes it possible to provide instant feedback on what he/she did not observe otherwise. Instant feedback plays a crucial role in learning. It helps the learner clearly know the adjustment needed. Furthermore, it helps the learner to know whether he/she achieved the goal or not. Evaluation system of language learning may also help the trainer to develop training courses that concentrate better on identified weakness and provide highly personalized learning experience. The feedback of our language learning application provides the advantages of both Constructivist and Behavioristic theories of language learning. The application acts as a virtual facilitator by providing instant feedback that emulates constructivism. Further, it implements behaviorism by identifying errors pertaining to intelligibility and guiding the learner to practice on specific pronunciations (Heil, et al, 2016).

1.2.3 User Profiling and Learning Retention Assessment

The content server in the cloud also maintains user profile for learning patterns. After completion of each session, the app sends performance data (e.g., list of mispronounced words) during that session, which is populated by the server into database. This enables the cloud server to generate different insights into user learning patterns, like most frequent mispronounced words, typical phonemes that the learner may have difficulty to pronounce, retention of learning over time, i.e., whether the learner's pronunciation improved for certain word and phoneme over time. The scope of data collection and server-side capabilities can be conveniently extended as needed due to the use of cloud-based approach, once the basic framework is available. Thus, we may also enhance both app and the server in future for many other insights through user's learning pattern profiling.

2. SYSTEM OVERVIEW

The iLEAP application focuses on the usability of the application, keeping a specific audience in mind: young kids of 4 - 8 age group. Hence the mobile application incorporates simple and intuitive ways to provide performance assessment, feedbacks and coaching on the reading session instantaneously.

2.1 System Architecture

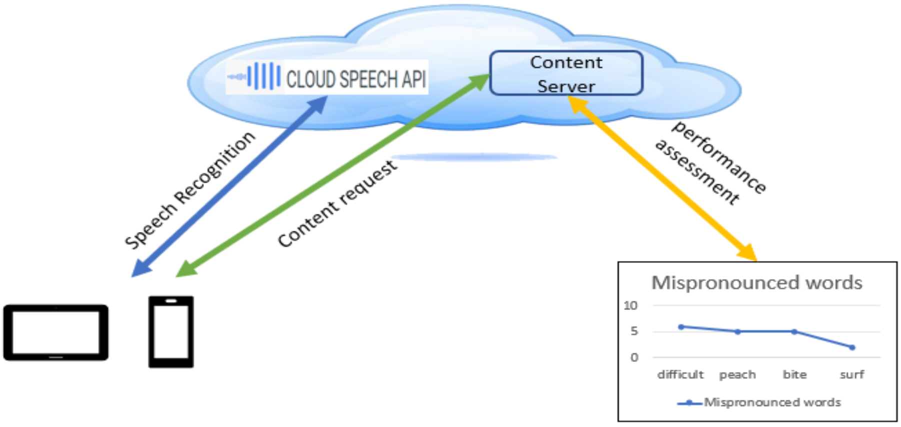


Figure 1. Overview of iLEAP System Architecture

Figure 1 illustrates the iLEAP system architecture. The application can be deployed on any mobile device, e.g., smartphones or tablets. The user is authenticated with content server and then the books that fit the authenticated account profile will be listed on the device. The title selected by the user will then be retrieved from server and the text content is displayed on the device. When learner starts reading the book, speech recognition service captures audio stream and sends the audio data to Google Cloud Speech for recognition. When the recognition result is received from the cloud server, the recognized text is compared with source text from the book for word by word comparison. However, there is a potential challenge of achieving accurate intelligibility assessment using today's deep learning based automatic speech recognition (ASR). These ASR algorithms are built to resolve even ambiguous pronunciation to the right word at the particular context. But, for language intelligibility assessment purpose, this self-correcting feature may reduce the pronunciation errors of the reader. To mitigate this problem, we also look at the confidence score of the recognition at the utterance level to enforce a high standard of correct pronunciation. Then, the learner will be given an instant feedback of his/her intelligibility in speaking the language in terms of highlighted text as the reading progress:

- Green highlight indicates the word pronunciation is accurate
- Yellow highlight indicates the word is mispronounced

These mis-pronounced words will be reviewed and rehearsed when the session ends. The content server also provides retention tracking. All the mispronounced words are updated in the database for learner's profile. This data can be used to analyze and profile the learner and evaluate the user performance. The analytics provides insights like words that the learner persistently fails to read, or individual phoneme in different words that the student experiences most difficulty in pronouncing accurately. It may also provide pattern of retention in the learning; whether the learner improved on certain word and phoneme that he/she had mispronounced earlier.

2.2 System components and enabling technologies

2.2.1 Speech Recognition

iLEAP uses Google Cloud Speech Streaming API to recognize AUDIO input. Streaming API enables it to perform speech recognition of continuous audio stream in real-time. Google cloud services provides gRPC stub for Android/Java platform. We implement speech recognition service using the gRPC stub APIs. For accounting purpose, the gRPC client stub needs authentication token to validate the account for the use of speech recognition service. Currently this service is available worldwide at $1.44 per hour, which is significantly lower when compared to hiring a personal language coach or tutor. This cost could eventually be eliminated with the arrival of more matured offline ASRs, e.g., the planned Google Offline ASR(Coldewey 2019).

2.2.2 Content and Profiling Server

The contents for reading session are dynamically retrieved from the content server that is deployed on cloud for 24/7 availability. In our project, Amazon cloud service is used and the Content server is implemented in Flask/Python with MySQL as backend database. This server provides RESTful APIs such that android app will be able to request reading content, request phoneme level comparison of words, update user profile in MySQL database for mispronounced words or get analytics on user profile for reading patterns.

2.2.3 User Interface

The user interface of the prototype is the most critical part of any learning apps designed for children at young age, it must be as simple as possible with the intention to avoid distraction due to unnecessarily complicated operations. Thus, in iLeap, most of the interactions are through intuitive components, such as buttons, layouts and views carry symbols that handily describe the objective of the interface. On completion of a reading session, it automatically summarizes all the mispronounced words from the session along with phoneme level intelligibility feedback, such that the app utters only individual phoneme that was mis-pronounced in case of homonyms. The coaching system simultaneously highlights correct way of lip gestures required to pronounce the phoneme accurately using visual animations through Emoji or Animoji.

2.2.4 Intelligibility Assessment

Speech intelligibility assessment is a complex process that may vary significantly from one human evaluator to another. In this research, we propose and adopt a more objective assessment methodology by determining the intelligibility based on outcome of speech recognition (Liu, et al, 2006). Following speech recognition, the assessment process is completed by an accurate comparison between speech-recognized spoken text and the original text. For instance, we need to compare the two texts to find the incorrect words that the learner spoke. Then, based on the result from the comparison, the learner will be given feedback of his/her intelligibility in speaking the language.

To identify the similarity/dissimilarity between two texts, we need to measure the distance between them. This can be achieved using various minimum distance finding algorithm, such as Levenshtein Distance, Hamming Distance, Longest Common Substring Distance and Jaro-Winkler Distance (Cohen, Ravikumar and Fienberg, 2003). In this research, we compare the recognized spoken text and the original text word-by-word using the Levenshtein algorithm. It calculates the minimum numbers of change, including deletion (Missed), insertion (Removed), and substitutions (Replaced), required to transform one string to the other. The time complexity of this algorithm is O (n*m), where n and m are the lengths of the two sentences being compared. The memory space complexity is O (n*m) because it memorizes in matrix. This could be a concerning factor considering we have to compare the sentence incrementally every time with speech recognized text if the sentence is uttered in multiple parts with pauses. However, it becomes less a concern nowadays as most of today's mobile devices can provide enough computing power and memory space for its operation, even for long sentences.

Table 1. Assessment through Levenshtein Algorithm

		five	**little**	**monkey**	**jumping**	**the**	**bad**
	0	1	2	3	4	5	6
five	1	0	1	2	3	4	5
little	2	1	↖ 0	1	2	3	4
monkeys	3	2	1	↖ 1	2	3	4
jumping	4	3	2	2	↖ 1	2	3
on	5	4	3	3	↑ 2	2	3
the	6	5	4	4	3	↖ 2	3
bed	7	6	5	5	4	3	↖ 3

We can extend this intelligibility assessment to other languages and at a phoneme level this is shown in section 2.2.5.

In Table 1, we illustrate the comparison between 2 sentences using the Levenshtein algorithm. For instance, the comparison between "five little monkeys jumping on the bed" and "five little monkey jumping the bad" identify the mismatch of words "monkeys/monkey" and "bed/bad", and also the missing of word "on".

2.2.5 Intelligibility assessment at phoneme level

We illustrate the phoneme level intelligibility assessment in Figure 2. The Levenshtein algorithm is again applied to the sequences of phonemes obtained from the CMU library for the actual word and the word returned by the ASR, thus identify the mispronounced phonemes, e.g., "eh" instead of "ah" shown in this example.

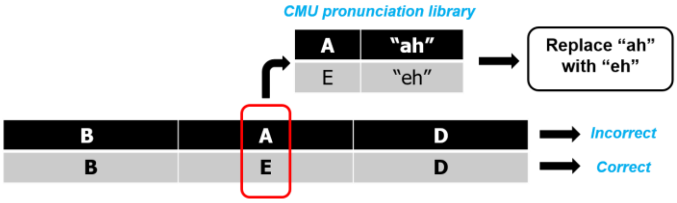


Figure 2. Phoneme-level Intelligibility assessment for English

We also illustrate the phoneme level intelligibility assessment for Deutsche in Table 2. We compare the pronunciation of the word “nicht” in Deutsche meaning “no”. We chose this word because there is a higher tendency of non-German speaking population to mispronounce “ch” as “k” or “sh”(Sonia, 2017) .

Table 2. Phoneme-level Intelligibility assessment

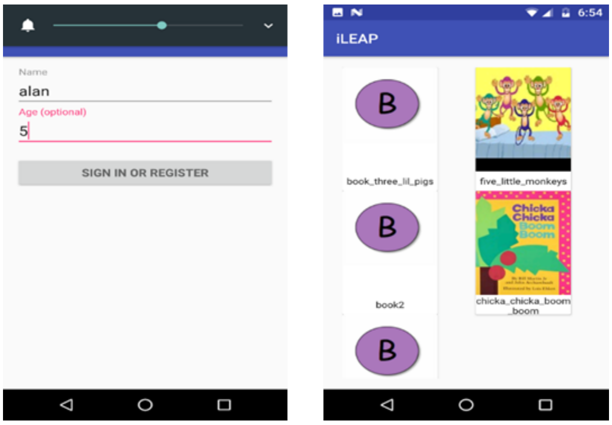
		n	I	ç	t
	0	1	2	3	4
n	1	0	1	2	3
i	2	1	0	1	2
k	3	2	1	1	2
t	4	3	2	2	1

3. EXPERIMENTAL RESULTS

The prototype has been developed to illustrate and evaluate the effectiveness of the mobile app enabled language learning. The following results validate our approach.

3.1 User Interface

Once the app is launched on the device, user lands on login page as shown in Figure 3. The authentication process verifies user profile on the backend server. After authenticating the learner, the application lists book titles that are relevant to learner’s profile. The profile level is derived at the server side based on learner’s age and how he progresses through various reading sessions. Server maintains books in generic hierarchical structure so that random titles can be displayed to the learner to expose them to new content/vocabulary and avoid repetition.



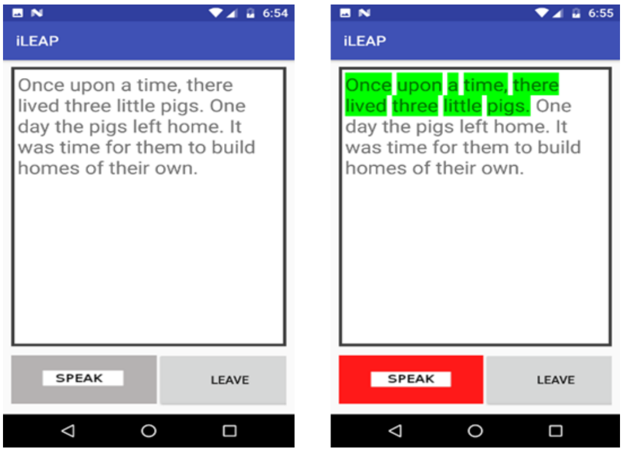
Name
alan

Age (optional)
5

Figure 3. Launch the App

3.2 Reading Progress with Accurate Pronunciation

When a book is selected by learner for reading, the contents are displayed as plain text. Once the audio recording is enabled with a button click, speech recognition results are matched with the original text in the background and text is instantly highlighted with appropriate color spans. As illustrated in Figure 4, if the recognized text matches with source text, the green background span highlights the portion of matched text.



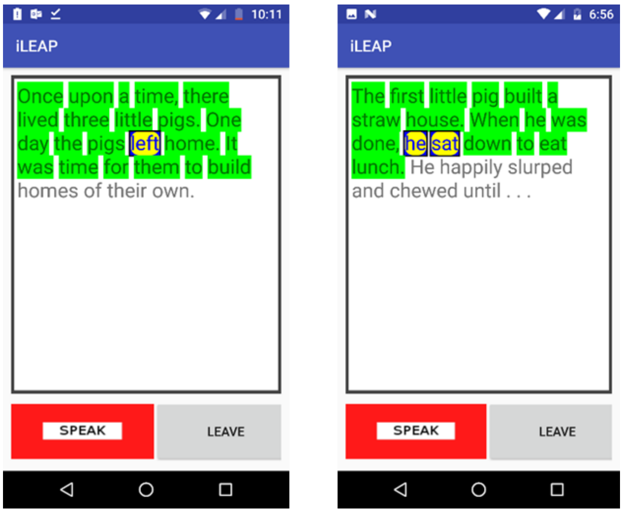
Once upon a time, there lived three little pigs. One day the pigs left home. It was time for them to build homes of their own.

Once upon a time, there lived three little pigs. One day the pigs left home. It was time for them to build homes of their own.

Figure 4. Reading progress without errors

3.3 Reading Progress with Dissimilarity Detection

If any word is mispronounced during the session, intelligibility assessment algorithm returns dissimilarity with original text. This dissimilarity is highlighted with yellow background on original text. The highlight also enables clickable interface on the word so that learner can click on the word to hear out correct pronunciation of the word using Android Text-To-Speech API. As illustrated in the Figure 5, when “left” was mispronounced as “lift”, the intelligibility assessment detects the mismatch between recognized text and the text is highlighted accordingly.



Once upon a time, there lived three little pigs. One day the pigs left home. It was time for them to build homes of their own.

The first little pig built a straw house. When he was done, he sat down to eat lunch. He happily slurped and chewed until . . .

Figure 5. Reading progress errors

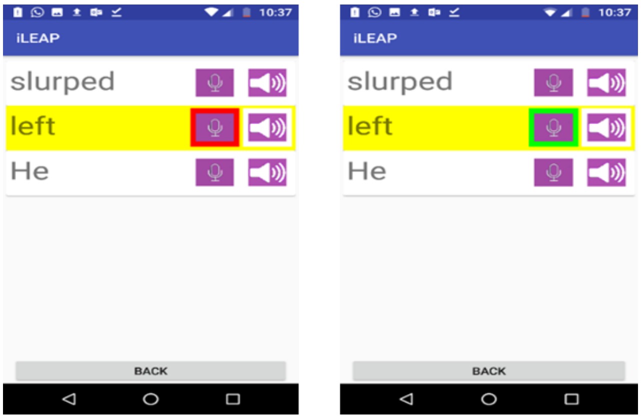


Figure 6. Correction Practice

3.4 Session Review and Correction Coaching

At the end of the session, all dissimilar words are displayed for practice as shown in Figure 6. The dissimilarity is mapped at phoneme level such animation shows lip movement for the missing phoneme. As shown in the figure, learner pronounced “lift” for “left”, the missing phoneme was identified as “EH”. The animation mimics lip movements to pronounce “EH”, along with Text-To-Speech utterance of the phoneme and entire word. The learner can practice again with the word that he/she failed to pronounce properly as shown in Figure 7. The mic button interface enables speech recognizer to accept audio input for speech-to-text translation. Intelligibility assessment feedback for the re-attempted word is also available in terms of background color of the mic button.

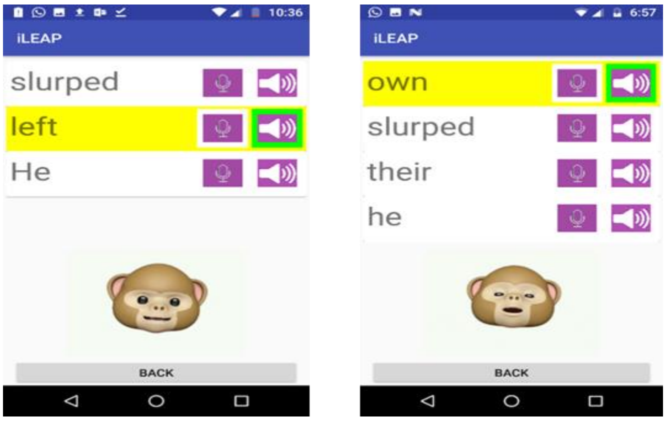


Figure 7. Correction Coaching

3.5 Analysis of Retention based on Learner’s Profile Data

The backend server implements a comprehensive database to store profile data for each student. The tables retain information such as frequency count of mispronounced words, frequency count of phonemes that found to be mismatching in recognized words. The analysis results can be displayed to show the student’s typical pronunciation errors at word and phoneme levels as illustrated in Figure 8 and Figure 9 (“The CMU Pronouncing Dictionary”). It can assist the classroom learning by providing the accurate and comprehensive list of assessment data to instructors. It is also used as evidence by iLEAP to automatically build dynamic training curriculum tailored to everyone’s learning patterns and needs based on his/her typical pronunciation errors, e.g., by recommending books that have the same words or words with the same phonemes.

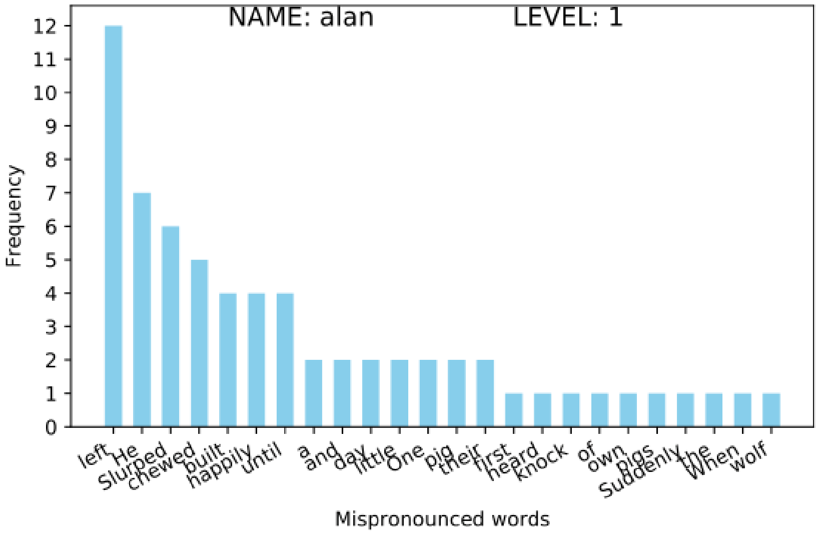


Figure 8. Performance analysis of frequency count of mispronounced words

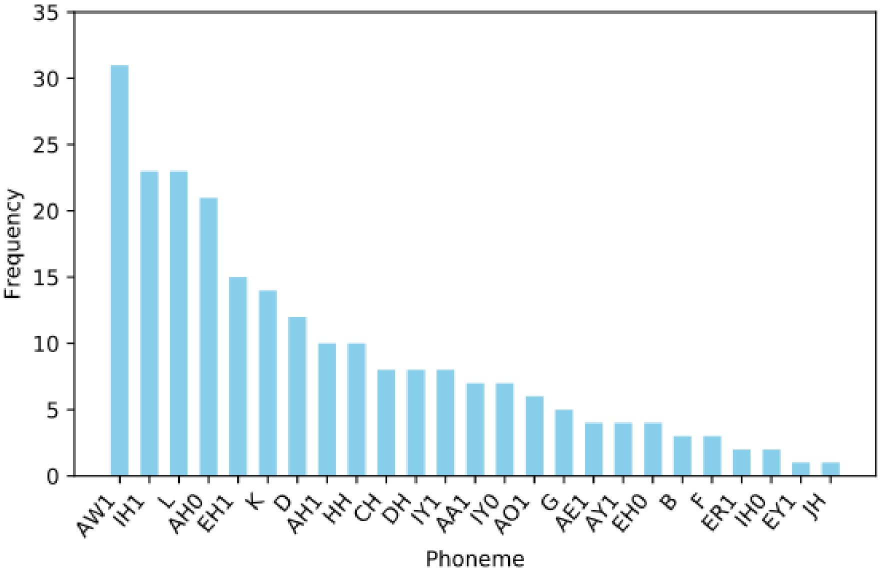


Figure 9. Performance analysis of frequency count of mispronounced words

Furthermore, for individual word, iLEAP can also find pattern of retention, which can provide evidence that learner improved on the word over time as illustrated in Figure 10.

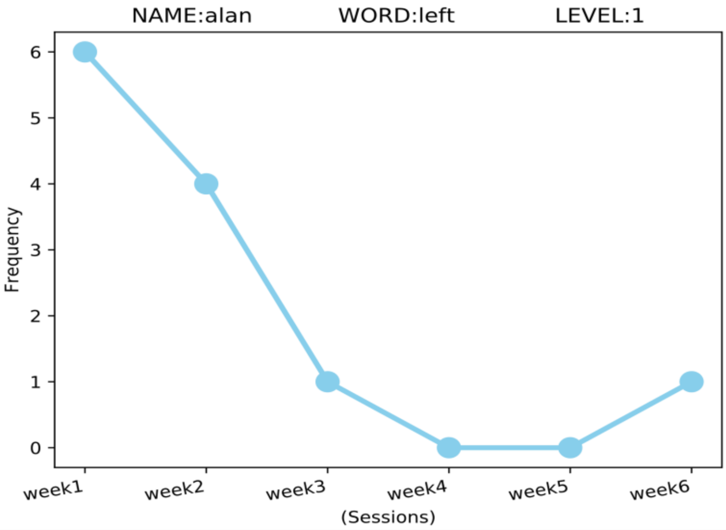


Figure 10. Performance tracking of retention of corrected pronunciation

4. CONCLUSIONS AND FUTURE WORKS

The prototype iLEAP solution confirms that advanced technologies in speech recognition, AI and mobile cloud computing can be leveraged to build a learning system for dual language learners. The system can provide a low cost, highly available and personalized tutoring with focus on reading and pronunciation skills of a learner who is attempting to learn English. Our experimental results demonstrate that the system is not only capable of providing immediate intelligibility assessment, but also tracking the learner's experience, which in long term can aid in improving the retention of the learning.

Even though the current system capabilities of iLEAP prototype are limited in terms of analyzing an individual's typical and atypical learning patterns, moving forward in future we could enhance backend system with No-SQL server, implement better analytics and profiling code that can generate a more detailed insight on learner's performance and trends in retention capabilities. Depending of those patterns, the system may better recommend a specific book that contains contents with a balance of learning new words and the retention of corrected words in a more engaging and supportive learning environment for young dual language learners.

Moreover, a randomized controlled trial (RCT), a group-randomized trial (GRT), is being planned to study and establish the effectiveness of iLeap as a pronunciation training/learning tool. During the RCT, half of the students will receive access to the iLeap training tool in conjunction with usual instruction (Experimental Group), while the other half of the students will receive instructions in the usual and customary manner, without the benefit of the iLeap training tool (Control Group). Upon successful completion of the trial, the iLeap-based training/learning solution could provide students more opportunities for practice and help them stay more engaged in their learning efforts to ensure the achievement of language competencies.

REFERENCES

CMU-Sphinx project: http://www.speech.cs.cmu.edu/, and https://cmusphinx.github.io/wiki/tutorial

"The CMU Pronouncing Dictionary." Accessed July 9, 2019. http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

Coldewey, Devin. 2019. "Google's New Voice Recognition System Works Instantly and Offline (If You Have a Pixel)." *TechCrunch* (blog). March 12, 2019. http://social.techcrunch.com/2019/03/12/googles-new-voice-recognition-system-works-instantly-and-offline-if-you-have-a-pixel/.

Google Cloud Speech. Available at: https://cloud.google.com/speech/.

Heil, C. R., Wu, J. S., Lee, J. J., & Schmidt, T. (2016). A Review of Mobile Language Learning Applications: Trends, Challenges, and Opportunities. *The EuroCALL Review*, *24*(2), 32–50.

Krasnova E., Bulgakova E. (2014) The Use of Speech Technology in Computer Assisted Language Learning Systems. In: Ronzhin A., Potapova R., Delic V. (eds) Speech and Computer. SPECOM 2014. Lecture Notes in Computer Science, vol 8773. Springer, Cham, Switzerland.

Liu, W. M., Jellyman, K. A., Mason, J. S. D., & Evans, N. W. D. (2006). Assessment of Objective Quality Measures for Speech Intelligibility Estimation. In 2006 IEEE ICASSP. https://doi.org/10.1109/ICASSP.2006.1660248

Neri, A., Cucchiarini, C. and Strik, H. (2003) Automatic speech recognition for second language learning: How and why it actually works. Speech Communication.

W. Cohen, W, Ravikumar, P. and E. Fienberg, S. (2003). A Comparison of String Metrics for Matching Names and Records. Proc of the KDD Workshop on Data Cleaning and Object Consolidation.