DEVELOPMENT OF A HUMAN-AI TEAMING BASED MOBILE LANGUAGE LEARNING SOLUTION FOR DUAL LANGUAGE LEARNERS IN EARLY AND SPECIAL EDUCATIONS

A Thesis submitted in partial fulfillment of the
requirements for the degree of
Master of Science

by

SAURABH SHUKLA

Bach., University of Pune, 2005

2018

Wright State University

WRIGHT STATE UNIVERSITY

GRADUATE SCHOOL

December 06, 2018

I HEREBY RECOMMEND THAT THE THESIS PREPARED UNDER MY SUPERVISION BY Saurabh Shukla ENTITLED Development of a Human-AI Teaming Based Mobile Language Learning Solution for Dual Language Learners in Early and Special Educations BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF Master of Science.

_____

Yong Pei, Ph. D.
Thesis Director

_____

Mateen M. Rizki, Ph. D.
Chair, Department of Computer Science and Engineering

Committee on Final Examination:

_____

Yong Pei, Ph. D.

_____

Mateen M. Rizki, Ph. D.

_____

Anna F. Lyon, Ed. D.

_____

Barry Milligan, Ph.D.
Interim Dean of the Graduate School

ABSTRACT

Shukla, Saurabh. M.S. Department of Computer Science and Engineering, Wright State University, 2018. Development of a human-AI teaming based mobile language learning solution for dual language learners in early and special educations

Learning English as a secondary language is often an overwhelming challenge for dual language learners (DLLs), whose first language (L1) is not English, especially for children in early education (PreK-3 age group). These early DLLs need to devote a considerable amount of time learning to speak and read the language, in order to gain the language proficiency to function and compete in the classroom. Fear of embarrassment when mispronouncing words in front of others may drive them to remain silent; effectively hampering their participation in the class and overall curricular growth.

The process of learning a new language can benefit greatly from the latest computing technologies, such as mobile computing, augmented reality and artificial intelligence. This research focuses on developing a human-AI teaming based mobile learning system for early DLLs. The objective is to provide a supportive and interactive platform for them to develop English reading and pronunciation skills through individual attention and interactive coaching.

In this thesis, we present an AR and AI-based mobile learning tool that provides: 1) automatic and accurate intelligibility analysis at various levels: letter, word, phrase and sentences, 2) immediate feedback and multimodal coaching on how to correct pronunciation, and 3) evidence-based dynamic training curriculum tailored for personalized learning patterns and needs, e.g., retention of corrected pronunciation and typical pronunciation errors. The use of visible and interactive virtual expert technology capable of intuitive AR-based interactions will greatly increase a student's acceptance and retention of 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 DLL as many of their parents don't speak English fluently and cannot offer the necessary help.

Ultimately, our human-AI teaming solution overcomes the shortfall of conventional computer-based language learning tools and serves as a supportive and team-based learning platform that is critical for optimizing the learning outcomes.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would first like to thank my thesis advisor, Dr. Yong Pei, for extending me the opportunity to work on this research. He always encouraged and guided me whenever I ran into a trouble spot or had a question about my research or writing. He consistently allowed this research to be my own work, but steered me in the right direction, stimulating innovative approaches in the thesis work whenever he thought I needed it.

I would also like to thank Dr. Anna F. Lyon for her valuable inputs. Her experience in the field of early education helped in understanding the problem that this research intends to solve. I also want to express my sincere gratitude to Dr. Mateen M. Rizki and Dr. Anna F. Lyon who were involved in the validation committee for this thesis.

Special thanks to my colleagues, Ashutosh Shivakumar, and Miteshkumar Vasoya, for their useful debate and exchanges of knowledge that enriched this experience.

Finally, I must express my very profound gratitude to my family especially to my spouse and kid, for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

1. INTRODUCTION

Learning English just like any other language can be equally challenging to dual language learners, both young and adults. 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 the restriction of time and staffing.

In this research, we have developed 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. This research is a continuation of a previous research project entitled “Development of a performance assessment system for language learning” (1). This research emphasizes developing assessment and tutoring system with second language learners in early education. 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 the school for the reading specialist since these students may require personalized attention which instructor cannot ensure due to the limitation of staffing and practice time. Moreover, it

helps address them the 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. Background of study

There are many learning apps already available, either web-based or on a mobile platform, for Dual Language Learners that provide personalized language training. Some of these applications use Flashcards, animation-based games (e.g. match words with pictures) to keep the learners engaged (2). They motivate the kids to memorize the vocabulary, but they hardly help in developing communication skills. Some popular applications like Babbel, Duolingo use translation and dictation to emulate traditional language classes. Learners read the text, listen to videos and then interpret and answer questions. But most of these applications have a focus on improving vocabulary and writing skills than speaking and accurate pronunciation. Below is the survey of the features provided by few of the popular commercial English learning applications for kids.

1.1.1. Learning through topic-based images and Audible books

Lingokids provide audiobooks with images, sounds, and animation to keep the kids engaged. It provides topic-based interactive content. Lingokids features teaching material from Oxford University Press that uses common but effective techniques like audio songs, worksheets, games. They provide a way of practicing vocabulary using memorization of words and phrases, structured as an interactive game (3). This increases the active participation of kids in learning new words and at the same time keep the kids motivated to do more.

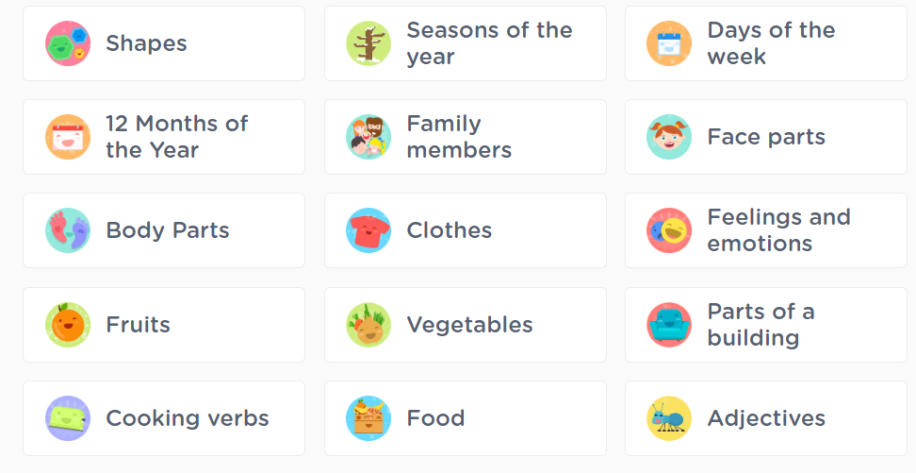

Figure 1.1: Lingokids topic-based content

1.1.2. Flash cards, interactive games, and Storytelling applications

Gus on the go is a popular iOS app that is available in 30 different languages for early language training. It narrates stories in visual images and animations. The flash cards and "match words with images" kind of puzzle games help increase vocabulary. The learner observes an image of a real-world object, reads its spelling and hears its pronunciation to associate the word with the image.

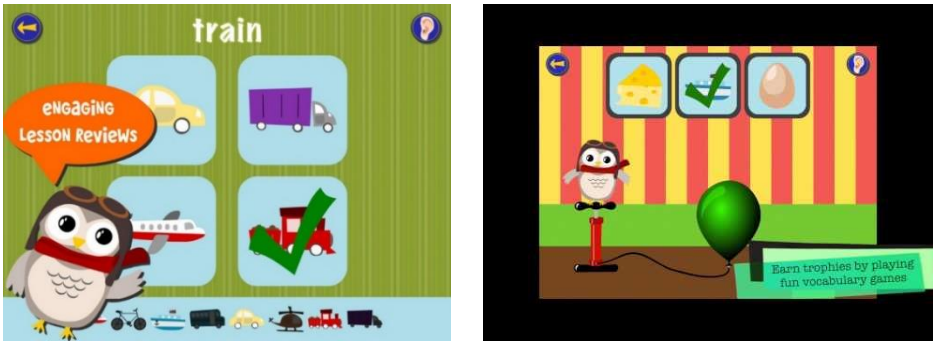

Figure 1.2: Gus on the go stories and flashcards

MindSnack is another app that provides interactive games and audio-visual interfaces for learning new English words and increase the vocabulary. The app has different channels to work on improving different aspects of language skill like grammar, spelling, vocabulary, connotation, synonyms.

Figure 1.3: Mindsnacks interactive games

The assessment of the performance is purely based on the capabilities of the learner to select the right card or interact with the game system accurately. Such apps do not have any means to assess the speaking abilities of the learner.

1.1.3. Online reading comprehension-based learning

Raz-kids is a web-based reading comprehension platform. It provides online reading content with repository comprising of hundreds of eBooks created to ensure text complexity and quality content. The books are categorized in 29 different levels

of reading difficulty. Students can access their leveled text through an interactive learning portal designed to keep them motivated and engaged. Every eBook is available online as well as in mobile formats and allows students to listen to, read at their own pace, and record themselves reading (4). The web-based content delivery platform provides interactive, kid-friendly environment. The audible books provide content highlighting as the reading session progress. Each session comes with a quiz interface to evaluate comprehension. Students qualify for a level upgrade depending on completing these quizzes. It provides teacher management portals that enable the instructors to monitor the progress of each individual student.

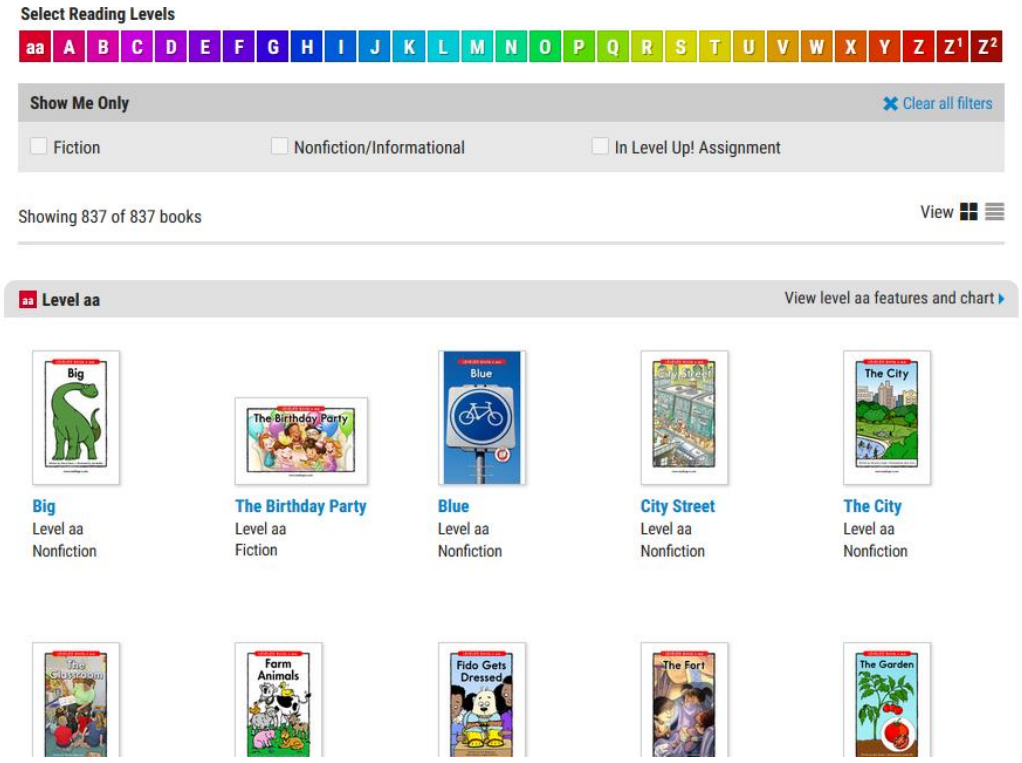

Figure 1.4: Raz-kids leveled book inventory

Although, this platform mostly focus on developing reading skills passively. Students are provided with interactive tools that they can use to take notes while listening to a reading session. They can listen to the accurate way of pronouncing individual word within a book content, but there is no real time evaluation of student's pronunciation skill.

1.2. Contribution of the Thesis

These applications use interactive features to maintain kids' attention. The studies suggest that words associated with actual objects or imagery techniques are learned more easily than those without. With multimedia applications, it is possible to provide, in addition to traditional definitions of words, different types of information, such as pictures and videos (5). However, most of these applications do not provide any kind of intelligibility and performance assessment of the learners' ability to grasp the language skills, especially the reading and pronunciation skills. Flashcards, audible storybooks, and games help the kids in increasing and retaining vocabulary. They improve writing more than speaking. But these commercial applications rarely provide means to assess speaking skills and quality of pronunciation which is critical for student's practices.

Learning a new language is not only about learning new words and formulating new phrases but it is also about learning how to accurately pronounce them. There is a need of an application that could also 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. This thesis research attempts to implement a mobile solution using advanced technologies like speech recognition, natural language processing and cloud computing that could aid dual language learners in developing English speaking and pronunciation skills. The solution will provide instant intelligibility assessment of the pronunciation.

1.3. Thesis Organization

The rest of this thesis is organized as follows: in chapter 2 we provide insight into how of the speech recognition systems work. Further, we review the efficiency and feasibility of various available automatic speech recognition services. In chapter 3 we describe various features in our proposed solution, system architecture, and functional overview. In chapter 4 we present experimental results. In chapter 5 we conclude on the thesis work and provide potential future work to further enhance the solution.

2. LITERATURE REVIEW

2.1. Speech recognition

Speech is a complex phenomenon. Speech structure as it is understood in current practice is a continuous audio stream where stable states mix with dynamically changed states (6). Phoneme classes are distinctly identified in these states. English words can be considered as a composition of several phonemes that have distinct sounds. In other words, phonemes are keywords to represent the word based on how it sounds or how it is pronounced.

In written English grammar, each word is composed of 5 vowels and 21 consonants. We can write an English word using these 26 distinct alphabetic representations. Similarly, we can pronounce any English word using phonemes. In general, NLP researchers have identified 40 distinct phonemes for English language comprising a symbol set called ARPAbet symbol set. This symbol set or a subset of it is commonly used in most of the Speech Recognition systems. For our project, we refer the CMU sphinx phoneme set that is derived from the ARPAbet symbol set. Table 2.1 shows these phonemes and the contextual word in which they are used as an example (7).

Phoneme	Example	Translation	Phoneme	Example	Translation
AA	odd	AA D	L	lee	L IY
AE	at	AE T	M	me	M IY
AH	hut	HH AH T	N	knee	N IY
AO	ought	AO T	NG	ping	P IH NG
AW	cow	K AW	OW	oat	OW T
AY	hide	HH AY D	OY	toy	T OY
B	be	B IY	P	pee	P IY
CH	cheese	CH IY Z	R	read	R IY D
D	dee	D IY	S	sea	S IY
EH	Ed	EH D	T	tea	T IY
ER	hurt	HH ER T	TH	theta	TH EY T AH
EY	ate	EY T	UH	hood	HH UH D
F	fee	F IY	UW	two	T UW
G	green	G R IY N	V	vee	V IY
HH	he	HH IY	W	we	W IY
IH	it	IH T	Y	yield	Y IY L D
IY	eat	IY T	Z	zee	Z IY
JH	gee	JH IY	ZH	seizure	S IY ZH ER
K	key	K IY			

Table 2.1: Phonemes from CMU Sphinx and their sample usage

The ASR engines work on AUDIO speech inputs that contain the pronunciation of the words to extract out phonemes from above phoneme set.

2.1.1. Components of Speech Recognition Engines

Speech recognition engines are made of following components –

- **Decoder**: This is a software component that process sound inputs and searches acoustic model for similar sounds. If a match is found, it determines phoneme for the corresponding sound piece. The decoder continues identifying phonemes for the sounds until it finds a pause in the input, which denotes the end of the word. At this stage, the series of

identified phonemes are looked up in the language model for corresponding best possible word (phoneme to word reverse lookup). A highest possible match is returned as a recognized word for the sound input.

- **Feature Extraction**: Analog to Digital Converter is used to remove noise from the input sound wave and digitize the input. The digital signal is broken into simple phoneme sounds using Fourier transformation.
- **Acoustic model**: It models the relationship between the audio signal and phonetic units in the language. This component tries to derive a potential phoneme based on digital audio samples. It uses complicated algorithm trained on speech corpus which is large speech data set to build a statistical representation of each phoneme. This representation is called the Hidden Markov Model. Each phoneme has associated HMM (8) (9).
- **Language model**: is used to predict the probability of a subsequent word based on the previously recognized word. Statistical language models use N-grams to probabilistically determine the word spoken based on context. Language models are also used to control search space. It defines the way search is performed.
- **Phonetic dictionary**: contains a mapping from words to phonemes. Each word is represented as a composition of the phonemes, e.g., *LEFT* → *L EH F T*. The dictionary is not the only method for mapping words to phones. You could also use some complex function learned with a machine learning algorithm (6).

2.1.2. Speech Recognition workflow

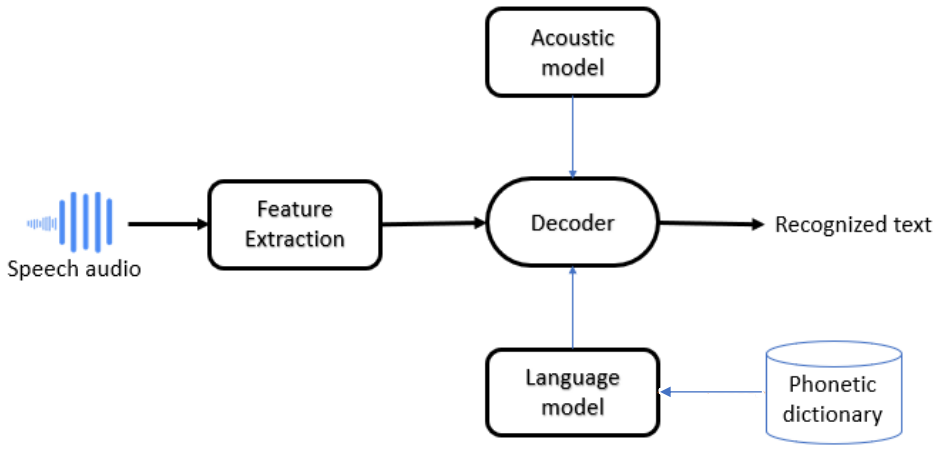

Figure 2.1: Workflow of the Speech recognition engine

- Speech is recorded by voice recorder in form of analog signal. This analog signal is converted by ADC into a digital signal that is sampled at 16kHz sampling rate.
- Fourier transformation breaks the complex signal into simple sound waves.
- These sound waves are then matched through the acoustic model to identify phonemes with HMM. The sequence of phonemes stacked up until the system identifies a pause in the input signal. The pause signifies completion of the word.
- On completion of the word, the sequence of identified phonemes is reverse looked up in phoneme dictionary for corresponding English word that may have been spoken. The language model also maintains the context of

previously identified words/phonemes in the N-gram model to determine the probability of word with a high likelihood that fits in the context.

- The decoder returns the sentence in text format as final output.

2.2. Automatic Speech Recognition services

There are many competitive speech recognition services available commercially. Most of them provide developer APIs on the cloud platform for high availability. Below is a list of few such leading speech-to-text service providers (10).

2.2.1. Google Cloud Speech API

Google is one of the leading inventors in Natural Language Processing and Speech Recognition. Their flagship "Google assistant" is currently an integral part of all latest android systems. Commercial products like Google home that are built using the natural language processing engine; being used to enable smart homes/equipment to be operated on voice commands. Google also provides an interface to their speech recognition engines as a hosted cloud service. These interfaces are available in the form of client libraries, command line, REST APIs or gRPC APIs. Depending on the need of the application, one can use simple REST API to send an HTTP request to Google Speech server with raw audio file as payload or make use of more complex but robust gRPC to avail features like streaming audio recognition.

Currently, Google Speech API supports 120 languages and variants. It can auto-detect language, auto-transcribe the punctuations and filters inappropriate contents in text results, supports real-time streaming audio input, speaker detection, and

many more to come. Their pricing model is reasonable. For audio speech-to-text API, first 60 minutes of audio processing is free per month. After 60 minutes, the pricing is $ 0.006/15 seconds.(i.e. $ 0.024 per minute) (11).

2.2.2. Amazon Transcribe

The Amazon Transcribe API can be used to analyze audio files stored in Amazon S3 and have the service return a text file of the transcribed speech (12). We can also send a live audio stream to Amazon Transcribe and receive a stream of transcripts in real time. Amazon Transcribe can be used for lots of common applications, including the transcription of customer service calls and generating subtitles on audio and video content. The service can transcribe audio files stored in common formats, like WAV and MP3, with timestamps for every word so that you can easily locate the audio in the original source by searching for the text. The pricing model of Amazon Transcribe is at a rate of $0.0004 per second. Usage is billed in one-second increments, with a minimum per request charge of 15 seconds.

2.2.3. Bing Speech API

Microsoft has moved the Bing Speech API under their cognitive services portfolio. The cloud-based Microsoft Bing Speech API provides developers with an easy way to create powerful speech-enabled features in their applications, like voice command control, user dialog using natural speech conversation, and speech transcription and dictation. The Microsoft Speech API supports both Speech-to-

Text and Text-to-Speech conversion. One can even customize the speech recognition models for unique vocabularies or accents (13).

2.2.4. IBM Watson speech to text

IBM Watson Speech to Text recognizes speech to text in real time with good accuracy. In order to have a high accuracy, it uses AI to combine language information with audio composition, accepting many types of audio formats with file sizes less than 100 MB. In addition, it returns partial recognition results when it becomes available. Compared to others API's, it supports a significantly lower number of languages; to date, it supports only eight different languages and the service needs to be paid for.

2.2.5. Android speech to text

Starting from API level 3, Android has inbuilt support for speech to text. Android provides classes and interfaces that can be extended in the application to use the built-in speech recognizer. The speech recognizer works in both online and offline modes. Thus, the applications that use Android speech recognizers do not require persistent internet connection. However, the accuracy of Android speech-to-text is not high due to the limitation of resources of the mobile device.

2.2.6. CMU Sphinx SR toolkit

The speech recognition toolkit is a result of research efforts at CMU. It can work on standalone hardware with low configurations(like mobile platforms).

PocketSphinx toolkit is designed to work in offline mode. It uses prebuilt statistical models for speech recognition. It supports multiple languages and can also build a model for any new language with enough speech corpus for training. It is available under a BSD-like license which enables developers to build and redistribute apps without any restrictions (6).

2.2.7. Kaldi

Kaldi is a speech recognition toolkit, freely available under the Apache License. It supports linear transforms, MMI, boosted MMI and MCE discriminative training, feature-space discriminative training, and deep neural network. It can transcribe most audio/video formats. The recognition models can infer punctuations, capitalization, and speaker segmentation. It can transcribe A 10-minute audio file in under 3 minutes. The pricing comes at 0.035/min (14).

2.3. Limitations of Speech Recognition systems

2.3.1. Ambient noise

The accuracy of the speech recognition engines heavily depends on the source input. If the quality of the input audio signal is not good, the resulting text output may be inaccurate or unreliable. Inferior microphones, noisy surroundings(streets, public places, and supermarkets, subway stations), large rooms with echo effects are typical sources of ambient noise where the ASR may not perform well.

2.3.2. Multiple Speakers and accents

Multiple overlapping sounds may impair the accuracy. Besides each English phoneme has associated HMM value based on speech corpus that the acoustic model is trained on. If the speech corpus contains a specific accent, then the system may not be able to accurately infer phonemes from a different accent that acoustic model is not exposed to.

2.3.3. Resource intensive computation

The language models and acoustic models of the speech recognition systems are built on statistical modeling systems and N-grams models. These predictive components need significant computational and memory resources in order to produce accurate speech recognition. Standard hardware might not be enough. Hence most of the commercial speech recognition systems are hosted in the cloud and provide services via client interfaces Abstracting out all the necessary resources.

2.3.4. Homonyms

In general, homonyms are words that sound similar, may or may not have a similar composition of phonemes/spelling but carry a different meaning. Linguistically Homonyms, homographs, and homophones fall under homonymy. Homonyms are words that strictly have the same spelling and pronunciation but carry a different meaning. Homographs do not have the restriction of same pronunciation.

Homophones, on the other hand, may have different spelling but have the same pronunciation. For instance, “To”, “Two” and “Too” are homophones. “Bow”(as in bow and arrow) and “Bow”(as in bow down) is an example of homograph. Speech recognition engines may translate such homonyms accurately because of the language models which maintain context. However, there is still scope of some error.

3. SYSTEM OVERVIEW

The iLEAP application focuses on the usability of the application, keeping a specific audience in mind: young kids of Pre-K – 3 age group. Hence, the solution leverages highly efficient, proven and tested solutions that offer speech recognition services for natural language processing and speech recognition. The mobile solution incorporates simple and intuitive ways to provide performance assessment on the reading session instantaneously.

3.1. System Features

The primary goal of this project is to support dual language learners for independent language learning. To achieve this goal, we identify the following key capabilities and features necessary for supporting effective pronunciation training/learning.

3.1.1. Emphasis on reading and pronunciation skills

The iLEAP system insists on developing the reading and pronunciation skills of the learners. 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 solution uses text-to-speech API provided by Android to show the correct way to pronounce the words. The learner

will be able to playback the words that he failed to pronounce correctly and practice on improving his/her pronunciation.

3.1.2. Intelligibility assessment, feedback and phoneme level correction

The assessment of the performance is done in real-time. The learner gets to know immediately if he/she mispronounced any word. We make use of android usability features Text highlighting, clickable spans to make the application easy and intuitive to use. The mispronounced word is compared with the original word further at phoneme level. On summary view, when word playback is requested, only the phonemes that diverged on the recognized word from original word are uttered, with help of visual animation that shows 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 play-back sound for "EH" with the corresponding animation followed by the utterance of the original word "LEFT"

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. The evaluation system of language learning may also help the trainer to develop

training courses that concentrate better on identified weakness and provide a 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 (15).

3.1.3. User profiling and learning retention assessment

The content server in the cloud also maintains user profile. After completion of each session, the app sends performance data (e.g., list of mispronounced words) during that session, which is updated by the server in a database. This enables the cloud server to generate different insights into the user profile, 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. The scope of data collection and server-side capabilities can be conveniently extended as needed due to the use of a cloud-based approach, once the basic framework is available. Thus, we may also enhance both the app and the server in future for many other insights in the user profile.

As described above, the system incorporates features like reading comprehension, vocabulary, interactive animation to keep learner motivated. In addition, it provides instant feedback on the performance which makes it portable, a one-to-one learning

platform for kids without human assistance. The learner can also use it beyond school hours to continue learning and improving the reading and pronunciation skills.

3.2. System Architecture

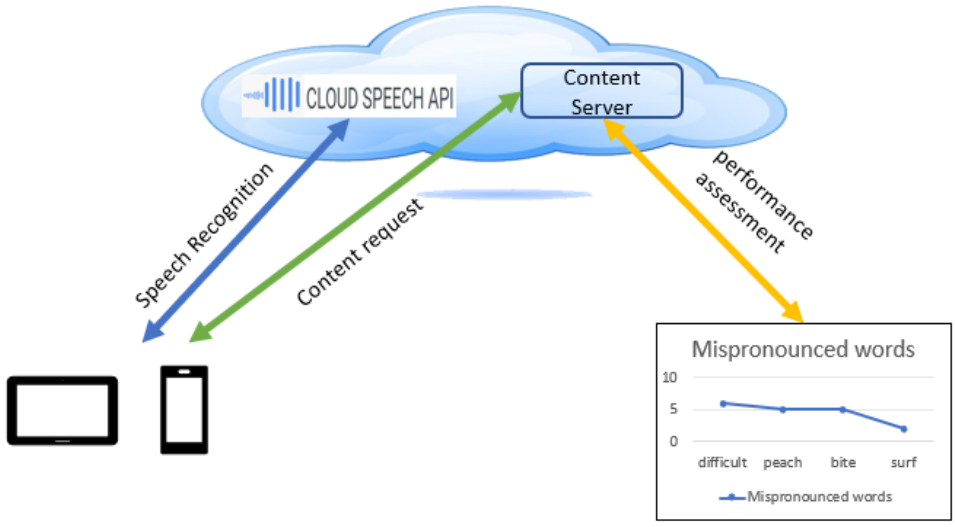

Figure 3.1: Overview of iLEAP System Architecture

Figure 3.1 illustrates the iLEAP system architecture. The application can be deployed on a mobile device or tablet. The application uses Google cloud speech API for speech-to-text translation and connects with a cloud-based content server for data requests like a list of book titles, the content of the book etc. Hence it needs a persistent internet connection. Below is a typical workflow of the app.

- Step 1: User is authenticated with the content server.
- Step 2: The book titles that match the authenticated account profile will be listed on the device.

- Step 3: User selects a book from the list. The title selected by the user will be retrieved from the server and the text content is displayed on the device.
- Step 4: User clicks “speak” button to start reading the rendered text.
- Step 5: Application enables the microphone on the device to record the audio signal. The captured audio data stream is sent to Google cloud Server using gRPC streaming recognition API.
- Step 6: When the recognition result is received from the cloud server, word-level comparison is down between the recognized text and source text using minimum distance finding algorithm
- Step 7: Feedback on learner’s intelligibility in speaking the language is instantly made available in terms of highlighted text as the reading progress –
    - Green highlight indicates the word pronunciation was accurate
    - Yellow highlight indicates the word was mispronounced
- Step 8: On completion of the reading session, aggregated distinct mispronounced words will be sent to the content server with a unique session id. The app also sends incorrect words from the recognized text that did not match with the original text.

Content server provides the REST API to compare two words at the phoneme level. Each incorrect word from the recognized text is compared with the original correct word and the REST API returns only mismatching phonemes. The mispronounced words can be 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 perform analytics on the learner’s profile

and evaluate the user performance. The analytics may provide insights like words that learner persistently fails to read or individual phoneme in different words that the student face most difficulty in pronouncing accurately. It may also provide a pattern of retention in the learning; whether the learner improved on a certain word that he/she faced difficulty in the beginning.

3.3. Components and enabling technologies

3.3.1. Speech Recognition service

iLEAP uses Google Cloud Speech Streaming API for automatic speech recognition. Streaming API enables it to perform speech recognition of continuous audio stream in real-time. Google cloud services provide gRPC (16) 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 in order to validate the account for the use of speech recognition service. As of now, this service is available worldwide at $1.44 per hour, which is significantly lower when compared to hiring a personal language coach or tutor.

3.3.2. Content and profiling server

The contents for the reading session are retrieved from the content server on the go. The content server nothing but a RESTful web server that can be deployed on Ubuntu Linux. The content server can be a dedicated hardware with internet connectivity or can be hosted on the cloud for 24/7 availability. In our project, we

use The Elastic Cloud instance deployed on Amazon Web Service(AWS) cloud. The experimental content server is implemented in Flask/Python with MySQL as the backend database. Flask is web microframework for python[ref]. It is very easy to deploy and very flexible at defining RESTful APIs on a web server. At present, the server bears minimal functionality to support the iLEAP android app. This server provides RESTful APIs such that android app will be able to request reading content, request phoneme level comparison of two words, update the user profile in MySQL database for mispronounced words or get analytics on the user profile for reading patterns.

3.3.3. User interface

The user interface of the prototype is the most critical part of any learning apps designed for children at young age, it need 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 system utters only individual phoneme that was mispronounced in case of homonyms. The coaching system simultaneously highlights a correct way of lip gestures required to pronounce the phoneme accurately using visual animations through Emoji or Animoji.

All the required resources like animation GIFs and phoneme pronunciation audio files are packaged with the Android app. Android can leverage third-party JARs like glide to render the GIF. We also extend default audio player interface to synchronize the individual phoneme audio playback with corresponding GIF rendering.

3.3.4. 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 the outcome of speech recognition (17). 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 the two texts, we need to measure the distance between them. This can be achieved using various minimum distance finding algorithms, such as Levenshtein Distance, Hamming Distance, Longest Common Substring Distance and Jaro-Winkler Distance (18). 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 the 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.

In Table 3.1, 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" computes similarity score of 71.

		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

Table 3.1: Assessment through Levenshtein Algorithm

4. EXPERIMENTAL RESULTS

A fully-functional prototype has been developed to illustrate and evaluate the effectiveness of the mobile app-enabled language learning. The following results provide validation for our approach.

4.1. User Interface

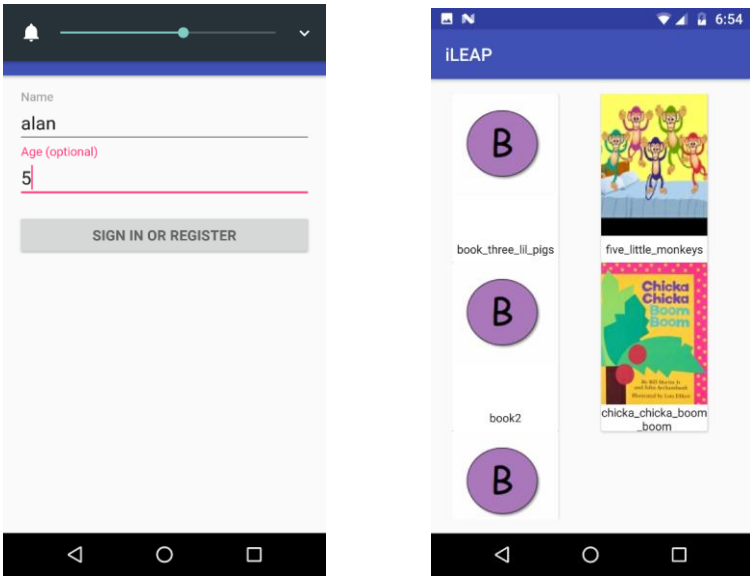

Figure 4.1: Launch the App

As illustrated in Figure 4.1, once the app is launched on the mobile device, the user lands on the login page as shown. The authentication process verifies the 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. The server maintains books in a generic hierarchical structure so that random titles can be

displayed to the learner in order to expose them to new content/vocabulary and avoid repetition.

4.2. Reading progress with accurate pronunciation

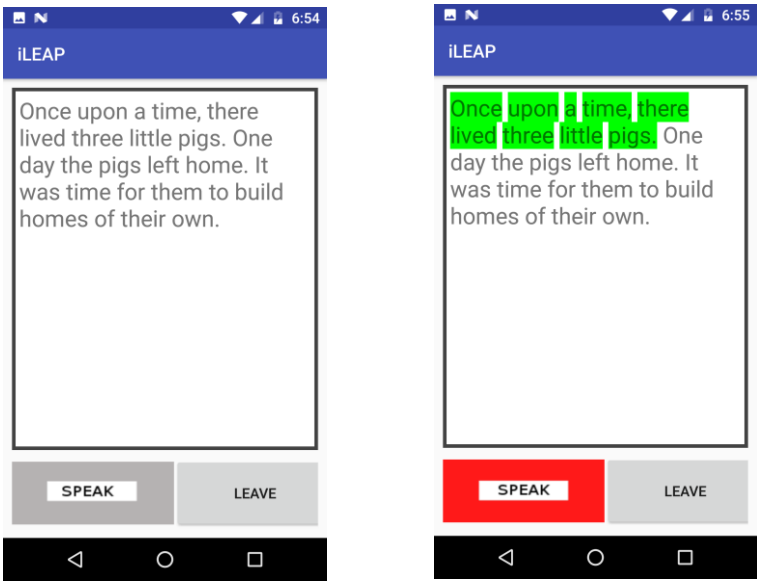

Figure 4.2: Reading progress without errors

When a book is selected by the 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 4.2, if the recognized text matches with the source text, the green background span highlights the portion of the matched text.

4.3. Reading progress with dissimilarity detection

If any word is mispronounced during the session, intelligibility assessment algorithm returns dissimilarity with the original text. This dissimilarity is highlighted with a yellow background on the original text. The highlight also enables clickable interface on the word. The learner can click on the word to hear out the correct pronunciation of the word using Android Text-To-Speech API. The text-to-speech will use the voice model(male or female) currently set by default on the device. However, the voice model can be changed from device configuration for voice with the desired accent.

As illustrated in Figure 4.3, when “left” was mispronounced as “lift”, the intelligibility assessment detects the mismatch between recognized text and the text is highlighted accordingly.

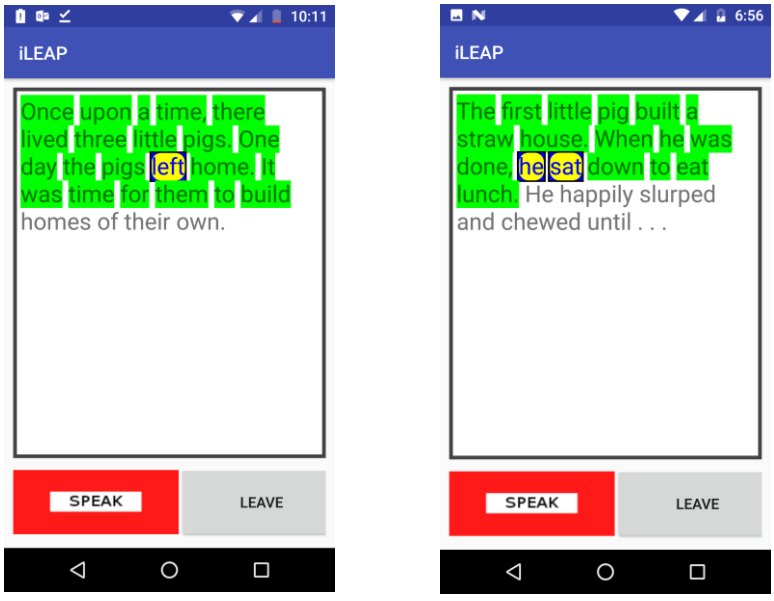

Figure 4.3: Reading progress with errors

4.4. Session review and correction coaching

At the end of the session, all dissimilar words are displayed for practice as shown in Figure 4.4. The dissimilarity is mapped at phoneme level by the content server and only phonemes that found to be missing will be returned to the app. The app contains resource files of all animation gifs and sounds corresponding to each of 39 phonemes. The GIFs are cartoon animations that illustrate exact lip gestures required to pronounce certain phoneme accurately. When the playback button is clicked for a word in the summary list, the corresponding animation shows lip movement for the missing phoneme. Corresponding audio for the individual phoneme will all be uttered in the background along with GIF. This produces an effect that animation is uttering the phoneme. As shown in the figure, learner pronounced “lift” for “left”, the missing phoneme was identified as “EH”. The Emoji animation mimics lip movements to pronounce “EH”, along with Text-To-Speech utterance of the phoneme and entire word to give multimodal coaching to the learners for enhanced learning outcome. The learner can repeat the practice with the word that he/she failed to pronounce properly.

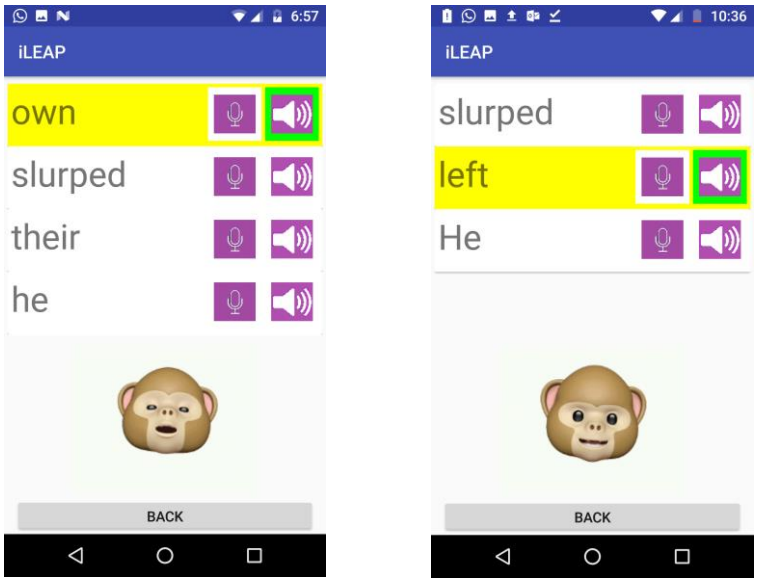

Figure 4.4: review and correction coaching

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 the background color of the mic button.

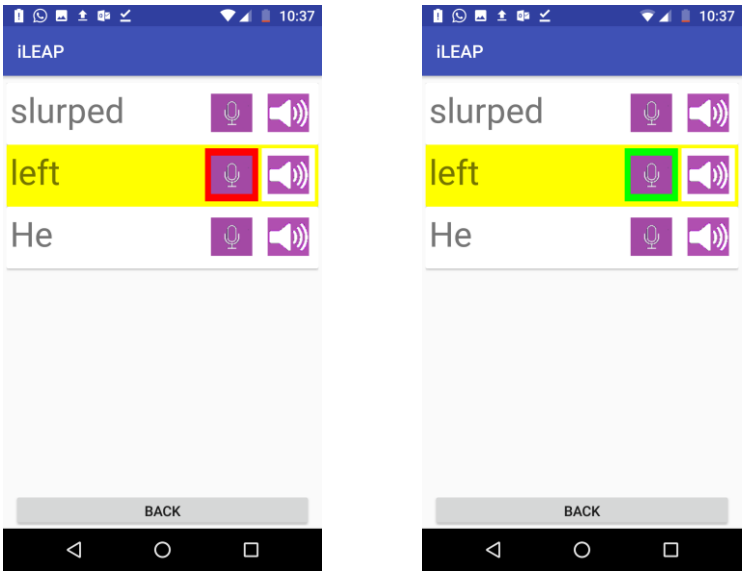

Figure 4.5: Correction Practices

4.5. 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, a frequency count of phonemes that found to be diverging 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 4.6. 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 every individual'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.

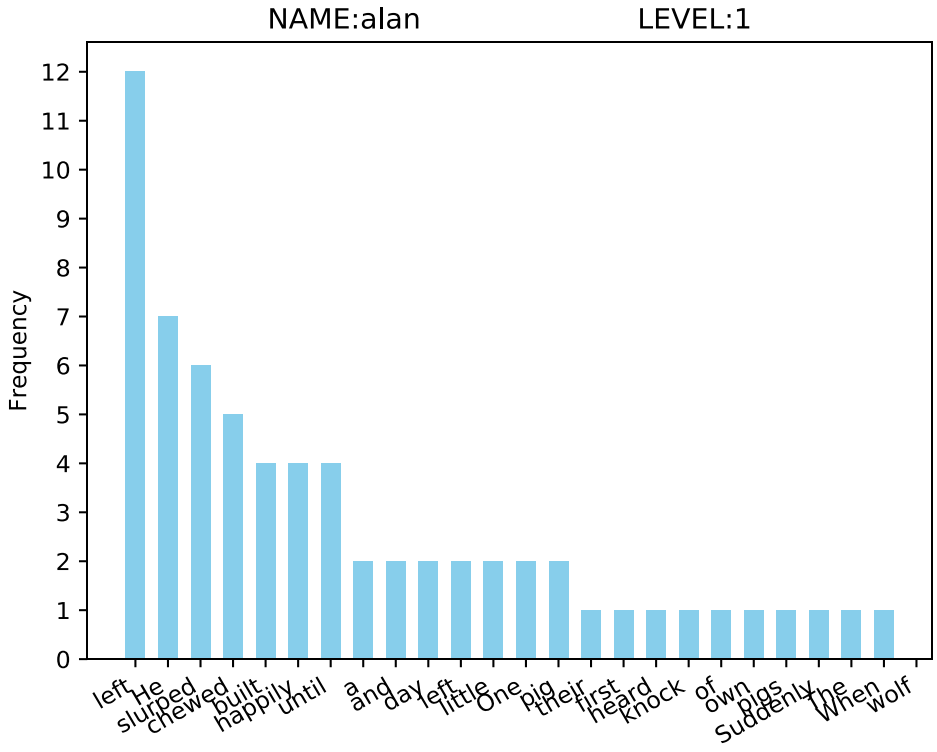

Figure 4.6: Performance analysis, e.g., a frequency count of mispronounced words

Furthermore, for an individual word, iLEAP can also find a pattern of retention, which can provide evidence that learner improved on the word over time as shown in Figure 13.

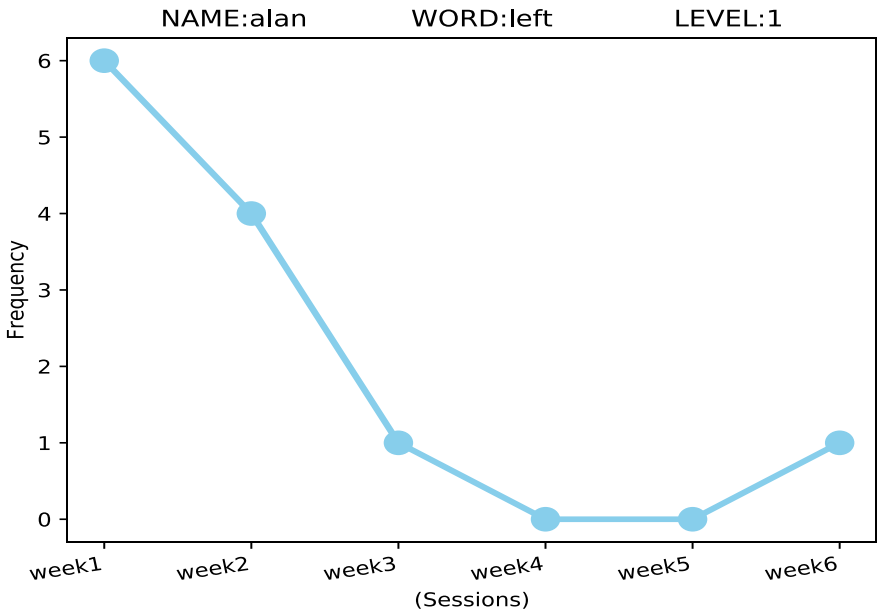

Figure 4.7: Performance tracking, e.g., retention of corrected pronunciation

5. CONCLUSION

The prototype iLEAP solution confirms that advanced technologies in speech recognition, AI and AR and mobile cloud computing can be leveraged to build a human-AI teaming-based learning system for dual language learners. The system can provide a low cost, highly available and personalized tutoring with a focus on reading and pronunciation skills for 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 the 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 on 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.

6. REFERENCES

1. kasrani, Imen. Development of a Performance Assessment System for Language Learning, Master of Science (MS), Wright State University, Computer Science and Engineering. [Online] 2017. http://rave.ohiolink.edu/etdc/view?acc_num=wright1515958862300186.

2. Tassinari, Tyler. online. [Online] https://elearningindustry.com/10-best-language-learning-apps-for-kids.

3. Lingokids. [Online] https://www.lingokids.com/english-for-kids.

4. [Online] raz-kids. https://www.raz-kids.com.

5. *Effects of Multimedia Annotations on Vocabulary Acquisition.* PLASS, DOROTHY M. CHUN JAN L. 2, 1996, The modern language journal, Vol. 80.

6. Basic concepts of speech recognition. [Online] https://cmusphinx.github.io/wiki/tutorialconcepts.

7. CMUdict. [Online] http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

8. [Online] http://www.voxforge.org/home/docs/faq/faq/what-is-an-acoustic-model.

9. Wikipedia acoustic model. [Online] https://en.wikipedia.org/wiki/Acoustic_model.

10. quora speech to text. [Online] https://www.quora.com/What-are-the-top-ten-speech-recognition-APIs.

11. Google speech-to-text. [Online] https://cloud.google.com/speech-to-text/.

12. [Online] Amazon transcribe. https://aws.amazon.com/transcribe/.

13. [Online] azure cognitive services. https://azure.microsoft.com/en-us/services/cognitive-services/speech-to-text/.

14. Kaldi speech. [Online] http://kaldi-asr.org/.

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

16. [Online] gRPC. https://grpc.io/.

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

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