Title	Speech Shadowing Support System in Language Learning
Author(s)	Lee, Carson
Citation	
Issue Date	2017-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/14166
Rights	
Description	Supervisor:長谷川　忍, 情報科学研究科, 修士

Speech Shadowing Support System in Language Learning

LEE Carson

School of Information Science
Japan Advanced Institute of Science and Technology

March 2017

Master's Thesis

Speech Shadowing Support System in Language Learning

S1510058 LEE Carson

Supervisor :	Hasegawa Shinobu
Main Examiner:	Iida Hiroyuki
Examiner :	Shirai Kiyoaki

School of Information Science
Japan Advanced Institute of Science and Technology

February 2017

Table of Contents

Abstract 1

1 Introduction 3

1.1 Motivation 3

1.2 Purpose of Thesis 4

1.3 Structure of Thesis 5

2. Background 6

2.1 Speech Shadowing 6

2.1.1 Traditional Speech Shadowing 6

2.2 Learning Model 7

3. System Design 9

3.1 Speech Shadowing System 10

3.1.1 Determining the difficulty of a speech 10

3.1.2 System Platform 11

3.1.3 User Management 12

3.1.4 Types of Scaffolding 13

3.2 Performance Evaluation Metrics 14

3.3 Evaluation Algorithms – Determining user level 16

4. System Development 18

4.1 System Requirements 18

4.2 Development Environment 18

4.3 External Libraries 19

4.4 User Interface 19

5. Case Study..24

5.1 Purpose of Case Study 24

5.2 Type of data to be gathered 24

5.3 Preliminary Evaluation..25

5.4 Control vs System User...26

5.5 Results 27

5.6 Result Analysis..29

5.5.1 Control Group 29

5.5.2 Test Group...31

5.7 Feedback Questionnaire 35

5.8 Discussion..36

6. Conclusion 37

6.1 Future Work 37

Acknowledgement 39

Publication List 40

References...41

Appendix A – Testing Manual / Instruction 43

Pre-test...43

Control Group 45

Test Group...45

Post Test...45

Appendix B - Speech Shadowing Language Learning Questionnaire 46

List of figures

Figure 1 Traditional Speech Shadowing Session setup 7
Figure 2 Phases in the Cognitive Apprenticeship Learning Process................8
Figure 3 Speech Shadowing System Use Case Diagram 9
Figure 4 Scaffolding 1,3, and 4 being used in a shadowing session..............13
Figure 5 Visualization of intonation difference 14
Figure 6 Visualisation of tempo difference 15
Figure 7 User’s Main Screen..20
Figure 8 Speech Selection Screen 20
Figure 9 Transcript Scaffolding Being Used..21
Figure 10 Mock-up of the two waveform after shadowing.............................22
Figure 11 Self-Evaluation Screen..23
Figure 12 Student 2 Performance Graph...30
Figure 13 Student 3 Performance Graph...30
Figure 14 Student 4 Performance Graph...31
Figure 15 Student 7 Performance Graph...31
Figure 16 Student 1 Performance Graph...32
Figure 17 Student 5 Performance Graph...33
Figure 18 Student 6 Performance Graph...33
Figure 19 Student 6 Performance Graph...34

List of tables

Table 1 User Level and Speech Difficulty Level Matching 13
Table 2 Variables 16
Table 3 Speech Shadowing System Minimum Requirements............................18
Table 4 Preliminary Evaluation Score 25
Table 5 Performance based on self-eval 26

Table 6 Group Members 26

Table 7 Session 1 Control Group Score 27

Table 8 Session 1 Test Group Score 28

Table 9 Session 2 Control Group Score 28

Table 10 Session 2 Test Group Score 28

Table 11 Session 3 Control Group Score 28

Table 12 Session 3 Test Group Score 29

Table 13 Questionnaire Results 35

Abstract

One trait that sets humans apart from other species is that our communication between species are much more advanced than any other on Earth. However, there are still no agreed upon theory as to how our language and communication evolved due to the lack of evidence. Regardless, humans begin to develop their verbal communication skills very early in life. The development of speech production throughout an individual's life starts from an infant's first babble and is transformed into fully developed speech by the age of five. It is a type of cognitive skill, and thus, we cannot teach it the same way we would teach sciences or history, as cognitive skill learning is the learning of a skill or knowledge that is hard to symbolize. From the moment we are born, we begin to cry as a means of communication. In a few more months we begin to babble, and soon after we form our first words. In just a short few words we begin to produce longer sentence as we attempt to express ourselves deeper. We can observe that a human’s first method of communication prior to obtaining the knowledge of language is the verbal method. We can see the importance of verbal communication in a language just from that observation. Despite of that, the way we are taught the verbal component of a language varies wildly. The way the language is taught differs from class to class, and even bigger difference in instructional methods can be seen on the larger scale. Various environmental and cultural factors also affect the way the language is taught and imparted. While the grammatical rules governing the structure of a sentence are usually strictly adhered to by most teachers and students, the same does not apply to the oral component. Even amongst countries that speak the English language natively, there exist a great variation in the accent. When speakers with a vastly different language pick up the English language, their English speech may sound very different from the original. For example, the Japanese language is largely monotonous, but English speeches can be greatly affected by the intonation. This makes the English spoken by most Japanese very hard to be comprehended by other English speakers as the Japanese English lacks the intonation context. These differences in accents results in miscommunication even when communicating using the same language. In addition, there exist many consonants and vowels that are mutually exclusive in both language, and thus, a student who learns to speak English via furigana often end up having a hard time to be understood by non-Japanese English. For example, a Japanese would often pronounce “eight” as “ei-to (エイト), fight as “figh-to (ファイト)”, or “the” as “za (ザ). Verbal communication is a major part of a language, but there are not many systems/solutions in the market that caters to self-learning of spoken language. A simple survey on the market place would show that the most popular language learning software focuses on the vocabulary and grammatical aspect. A few of these software contains modules where individual word pronunciations are evaluated as well. However, as mentioned earlier, the intonation of words in an English sentence can affect the meaning, and thus, improving single word pronunciation is inadequate for fluency in language. Speech recognition software can help with pronunciation of longer sentences, but it still does not take into account the intonation. One of the teaching methods that can resolve this problem is Speech Shadowing. It is an experimental technique where a subject repeats speech immediately after hearing it. The process is guided by an instructor who will evaluate the shadowed speech and provide feedback on how to make improvements. However, it is a time consuming method as it requires 1-on-1 tutoring and thus it is not suitable for a large class.

One way we can solve this problem is by applying technology. If we can replace the instructor in terms of evaluating the shadowed speech and provide feedback for improvement, we can greatly increase the adoption rate of this learning method. Depending on the exact technology and technique applied, the system might even reduce training time required. In this paper, we present our approach to utilizing this method for a self-supported learning system and how to utilize technology to improve its efficiency over traditional speech shadowing methods. Using the Cognitive Apprenticeship Model, we describe how the system supports the user in learning via Speech Shadowing. We also explain how the system provide contents and how the system sorts the contents according their difficulty levels. The system would run on a mobile platform to ensure maximum flexibly in self-learning as the user would be able to learn using the system wherever their smartphone/tablet goes. The user management system would also ensure that users are given an approximation of their current progress so that they can self-motivate and also not attempt speeches that are too far above their proficiency level and in turn get discouraged. The system also provides several forms of scaffolding (support) to help weaker users improve. The system uses self-evaluation and it provides adequate support for users to accurately self-evaluate. This is in the form of audio waveforms. By comparing certain traits in 2 audio waveform, users can abstract valuable information from the raw data. An algorithm was defined to evaluate the user's proficiency level based on the data gathered by the system during the self-evaluation process. This data is also used by the system to determine the types of support needed by the user should he/she attempt speeches at levels that are different that his current level. The paper also describes the system development process on the Android platform and examples of a standard use-case flow with the system interface. At the end, a case study was conducted to validate the effectiveness of the system described. We find that the implemented system can provide coaching similar to that of a human instructor. In the final section of the paper, future works and improvements on the system is describe.

1 Introduction

1.1 Motivation

One of the topic of scholarly discussion for several centuries is the origin of language in the human species. Yet till this day, there is no consensus on the actual origin or age of the human language. This is due to the lack of direct evidence. However, what is known is that communication between humans are much more advance than any other species on Earth, and amidst our means of communication, verbal communication stands as one of the most important one.

The development of speech production throughout an individual's life starts from an infant's first babble and is transformed into fully developed speech by the age of five [1]. It is a type of cognitive skill, and thus, we cannot teach it the same way we would teach sciences or history, as cognitive skill learning is the learning of a skill or knowledge that is hard to symbolize.

Today, the English language is the de-facto lingua franca. Despite the widespread usage of English, there exists many variations of the English dialects, such as British English, Cockney English, American English, Engrish (generally refers to poor Japanese influenced English), Manglish (Malaysian English), and many more. The more formal dialects such as British English and American English are often used as the standard for major English proficiency test such as IELTS and TOEFL. Other dialects have evolved from their original one often due to cultural and environmental influences. For example, Manglish is a result of assimilating the many languages spoken in Malaysia into the English language. Another example would be Japanese English, where students often learn English the aid of furigana. There exist many consonants and vowels that are mutually exclusive in both language, and thus, a student who learns to speak English via furigana often end up having a hard time to be understood by non-Japanese

English [2]. For example, a Japanese would often pronounce “eight” as “ei-to (エイト), fight as “figh-to (ファイト)”, or “the” as “za (ザ)”.

Another reason for doing this research is to reduce miscommunication due to different accents/dialects. Looking at the aviation industry, we can observe that many accidents have resulted from communication error. The nuances of a language can be complicated and the same word can carry multiple meanings. Depending on how it is delivered, the message conveyed might vary [3].

Furthermore, in this digital age, information can be disseminated very quickly through the internet and thus many people can spend their downtime (riding on a bus/train, waiting in line, etc.) to absorb more information via their mobile devices. This allows people to learn almost anywhere and anytime. However, some domains are not as easy to be learnt without the presence of an instructor or teacher. There are many applications that cater to language learning. However, the amount of smartphone applications that focuses on improving a learner’s speaking skill is also very limited. Most of these applications focuses on the reading/writing aspect, and the speaking aspect is usually very simple (such as pronunciation of a single word at a time). In teaching a student to speak a foreign language, most attention is devoted to the correct pronunciation of sounds and isolated words. Generally speaking, much less attention is paid to a correct production of intonation [4].

1.2 Purpose of Thesis

In this research, the aim is to solve the problems such as miscommunications due to accents and the high cost of traditional Speech Shadowing. We propose to use the cognitive apprenticeship theory as a basis for the learning support system. By capitalizing on the advancement of hardware and software capabilities of modern smartphones, the system would also be developed on the Android platform. Furthermore, we propose to evaluate and analyse the learner’s performance by using components of a speech. The scope of this research will cover the basics of the traditional Speech Shadowing method, the learning model used, the usage of technology to replace the traditional model, the development of a system, and a case-study to determine the effectiveness of a system designed to improve a user’s speaking skill in the English language via Speech Shadowing.

1.3 Structure of Thesis

This thesis is organised into 6 chapters. In chapter 1, the motivation and purpose of this thesis is described. In Section 2, the background of the research is elaborated. In Section 3, we will describe our approach to solving the problems described earlier, and their algorithms. Section 4 will detail the system development while Section 5 covers the case study purposes and results. Finally, in Section 6 will be the conclusion to this paper, summarizing it, and describing future work that can be done to make the system more efficient.

2. Background

2.1 Speech Shadowing

One way to improve a user’s speaking ability is via Speech Shadowing. Speech shadowing is an experimental technique where a subject repeats speech immediately after hearing it, usually through headphones to reduce noise and/or speech jamming. The reaction time between hearing a word and pronouncing it can be as short as 254ms or even 150ms [5]. While a person is only asked to repeat words, they also automatically process their syntax and semantics. Words repeated during the shadowing practice imitate the parlance of the overheard words more than the same words read aloud by that subject.

We can also observe a similar behaviour in children as they begin to develop their speaking ability. They are often predisposed to imitate/shadow words and speech as a way to guide themselves to enter their cultural community [6]. Since children utilize this method to learn a language, it could be possible to utilize the same method for adults. In fact, learning the patterns of intonation is thought to take place unconsciously by mere imitation. That is, by listening to, and repeating model utterances the foreign-language learner has to acquire a proper intonation.

2.1.1 Traditional Speech Shadowing

In the traditional speech shadowing method, an instructor is needed to sit there to evaluate the student performing speech shadowing. Fig. 1 illustrates the usual steps for a speech shadowing session and they are as follows:

1. Playback of a speech/conversation recording

2. Student performs speech shadowing (repeats the heard speech with minimal delay as clearly and loudly as possible)
3. Instructor listens to the shadowed speech and provides evaluation/feedback to the student
4. The student attempts to improve based on the given feedback and retries the process on a later date.

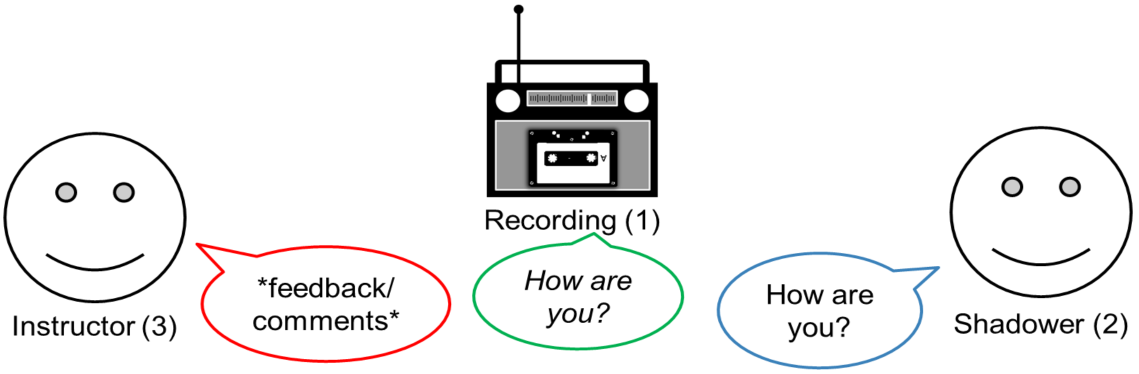


Figure 1 Traditional Speech Shadowing Session setup

2.2 Learning Model

The learning model used in this research would be the Cognitive Apprenticeship Theory. It is the process where a master of a skill teaches it to an apprentice via 5 steps/stages, as seen in Fig 2, which is modelling, coaching, reflection, articulation and exploration [7].

- Modelling – Demonstrating the thinking process
- Coaching – Assisting and supporting student cognitive activities as needed (includes scaffolding)
- Reflection – Self-analysis and assessment
- Articulation – Verbalizing the results of reflection
- Exploration – Formation and testing of one’s own hypothesis

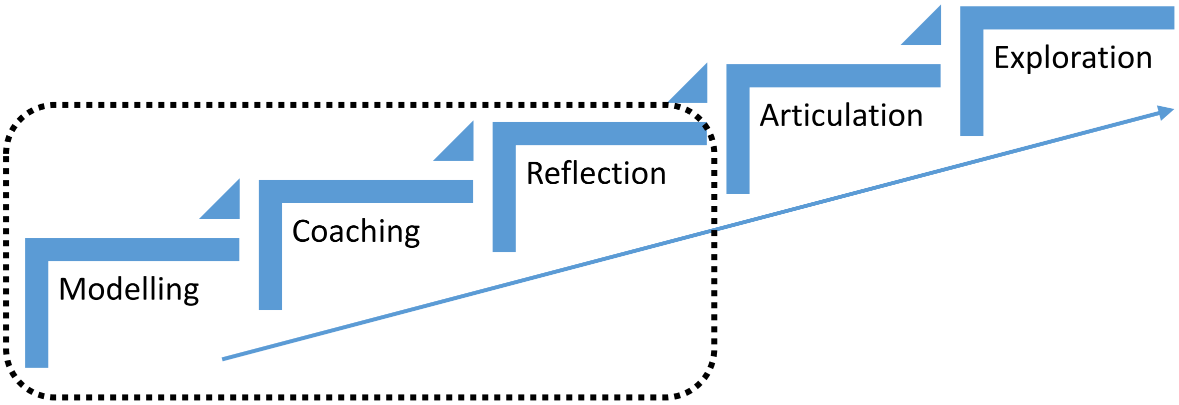


Figure 2 Phases in the Cognitive Apprenticeship Learning Process

The dashed line box in Fig 2. Illustrates the focus of this research, which is modelling, coaching, and reflection, whereby the original speech would be the model, the scaffolding being the coach, and self-evaluation being the reflection.

Coaching would be done via scaffolding with the 4 elements being used to control the difficulty. The 4 elements would be discussed in Section 3.1. Initially the user would be subject to a speech shadowing session to judge their own level and a speech without any scaffolding. After the initial rating, the user will then be given scaffolding suited to his level.

At this phase of the research, reflection would be self-evaluation. The user would be given some visual aids such as the audio waveform in order to evaluate his own performance and then he would answer a questionnaire. Feedback such as graphs will then be provided to show the user his current performance in various aspect of speech such as intonation, tempo, and pronunciation. The user can also track his past performances. These metrics would be fed back to the system in order to determine the coaching needed for the next shadowing session.

3. System Design

Due to the impracticality of the traditional speech shadowing for language learning on a larger scale, we propose a system that is able to replace the role of the instructor of the traditional method. At the same time, we want the system to provide a more tailored learning method for the student using it, so that he/she may learn and improve faster. The lack of an instructor also allows the student to learn independently, and due to the simplicity of our proposed system, the system can also be implemented on a mobile system, allowing students to learn anywhere and anytime. In Fig. 3, we can see the use case diagram for the system. The user will be able to check his/her past-performance from the system, do speech shadowing, and also receive scaffolding during his/her sessions to improve the training process. With every speech shadowing session, the user would also perform self-evaluation.

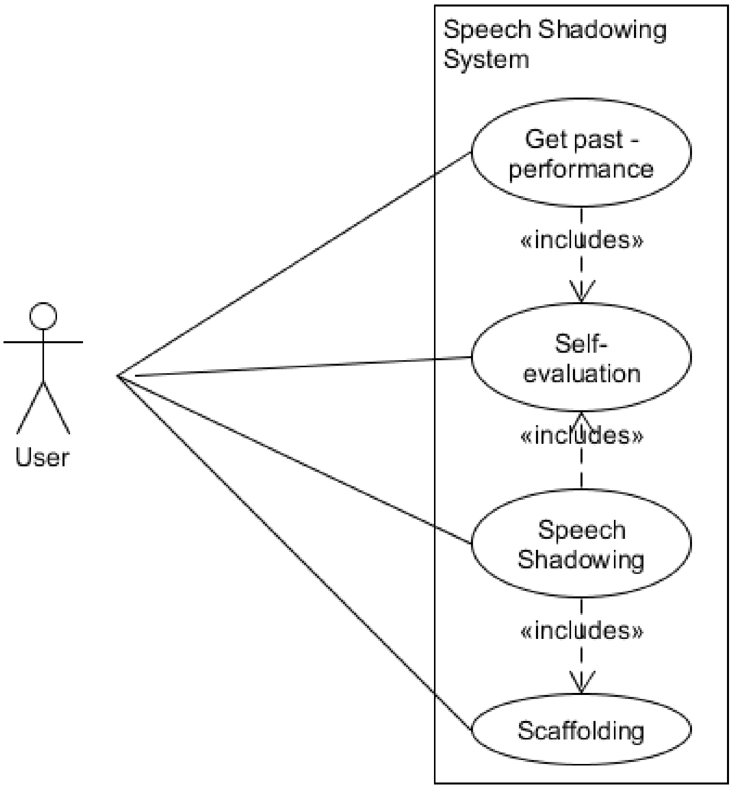


Figure 3 Speech Shadowing System Use Case Diagram

3.1 Speech Shadowing System

The system would contain recordings of speeches to be listened to, and the speeches will be sorted by difficulty levels according to their length, speed, and difficulty of the words or sentences. The system would also pickup and record the speech shadowed by the student so that it can be analysed to provide feedback and evaluation.

3.1.1 Determining the difficulty of a speech

The difficulty level of the speeches will be determined by the following elements of speech:

- Length of speech
- Speed/tempo of speech
- Difficulty of words used
- Number of stresses/intonation in sentence

The reason the elements are chosen are explained as follows. The length of speech can directly affect the difficulty of the speech as it increases the cognitive load as it becomes longer. The speed and tempo of a speech also affects the difficulty of a speech as speech rate (the number of words spoken per minute) has been used extensively in the previous research of oral fluency [8] [9] [10]. Previous research also found that speech rate positively correlated with other measures of fluency, such as length of speech without pauses, hesitations, or repeats [11] [12]. Difficulty of words that appear is also taken into consideration as it can affect the understanding of a shadowed speech.

The number of stresses and intonation in a sentence can affect the difficulty of a speech because linguistic, syntactic and semantic information is more easily conveyed when a speaker produces the correct variations in pitch in a speech utterance [13]. Of all the elements of a target language, the intonation appears to be the most difficult to acquire [14]. First, because the intonation in infants is learned at a very early stage in the language-acquisition process [15], it is most resistant to change. Second, as a result of the fact that suprasegmental patterns are particularly deep-rooted, foreign language learners often superimpose the prosodic features of their mother language on the sounds of the foreign language. For this reason, foreign-language learners are often not aware of any differences in intonation between the mother language and the foreign language [4]. This makes the number of stresses in a sentence directly related to the difficulty of shadowing a sentence.

3.1.2 System Platform

We propose that the system runs on a smartphone so that it can make the learning process more accessible as year-by-year digital media audiences are increasingly coming from mobile devices [16]. Setting up a headset is also easier and less costly compared to a desktop-based system as most smartphone owner would already have access to a headset. This also ensures students can learn on the go, although they should use the system in an isolated environment to avoid disturbing others.

3.1.3 User Management

System users will have an account created for progress tracking purposes. First time user of the system would take a standardised test and answer a short questionnaire to determine his/her initial level and proficiency (system initialization). The test would be a speech shadowing session without any support from the system. The difficulty of the speech would also be a predetermined medium level speech.

Under a normal use-case condition (post-initialization), students would login to the system and be presented with a list of recommended speeches to shadow, which are determined by the student's proficiency and level. The amount and type of scaffolding provided during a shadowing session is affected by the student's proficiency and level along with the difficulty of the speech attempted. Scaffolding availability is determined via a simple formula of **User Level /** 2, as described in Table 1.

Take for example Student A is rated by the system as a level 6 user (out of 10 possible levels, with 1 being lowest and 10 being highest) attempts a speech of difficulty level 2 (out of 5 difficulties with 1 being easiest and 5 being the hardest). Student A would get no scaffolding as his proficiency should be sufficient to attempt the speech with ease. However, if Student A attempts a level 5 difficulty speech, all scaffolding would be activated to help Student A with his shadowing attempt. In the optimal scenario, Student A should be attempting speeches with difficulty level that matches his own proficiency level, as the effect of learning via speech shadowing can be affected by having too much scaffolding.

Table 1 User Level and Speech Difficulty Level Matching

(User Level) / 2	**Scaffolding**	**Notes**
> speech level	No scaffolding	Scaffolding provided depends on user's proficiency on speech elements as well
= speech level	Partial Scaffolding	
< speech level	More / All Scaffolding	

3.1.4 Types of Scaffolding

Based on the 4 elements of speech that is used to determine the difficulty level, the system also correspondingly provides 4 types of scaffolding. Fig. 4 shows the types of scaffolding that is provided by the system being used:

1. Speech transcript (helps with overall understanding of the speech)
2. Pronunciation help (helps with individual words of speech)
3. Highlighting sentence stress points (helps with understanding points of intonation)
4. Speed control for recordings (helps with overall difficulty of speech)

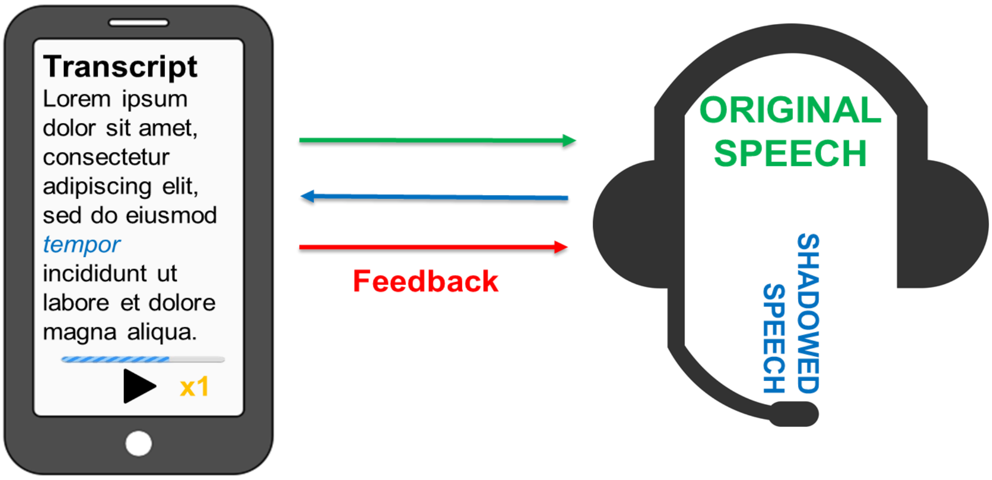

Transcript
Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.
x1
Feedback
ORIGINAL SPEECH
SHADOWED SPEECH

Figure 4 Scaffolding 1,3, and 4 being used in a shadowing session

3.2 Performance Evaluation Metrics

In order to provide the student with a valuable feedback and evaluation without an instructor, a way to grade the speech shadowing session needs to be devised. Using 3 metrics, the user's performance can be measured more accurately and the training time needed can be shortened as the student knows what he has to focus on to improve. The 3 metrics that is used in this system are:

- Intonation
- Pronunciation
- Tempo

The user would evaluate the 3 metrics on his own by comparing his shadowed speech to the original recording. Using a simple questionnaire, the student would rate his own performance compared to the sample recording. The system will provide some visualisation of the data in order to make the process easier.

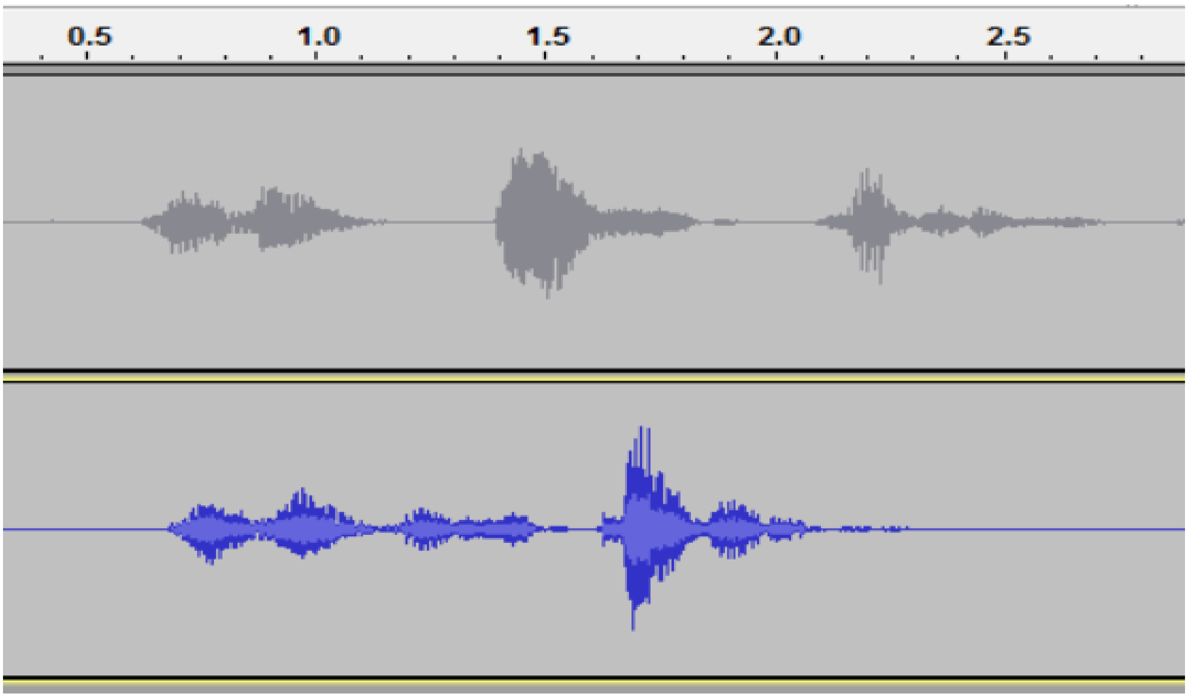


Figure 5 Visualization of intonation difference

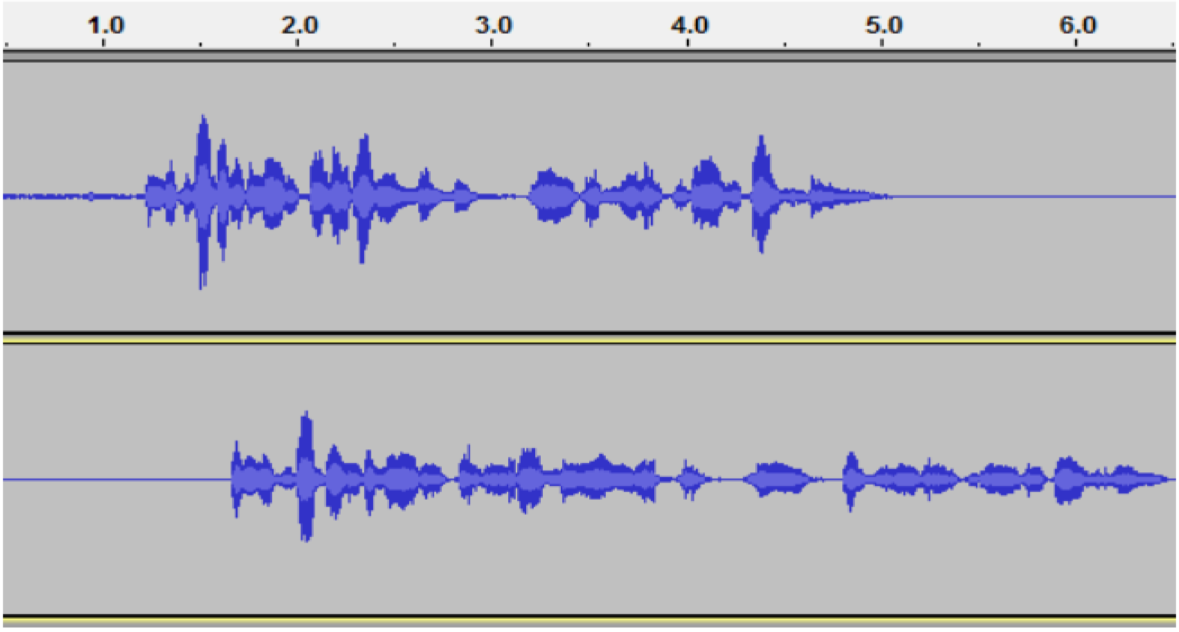


Figure 6 Visualisation of tempo difference

In Fig. 5, we can see an example of intonation difference. While the number of peaks and shapes match, the amplitude of both waveforms are different. Using this information, students can perform self-evaluation more accurately. Fig. 6 shows tempo difference, and to a certain extent intonation difference as well. We know this because the time taken to form all the peaks are different in both waveforms, along with some of the amplitudes. For pronunciation difference, the shape of the waveform would be completely different from the original (unpronounced/missed words fall into this category too). An explanation/tutorial would be provided to users prior to self-evaluation so that they understand how to fully utilize the given information.

After the evaluation is done, the system would use the data to determine if a user has levelled up and thus have some of the scaffolding removed. The data would also be archived so that users can keep track of their past performance and pinpoint where their weakness is.

3.3 Evaluation Algorithms – Determining user level

In order to determine the user level, a number of variables are taken into account. The variables are shown in Table 2. The variables are then plugged into an equation to solve for the user level. The variables are first calculated during system initialization. They are then updated every time the user attempts another session of speech shadowing.

S_i, S_t, and S_p are scores that are recorded after the user does self-evaluation. Each score has their corresponding weightage W, which is a constant that will be multiplied when calculating the score of a session, S_{cs}. The weightage constants are subject to change and will be revised throughout the system development in order to optimize the equation.

Table 2 Variables

Variables	**Definition**
S_i	Score – intonation
S_t	Score – tempo
S_p	Score – pronunciation
S_{cs}	Score – current session
S_{pp}	Score – past performance
N_s	Total Number of Sessions
W_i	Weightage – intonation
W_t	Weightage – tempo
W_p	Weightage – pronunciation
UL	User Level

In Equation 1, the score of the current session, S_{cs}, is calculated and used to determine the user level. It is worth noting that S_{cs} ranges between 0 – 100, while user level ranges from 1 – 10. After determining the user level, the current session score, S_{cs}, would be registered as the score of past performance, S_{pp}.

Equation 1 Post System Initialization

$$S_{cs} = (S_i \times W_i) + (S_t \times W_t) + (S_p \times W_p)$$

$$UL = \frac{S_{cs}}{10}$$

$$S_{pp} = S_{cs}$$

After the system is initialized, Equation 2 would be used in all future calculation of the scores and user level. The only difference between Equation 1 and Equation 2 is that the score of past performances, S_{pp}, is taken into consideration.

Equation 2 Next Iterations

$$S_{cs} = (S_i \times W_i) + (S_t \times W_t) + (S_p \times W_p)$$

$$UL = \frac{(S_{cs} + S_{pp}) \times {}^{1}/_{N_s}}{10}$$

$$S_{pp} = \frac{S_{cs} + S_{pp}}{2}$$

4. System Development

4.1 System Requirements

This system was designed for the Android operating system primarily to take advantage of the vast libraries available for audio processing, as well as the open source nature of the libraries. The target device needs to be running at least Kit Kat, because of the needed libraries for generating the audio waveform. The detailed system requirements are listed in Table 3.

Table 3 Speech Shadowing System Minimum Requirements

Minimum OS version	Android API level 19 (Kit Kat)
Audio Output	Auxiliary 3.5mm
Audio Input	Built-in mic
Processor	1.2GHz quad-core Qualcomm Snapdragon 40
RAM	512MB
Storage	At least 100MB

4.2 Development Environment

The development environment used for this system is **Android Studio 2.2.3**. Being Android, the language used for the development of this system is **JAVA**, while **SQL** is used for handling the **SQLite** database. **Mozilla Firefox** with **SQLite add-on** was used to extract data from the database on the phone.

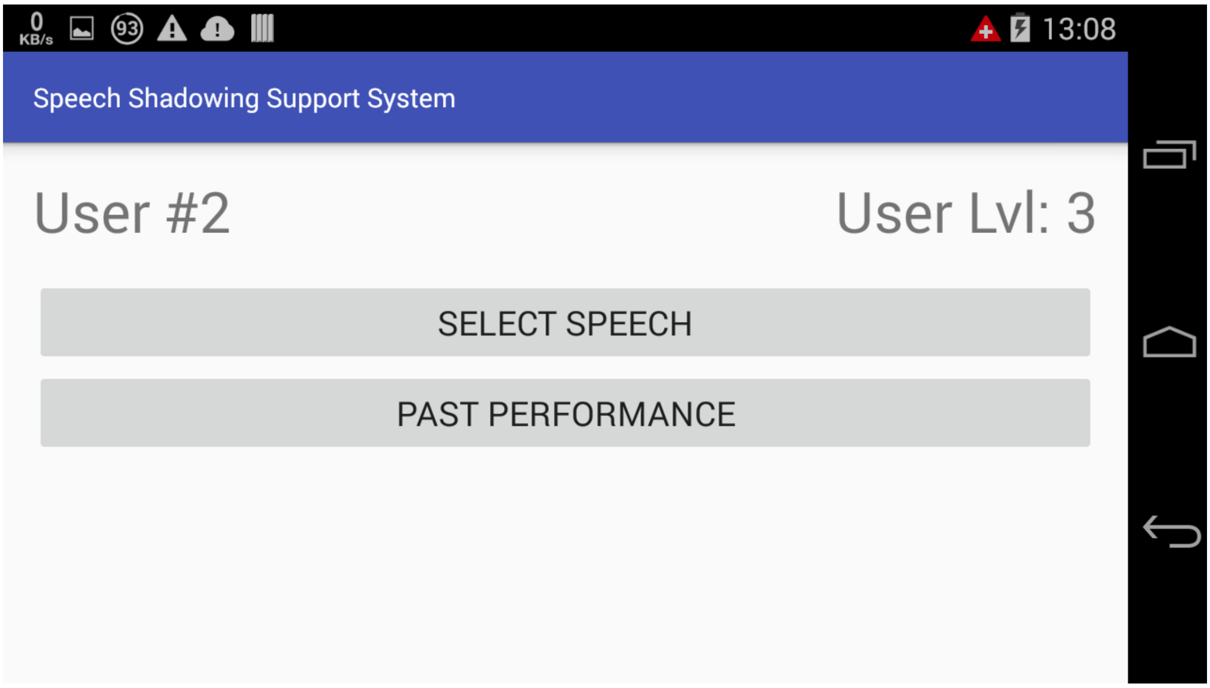
Audacity 2.1.2 was used to edit and handle the audio files used in the system. **Notepad++ v6.7.8.2** was used for creating and editing the transcripts used by the system. A **Motorola Moto G (2013)** was used as the test device in this thesis.

4.3 External Libraries

Solitaire's **waveform-android** library was used in the system for drawing the audio waveforms. (https://github.com/Semantive/waveform-android)

4.4 User Interface

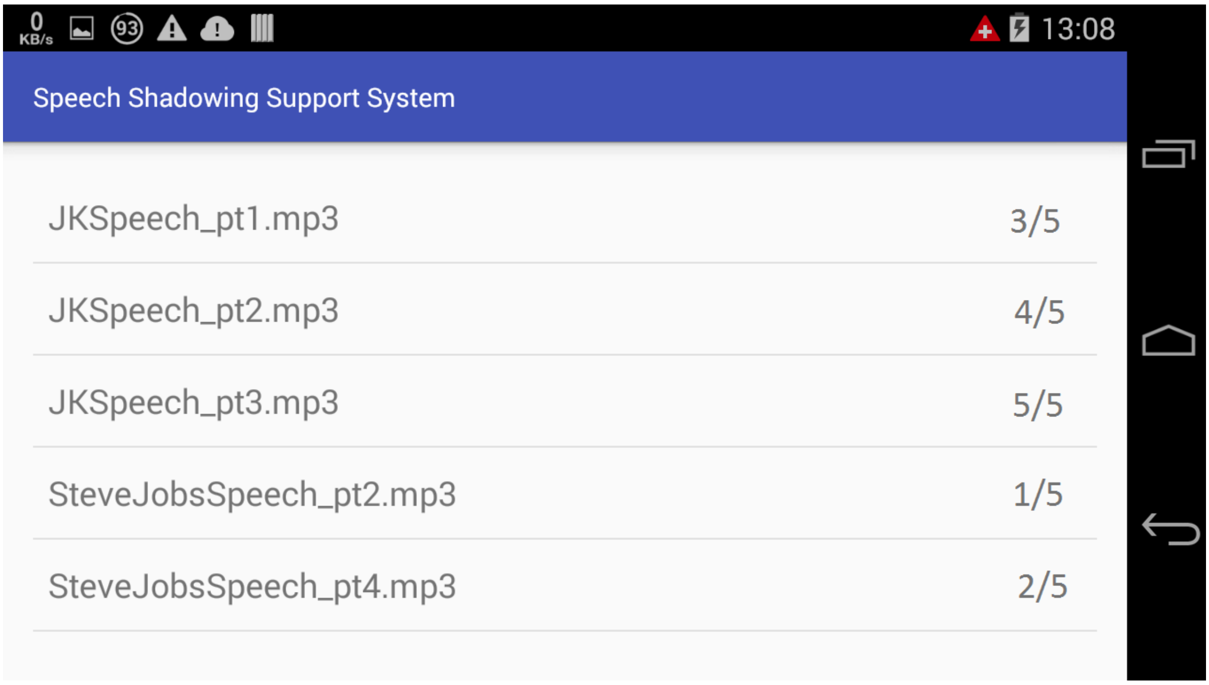
For the purpose of this research, a prototype was developed in order to evaluate the effectiveness of the system. Therefore, non-critical systems are not implemented as they do not affect the actual speech shadowing learning support system when case-studies are done under controlled environment. The following figures also show a rough process flow for speech shadowing.



User #2 User Lvl: 3

Figure 7 User's Main Screen

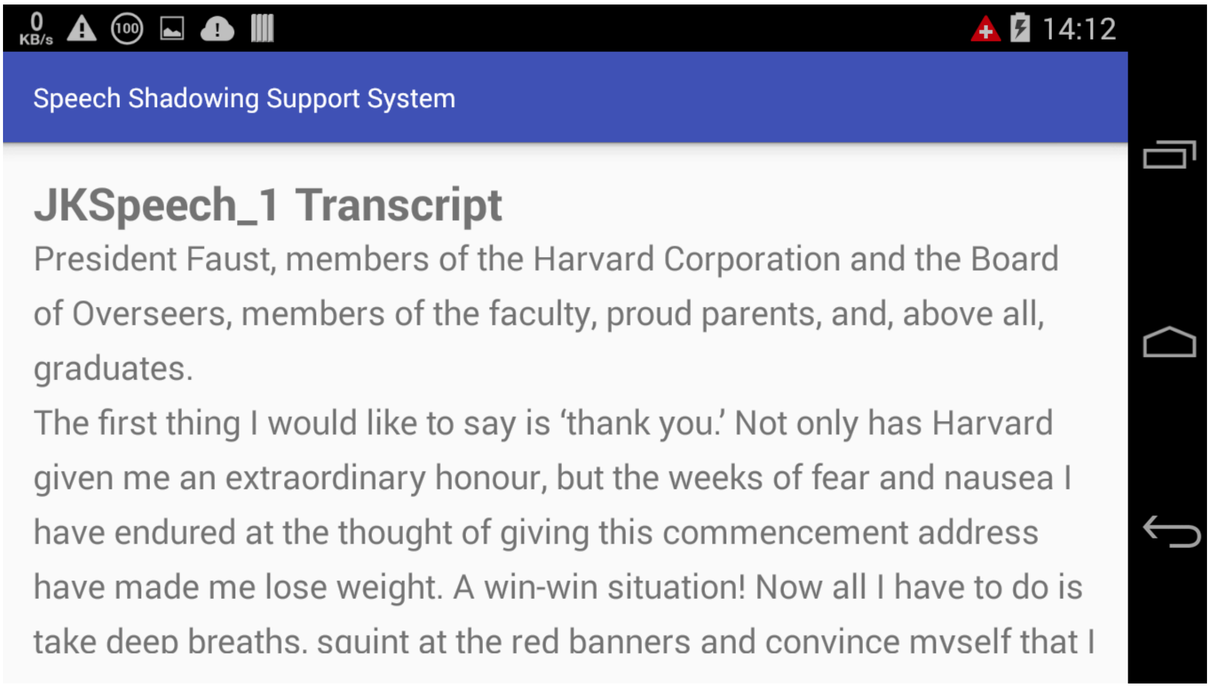
In Fig. 7, we can see that the user ID is displayed on the top left corner while the user's current level is displayed on the top right. This way the user can check his/her level easily before selecting a speech.



JKSpeech_pt1.mp3	3/5
JKSpeech_pt2.mp3	4/5
JKSpeech_pt3.mp3	5/5
SteveJobsSpeech_pt2.mp3	1/5
SteveJobsSpeech_pt4.mp3	2/5

Figure 8 Speech Selection Screen

The speech selection interface is shown in Fig 8. In future work, more metadata would be displayed as well, but it is excluded for now as the metadata is not critical to the learning process.



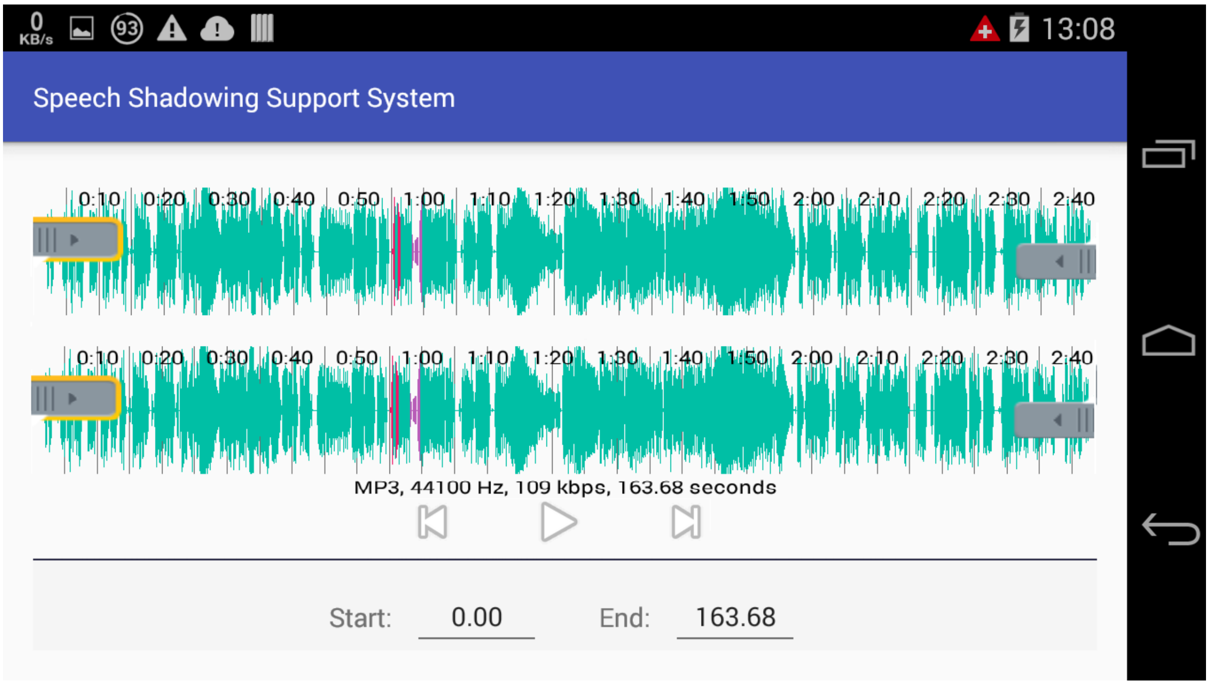
JKSpeech_1 Transcript

President Faust, members of the Harvard Corporation and the Board of Overseers, members of the faculty, proud parents, and, above all, graduates.

The first thing I would like to say is 'thank you.' Not only has Harvard given me an extraordinary honour, but the weeks of fear and nausea I have endured at the thought of giving this commencement address have made me lose weight. A win-win situation! Now all I have to do is take deep breaths, squint at the red banners and convince myself that I

Figure 9 Transcript Scaffolding Being Used

Figure 9 shows how one of the type of scaffolding(transcript) is applied. This type of scaffolding is only applied when the user level is critically lower than the level of the speech.



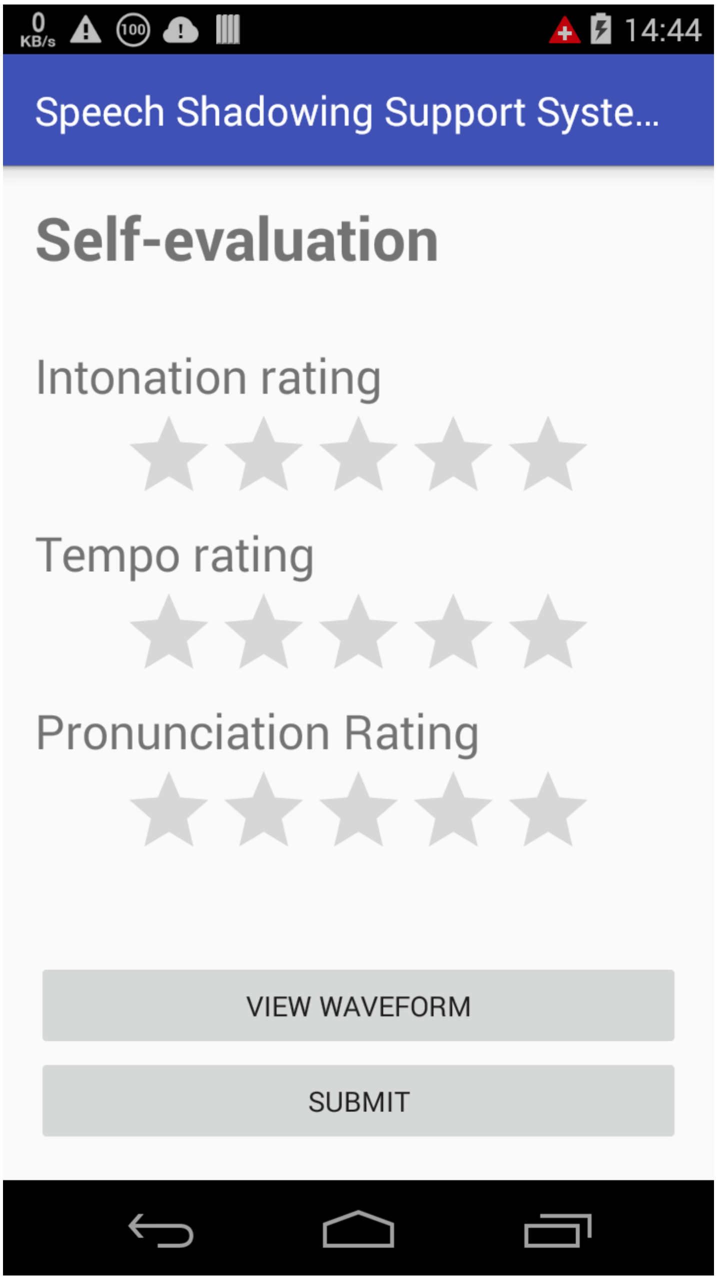
Speech Shadowing Support System

MP3, 44100 Hz, 109 kbps, 163.68 seconds

Start: 0.00 End: 163.68

Figure 10 Mock-up of the two waveform after shadowing

Figure 10 is a mock-up of the two waveform that are generated after a speech shadowing session, with the top being the waveform from the original speech while the bottom being a waveform generated by the user.



Self-evaluation

Intonation rating

☆☆☆☆☆

Tempo rating

☆☆☆☆☆

Pronunciation Rating

☆☆☆☆☆

Figure 11 Self-Evaluation Screen

In figure 11, a user would perform self-evaluation, after he/she has taken a look at the waveform to visualise his/her performance. (this screen is captured in portrait orientation to display the full UI)

After submitting his/her evaluation, the user can begin the whole speech shadowing process anew.

5. Case Study

5.1 Purpose of Case Study

A case study was conducted to evaluate the system’s effectiveness. 8 students from JAIST were selected to participate in this case study. Following a standardised procedure, the students were first given a preliminary test to assess their competence level. The instruction manual can be found at the end of this paper in Appendix A.

This case study was done to determine if the speech shadowing support system is capable of replacing the instructor in a traditional speech shadowing method.

5.2 Type of data to be gathered

During this case study, a few type of data was captured, namely the 3 component score described in Section 3.2, which are intonation, tempo, and pronunciation. Furthermore, the usage of scaffolding in the system is also recorded. In addition, the shadowed sound file is also saved in order to do a more detailed analysis on the audio waveform.

5.3 Preliminary Evaluation

Using a short speech (part of Steve Job's Standard 2005 Commencement Speech), students were instructed to shadow the speech. The speech had the tempo slowed down by 10% after numerous students found the original speed too fast to shadow. The procedure is as follows:

1. Read the speech transcript given
2. Listen to the original recording
3. Attempt to shadow
4. Feedback by examiner
5. Evaluation/score is recorded

The following tables contains the result from the preliminary evaluation:

Table 4 Preliminary Evaluation Score

STUDENT	TOEIC SCORE	INTONATION	TEMPO	PRONUNCIATION	OVERALL
1	800	2	2	3	23.333
2	735	4	6	4	46.667
3	760	2	4	2	26.667
4	455	1	2	4	23.333
5	890	2	2	3	23.333
6	810	2	3	5	33.333
7	890	3	4	4	36.667
8	680	4	4	4	40.000

Table 4 shows the scores of the students as graded by an instructor. Table 5 below shows the scores of the students as graded by themselves (self-evaluation). As the control group would continue to be graded by the instructor while the test group would be using self-evaluation, and at this point of the case study, they were yet to be separated into the two groups, both type of evaluation was recorded.

Table 5 Performance based on self-eval

STUDENT	TOEIC SCORE	INTONATION	TEMPO	PRONUNCIATION	OVERALL
1	800	1	2	3	20.000
2	735	2	5	5	40.000
3	760	2	4	2	26.667
4	455	1	4	5	33.333
5	890	2	2	3	23.333
6	810	4	3	8	50.000
7	890	6	5	6	56.667
8	680	4	5	5	46.667

5.4 Control vs System User

The students were split into 2 groups of 4 based on their overall scores. The control group would undergo the traditional speech shadowing method while the test group would use the speech shadowing system. They are split as shown in Table 6.

Table 6 Group Members

Control Group (Traditional)	*Test Group (System)*
2	1
3	5
4	6
7	8

After they were split up, both groups underwent 3 training sessions over the course of 3 days.

5.5 Results

The following are the results obtained after the case study. For test groups, one extra column of data is recorded, which is the type of scaffolding applied. The blue-coloured tables are control groups, and were evaluated by the instructor. The green-coloured tables with the extra scaffolding are the test groups, and they were self-evaluated. The Intonation, Tempo, and Pronunciation are recorded as it is by the instructor and the system while the overall score was calculated using the Equation 2, as described in Section 3.3. The scaffolding column shows what type of scaffolding was provided and used by the students during their shadowing session. “slow” refers to playback speed slowdown scaffolding, “trans” refers to providing the speech transcript scaffolding.

From the results below, we can observe that students have shown improvement in their scores regardless of the group they are in. Students in the test group who received scaffolding did improve slightly better than the student who did not receive scaffolding from the system. The detailed analysis of the results is discussed in the next section, Section 5.6.

Table 7 Session 1 Control Group Score

Student	Intonation	Tempo	Pronunciation	Overall
2	2	2	5	30.000
3	2	6	2	33.333
4	2	1	2	16.667
7	4	5	4	43.333

Table 8 Session 1 Test Group Score

Student	Scaffolding	Intonation	Tempo	Pronunciation	Overall
1	Slow	1	2	2	16.667
5		3	3	3	30.000
6		6	5	7	60.000
8		3	6	4	43.333

Table 9 Session 2 Control Group Score

Student	Intonation	Tempo	Pronunciation	Overall
2	4	2	2	26.667
3	2	5	2	30.000
4	6	5	5	53.333
7	4	5	4	43.333

Table 10 Session 2 Test Group Score

Student	Scaffolding	Intonation	Tempo	Pronunciation	Overall
1	slow w/ trans	3	5	3	36.667
5	slow	4	4	3	36.667
6		7	5	8	66.667
8	slow	5	7	5	56.667

Table 11 Session 3 Control Group Score

Student	Intonation	Tempo	Pronunciation	Overall
2	4	4	5	43.333
3	2	7	3	40.000
4	6	6	5	56.667
7	6	7	7	66.667

Table 12 Session 3 Test Group Score

Student	Scaffolding	Intonation	Tempo	Pronunciation	Overall
1		2	5	3	33.333
5		4	5	3	40.000
6		8	6	8	73.333
8		6	8	6	66.667

5.6 Result Analysis

5.5.1 Control Group

Under the control group, students were given feedback by an instructor to help them improve. While the instructor gave very objective feedback, it can be seen that students may interpret it slightly differently than intended.

For example, Student 2 was given the feedback that he should focus on his intonation during session 1. During session 2, Student 2 did show marked improvement on his intonation score, however the over-zealous focus on one component caused his pronunciation score to fall, and his tempo to show no improvements. After being given another feedback that all aspects should be focused on equally, Student 2 then shifted his focus to his weaker components, which is tempo and pronunciation.

Another interesting observation is Student 4. Student 4 showed amazing interest in Speech Shadowing for language learning, and went on to attempt Speech Shadowing on her own between the sessions. This resulted in a remarkable increase in score during the 2nd session.

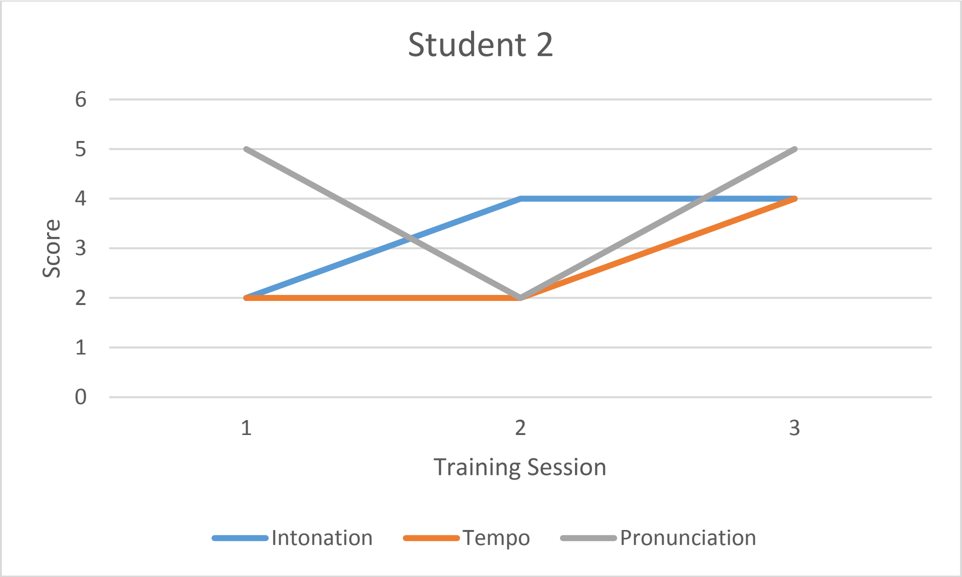


Figure 12 *Student 2 Performance Graph*

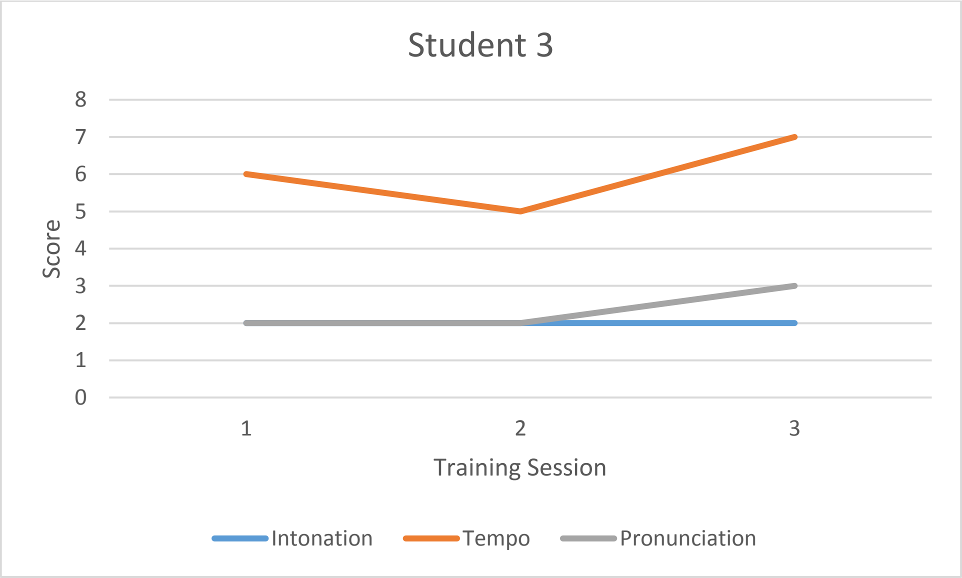


Figure 13 *Student 3 Performance Graph*

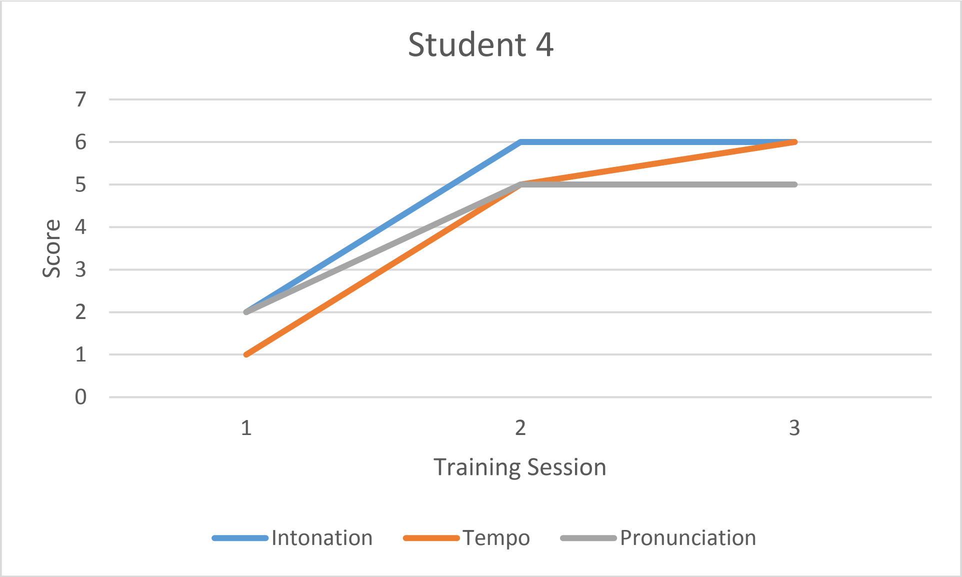


Figure 14 Student 4 Performance Graph

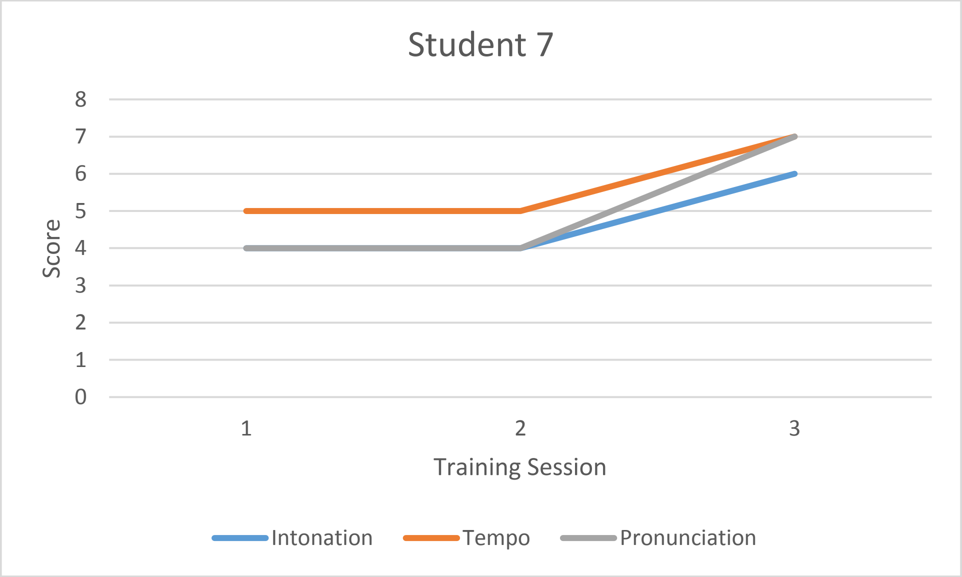


Figure 15 Student 7 Performance Graph

5.5.2 Test Group

Under the test group, students were given careful instructions on how to do self-evaluation by using the audio waveform. I was also present to monitor and provide additional tips on how to abstract useful information from the audio waveform. The use of any scaffolding by the system was also recorded and considered when analysing their results.

Student 1 did very poorly even on the preliminary test, and thus the system provided playback speed reduction as a scaffolding on her first attempt. Despite that, Student 1 still did not show much improvement. Due to the low scores, the system provided Student 1 with every scaffolding available at present (transcript, highlighted intonation stress points, slowed playback). With all the scaffolding in place, Student 1 showed major improvement on the 2nd session. The system then retracted most scaffolding for the 3rd session and Student 1 managed to barely maintain her score.

Due to the low intonation score of both Student 5 and 8 during their 1st session, the system slowed the playback speed for the 2nd session and both students showed improvement in their overall score.

Student 6 on the other hand showed very good performance from the 1st session onwards, thus the system did not provide scaffolding.

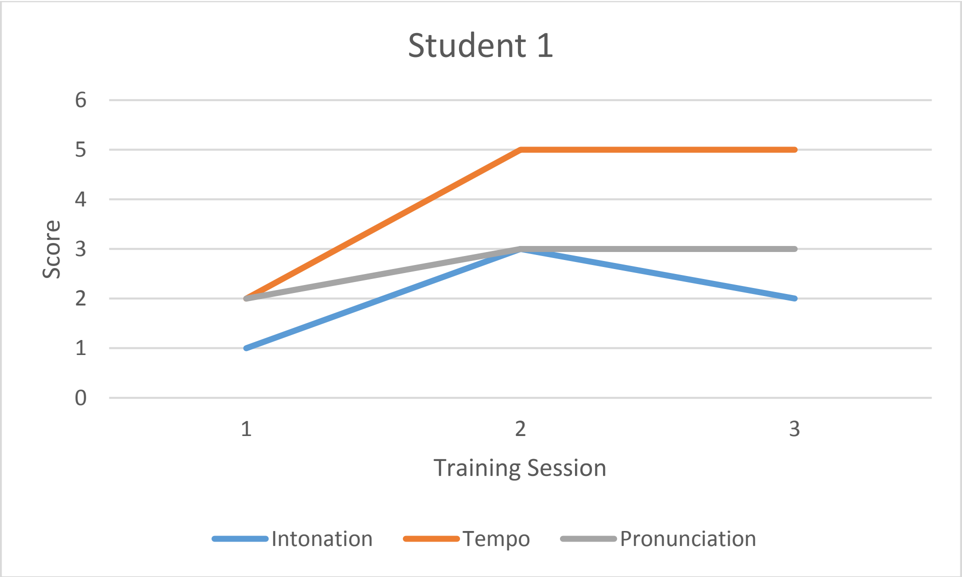


Figure 16 Student 1 Performance Graph

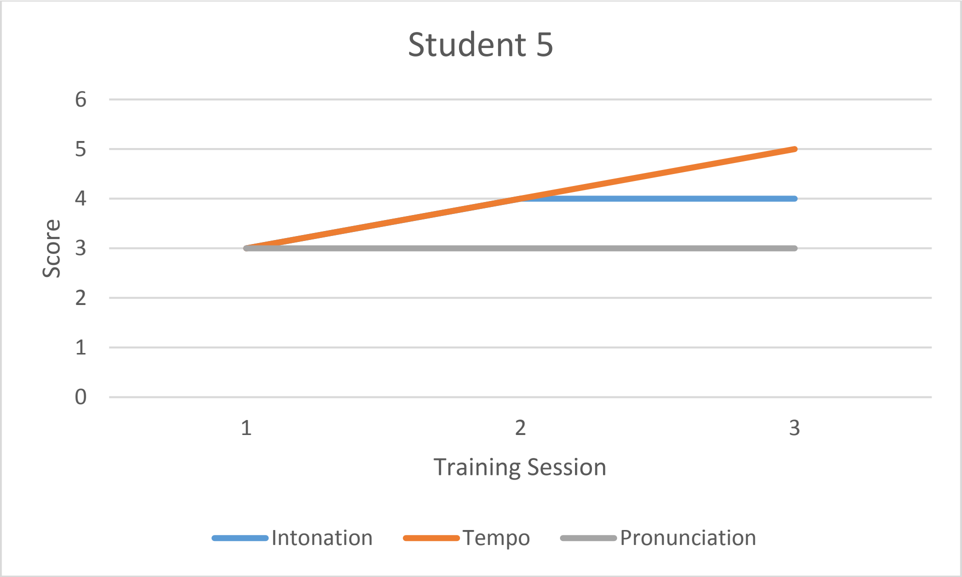


Figure 17 *Student 5 Performance Graph*

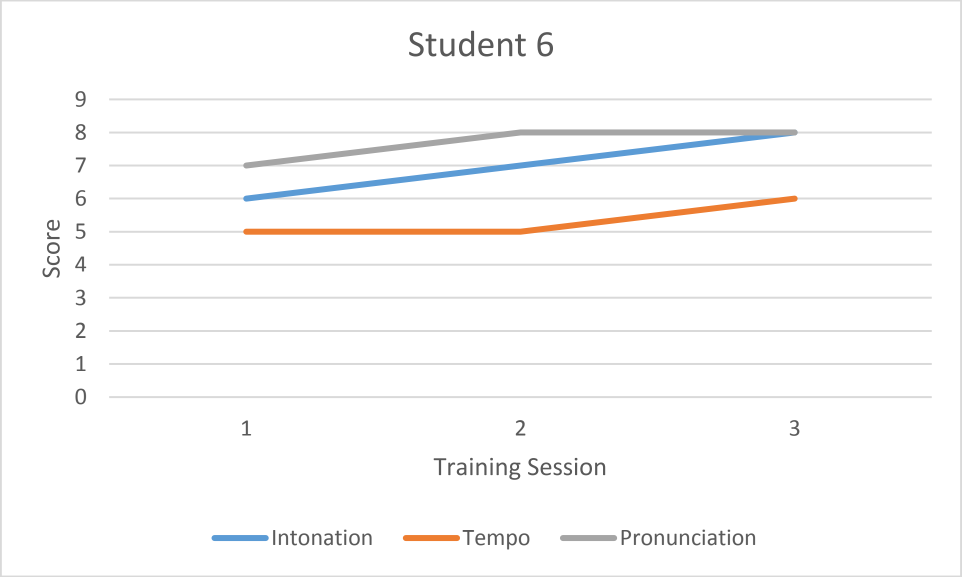


Figure 18 *Student 6 Performance Graph*

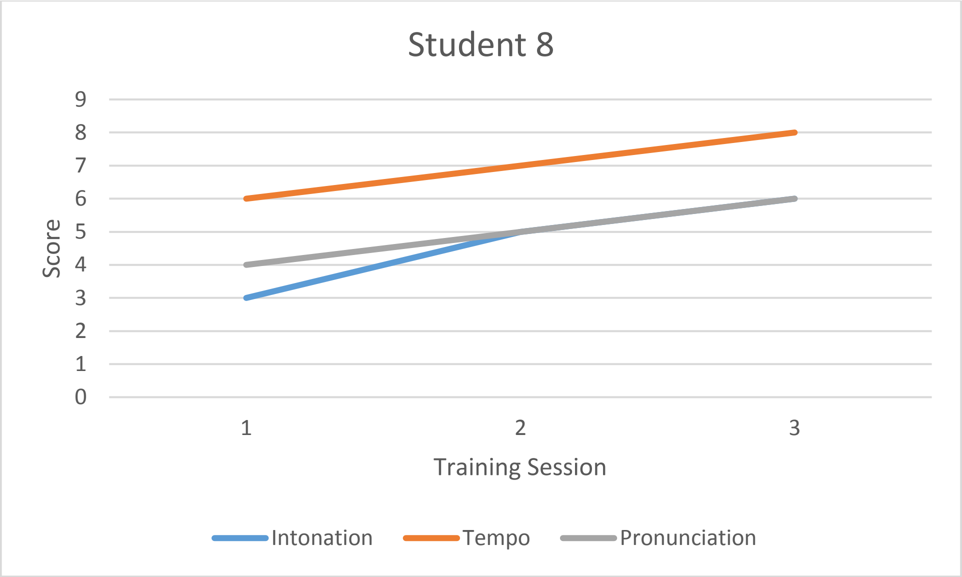


Figure 19 Student 6 Performance Graph

5.7 Feedback Questionnaire

After the 3 sessions, the students were given a questionnaire to evaluate the learning process. The questionnaire can be found in Appendix B. The following are the results from the questionnaire. They are likert scale questions from 1 – 5, with 1 being Strong Disagree and 5 being Strong Agree.

Table 13 Questionnaire Results

Speech shadowing sessions improved your speaking skills	4	5	4	4	4	4	4	4
Reducing the speed of the speech helps	4	3	5	3	4	5	3	5
Having the transcript during shadowing helps	4	4	4	5	5	5	4	4
Having words that are stressed highlighted in the transcript helps	4	3	3	4	5	3	5	3
Seeing your past performance helps	4	5	3	2	4	2	2	3
Knowing the component score is better than just overall score (i.e. intonation score, tempo score)	4	5	4	4	4	5	5	5
Having your performance in a visual/graph form is better than verbal comments	4	2	4	4	4	3	3	4
(For system users) The audio waveform is helpful in visualising your performance	3	2	3	4	4	3	4	4

5.8 Discussion

From the data gathered, we can see that both methods provide roughly the same improvement over the course of 3 sessions. We can conclude that the system is as effective as the traditional method from this case study. One point to note is that some students have commented that the audio waveform is not helpful in visualising their performance.

Another interesting point is that the students using the system evaluate themselves very differently from their actual performance. While some did indeed perform well and rated themselves accordingly, some students rated themselves much lower than an instructor would score them. Conversely, students who performed rather weak rated themselves much high than the instructor would. Despite that, the students have all shown noticeable improvement over the 3 sessions.

One last point to note is that the students TOEIC score does not correlate to their performance. In fact, some of the highest TOEIC scoring students did the worst in both the traditional method and system.

6. Conclusion

In this research we proposed a Speech Shadowing Learning Support System in language learning to solve and overcome the problems faced in traditional speech shadowing, namely the high cost of training. This was proposed because of the lack of systems available to provide training/coaching in spoken language learning for the slightly more advanced learner. During the research, a basic speech shadowing learning system was developed on Android and was used to conduct a case study. The results show that the system is at least capable of providing the same level of improvement as an instructor would, over the course of 3 sessions. Feedbacks from the students also show that some parts of the system are desirable and helpful for speech shadowing.

6.1 Future Work

One notable comment received from the students is that they find it hard to identify their weak points when shadowing. A method should be implemented to highlight their weak points and also provide them with good examples and tips on how to achieve it. For example, if their weak point is pronunciation, the system should show them a comparison of their pronunciation vs a native speaker's pronunciation. The system can also provide videos on lip/tongue movement when the voice is generated by the native speaker.

Students have commented that while having the audio waveform helps them in visualising the data, it is still very abstract and thus it is not easy nor intuitive to use for self-evaluation. In the future, the system should have automated evaluation, thus removing one uncertainty variable when evaluating the effectiveness of the system. Big Data Analysis is one way this automation can be achieved; however, this method requires a very large pool of data. Another method that can be used is signal processing.

One more point to address in the future is the normalization and boosting of the shadowed speech. During the case study, some students spoke much softer than others, and thus the picked up audio waveform have much lower amplitude than the original. Using Audacity, I was able to boost and normalize the shadowed speech so that the waveforms are much more similar, thus making comparison and evaluation easier. In future works, an automatic microphone gain function should be implemented into the system.

Furthermore, if Natural Language Processing and voice recognition is implemented in the system, there can be one more metric to evaluate the user’s performance by. By comparing the captured words to the transcript, the difference in words in the transcript can be evaluated and be used as a performance indicator.

Acknowledgement

Upon the closure of my thesis, I cannot help but feel ephemeral as it only felt like yesterday when I first stepped into JAIST as a new student. I still remember the classes I attended during my first year. The vast and deep knowledge that the professors here possess in their field amazed me every single time, and the other students from every corner of the world humble me with their tenacity and talent in their studies and research. From studying to working part-time jobs in Japan, this great nation has imparted on me a great deal of knowledge and ethics. These two years have changed me both physically and mentally.

I would like to thank all my friends who helped me through thick and thin, who supported me when I was at my lowest point. To the friends that I made here in Japan, I want you to know that your friendship has proved invaluable to me. Being alone in a foreign land and barely speaking the language would have been too much for me to shoulder. And to my foster family, thank you for making me feel at home and having family even in Japan.

Last but not least, I would like to extend my most heartfelt gratitude to my supervising professor. Hasegawa-sensei has been very patient and understanding to me despite my lack of progress. He has given me far too many opportunities to do great things, things that I would not even have dreamt of doing just 2 years prior. In fact, if it wasn’t for him, I would not have had the financial means to attend such a wonderful graduate school. Words cannot express my gratitude towards Hasegawa-sensei. Thank you for going above the call of duty for me.

Publication List

1. **Carson Lee** and Shinobu Hasegawa: Speech Shadowing Support System in Language Learning. *The 9th International Conference on Mobile, Hybrid, and On-line Learning.* Nice, France. March 2017. (Accepted)

References

[1] Kevin Shockley, L. S. Imitation in shadowing words. Perception & Psychophysics , pp. 422-429, 2004.

[2] Marslen-Wilson, W., Linguistic Structure and Speech Shadowing at Very Short Latencies. Nature Vol, pp. 244, 522-523, 1973.

[3] McMillan, David. Miscommunications in Air Traffic Control. Diss. Queensland University of Technology, 1998.

[4] Gerard W. G. Spaai and Dik J. Hermes, A Visual Display For the Teaching of Intonation. CALICO Journal, Volume 10 Number 3, 1993.

[5] Shudong, Wang, Michael Higgins, and Yukiko Shima. "Teaching English pronunciation for Japanese learners of English online." JALT CALL Journal 1.1 (2005): pp. 39-47.

[6] Trevor A. Harley and L. J., Decline and fall: A biological, developmental, and psycholinguistic account of deliberative language processes and ageing. Aphasiology, 2011.

[7] Dennen, Vanessa P., and Kerry J. Burner. "The cognitive apprenticeship model in educational practice." Handbook of research on educational communications and technology 3 pp. 425-439, 2008.

[8] R. Ejzenberg and H. Riggenbach, "The juggling act of oral fluency: A psycho-sociolinguistic metaphor," Perspectives on Fluency, Ann Arbor University of Michigan, pp. 287-314, 2000.

[9] Freed B. Riggenbach H. ‘Is fluency, like beauty, in the eyes (and ears) of the beholder?’, Perspectives on Fluency, Ann Arbor University of Michigan pp. 287-314, 2000.

[10] Lennon P. ‘Investigating fluency in EFL: A quantitative approach,’ , Language Learning, 1990, vol. 40, pp. 387-417, 1990.

[11] Freed, B.F., Segalowitz, N. and Dewey, D.P., Context of learning and second language fluency in French: Comparing regular classroom, study abroad, and intensive domestic immersion programs. Studies in second language acquisition, 26(02), pp.275-301, 2004.

[12] Segalowitz N , Freed B . ‘Context, contact, and cognition in oral fluency acquisition,’ , Studies in Second Language , 2004 , vol. 27, pp. 175 -201, 2004.

[13] Crystal, D. Prosodic Systems and Intonation in English. Cambridge: Cambridge University Press, 1969.

[14] Leon, P. R., and P. Martin, "Applied Linguistics and the Teaching of Intonation." The Modern Language Journal, 56, 3, pp. 139-144, 1972.

[15] Lieberman, P. (1967). Intonation, Perception and Language. Cambridge, MIT Press. ______, and S. B. Michaels, "Some Aspects of Fundamental Frequency and Envelope Amplitude as Related to the Emotional Content of Speech." Journal of the Acoustical Society of America, 34, pp. 922-927, 1962.

[16] comScore, Cross-Platform - Future in focus 2016.

Appendix A – Testing Manual / Instruction

Student volunteers for the case study must meet the following requirements:

- Have at least simple conversational skill in English
- Have TOEIC scores less than 450
- Non-native English speaker

Accepted students would be evaluated by an instructor in order to assess their proficiency level.

This case-study would last for 4 days (inclusive of the pre-test).

The volunteers' personal information will not be disclosed, however, the audio recordings of the system users will be retained for data analysis.

Pre-test

The speech used for this case study would be Steve Jobs' Stanford Commencement (2005) Speech. The speech would be split into 5 segments, and the 2nd segment would be used for the pre-test. The following is the transcript for the segment of speech used:

> "Let me give you one example: Reed College at that time offered perhaps the best calligraphy instruction in the country. Throughout the campus every poster, every label on every drawer, was beautifully hand calligraphed. Because I had dropped out and didn't have to take the normal classes, I decided to take a calligraphy class to learn how to do this. I learned about serif and san serif typefaces, about varying the amount of space between different letter combinations, about what makes great

> typography great. It was beautiful, historical, artistically subtle in a way that science can't capture, and I found it fascinating.
>
> None of this had even a hope of any practical application in my life. But 10 years later, when we were designing the first Macintosh computer, it all came back to me. And we designed it all into the Mac. It was the first computer with beautiful typography. If I had never dropped in on that single course in college, the Mac would have never had multiple typefaces or proportionally spaced fonts. And since Windows just copied the Mac, it's likely that no personal computer would have them.
>
> If I had never dropped out, I would have never dropped in on this calligraphy class, and personal computers might not have the wonderful typography that they do. Of course it was impossible to connect the dots looking forward when I was in college. But it was very, very clear looking backwards 10 years later.
>
> Again, you can't connect the dots looking forward; you can only connect them looking backwards. So you have to trust that the dots will somehow connect in your future. You have to trust in something — your gut, destiny, life, karma, whatever. Because believing that the dots will connect down the road will give you the confidence to follow your heart even when it leads you off the well-worn path and that will make all the difference."

On completion of the pre-test, volunteers would be split into 2 groups. 1 group for control (traditional speech shadowing), another group for test (system users).

Control Group

Volunteers in the control group would undergo speech shadowing under the guidance of an instructor. Volunteers will be given feedback on their performance and tips for improvement.

Test Group

Volunteers under the test group would undergo speech shadowing by using the provided system. Please ensure that the volunteers have the headphones on properly before beginning. Volunteers should also speak as loudly and clearly as they can directly into the mic.

An instructor will be there to assist the volunteer on the steps to perform effective self-evaluation. Use of any scaffolding will be recorded.

Post Test

Upon completion of the third and final speech shadowing session, volunteers will answer a questionnaire. Additional comments will also be recorded in the second half of the questionnaire.

Appendix B - Speech Shadowing Language Learning Questionnaire

Helpfulness of the training

Choose from a scale of 1 – 5

1 being strongly disagree and 5 being strong agree

	Questions	1	2	3	4	5
Q1	Speech shadowing sessions improved your speaking skills					
Q2	Reducing the speed of the speech helps					
Q3	Having the transcript during shadowing helps					
Q4	Having words that are stressed highlighted in the transcript helps					
Q5	Seeing your past performance helps					
Q6	Knowing the component score is better than just overall score (i.e. intonation score, tempo score)					
Q7	Having your performance in a visual/graph form is better than verbal comments					
Q8	(For system users) The audio waveform is helpful in visualising your performance					

Comments/Feedback

Please let us know your comments and feedback regarding this case-study/system

Q1. What additional support would you like to have while undergoing speech shadowing?

Q2. Do you have any other comment about the speech shadowing learning process?

Q3. Additional comments