Title	iTonation : an application to help learners practice English using their own voice
Sub Title	
Author	李, 相旻(Lee, Sangmin) 大川, 恵子(Ōkawa, Keiko)
Publisher	慶應義塾大学大学院メディアデザイン研究科
Publication year	2020
Jtitle	
JaLC DOI	
Abstract	
Notes	修士学位論文. 2020年度メディアデザイン学 第793号
Genre	Thesis or Dissertation
URL	https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=KO40001001-00002020-0793

Master's Thesis
Academic Year 2020

iTonation: An Application to Help Learners Practice English Using Their Own Voice

Keio University
Graduate School of Media Design

Sangmin Lee

A Master's Thesis
submitted to Keio University Graduate School of Media Design
in partial fulfillment of the requirements for the degree of
Master of Media Design

Sangmin Lee

Master's Thesis Advisory Committee:

Professor Keiko Okawa	(Main Research Supervisor)
Senior Assistant Professor Chihiro Sato	(Sub Research Supervisor)

Master's Thesis Review Committee:

Professor Keiko Okawa	(Chair)
Senior Assistant Professor Chihiro Sato	(Co-Reviewer)
Professor Matthew Waldman	(Co-Reviewer)

Abstract of Master's Thesis of Academic Year 2020

iTonation: An Application to Help Learners Practice English Using Their Own Voice

Category: Design

Summary

This research presents the design process of 'iTonation', an application software intended to help non-native English learners in elementary schools in Korea train their prosody more easily by using their own voice.

Traditionally, non-native learners often practiced prosody by trying to imitate what they hear from native speakers or listening materials like CD or DVD (so called 'repeat after me'). However, not every non-native learners will always have native speakers close by to help them. Furthermore, unlike other phonetic features, if learners make mistakes on prosody, it's difficult to explicitly point out the mistakes and provide feedbacks.

According to a study in Korea 2010, non-native learners could acquire native prosody more efficiently by listening to their own voice, albeit converted to have native prosody, with the help of 'Praat', a computer software specialized in phonetics. However, the presented method required users to be confident in using Praat and voice samples had to be transcribed manually to conduct prosody conversions, requiring a certain amount of skills related to computer and phonetics. Moreover, in order to provide models for prosody conversions, native speakers still had to provide voice samples, with the same manual transcriptions. Lastly, the previous works focused on university students in Korea, and there was no studies that tested this method on elementary schools.

iTonation intends to expand upon the existing method by combining speech recognition and synthesis and combined to a single user workflow to make it more accessible on more non-native English learners. Speech recognition have vastly improved in recent years, being able to provide timestamps for spoken words,

which can be a great aid in replacing manual transcription. Also, synthesized speech have gained similar improvements, which may replace human speakers as the model for prosody conversion. Thus, the ultimate goal of iTonation is to be a tool of witch children can easily practice their prosody autonomously without the assistance or intervention from teachers or native speakers.

Based on the aforementioned components above, the researcher was able to create a working prototype for the application. Also, during the prototype's user tests on four students from a Korean elementary school, the children were able to run through the process of prosody conversion without the intervention from the researcher except for some minor technical issues. The children's reactions upon hearing their voices converted with native prosody were generally favorable. However, not every participants showed enthusiasm when asked about whether they would like to continue using the application for their actual English learning activities. This may indicate that supplementary means for motivating children to continue using the application may be required at the current stage of iTonation's development.

Nevertheless, this thesis argues that even at the current stage, iTonation was able to provide the basis for the autonomous environment for children to practice their English prosody.

Keywords:

English, prosody, CALL, Praat, speech synthesis, speech recognition, children, elementary school

Keio University Graduate School of Media Design

Sangmin Lee

Contents

List of Figures

List of Tables

Acknowledgements

I am in debt to Professor Keiko Okawa for guiding not only about research but with many aspects of my life. I'd also like to thank Assistant Professor Chihiro Sato for advices on many aspects of my research. Also, I'd like to acknowledge Mr. Yuta Goto, an alumni from Keio Graduate School of Media Design, for inspiring me to pursue my research goals. Lastly, I'd like to express my gratitude toward my family for their support, especially my mother who was instrumental in preparation for my research.

Chapter 1
Introduction

1.1. Background

1.1.1 Prosody

Prosody refers to the rhythmic and intonational aspect of a spoken language. Needless to say, it is a highly important factor in learning English, one of the most widely used language for communications. However, since every language has it own unique prosody, it's often difficult to acquire native English prosody for non-native learners.

Traditionally, non-native learners often practiced prosody by trying to imitate what they hear from native speakers or listening materials like CD or DVD (so called 'repeat after me'), especially during classroom environments where many learners get to know the language for the fist time [4]. However, not every non-native learners will always have native speakers close by to help them. Furthermore, unlike other phonetic features, if learners make mistakes on prosody, it's difficult to explicitly point out the mistakes and provide feedbacks.

1.1.2 Situation in Korean Elementary Schools

Moreover, regarding teaching English for younger learners, many public elementary schools in countries like Korea are facing difficulties with overall spoken English, not just prosody. In Korea, school students take up the major part in overall number of English learners. Korean curriculum for elementary school English always emphasized on the importance of speaking since its inception, and many efforts have been made including developing several teaching materials or training teachers for oral education. Despite the efforts however, the reality is that

the current level of elementary school English does not meet the requirements of students and its parents. [5]

1.1.3 Utilizing Computers

Recent developments in multimedia technology such as speech recognition and synthesis made it possible to utilize computers to aid in learning English, resulting in a concept named CALL(Computer-Assisted Language Learning). Furthermore, a number of research have been conducted regarding the alternative ways of training prosody for non-native learners. For instance, a research conducted in Korea in 2010 has shown that it's possible for non-native English learners to train their prosody through listening to their own voice, albeit converted to have prosodic features found on native speakers. [4]

However, at the current stage, this method still required native speakers to provide voice samples to be used as the model for converting prosody. Also, each voice samples required manual transcription with timestamps, and users had to be comfortable with dealing with computer softwares, which is not always the case, rendering the process time-consuming and complicated. Moreover, the research primarily focused on mature learners like university students in Korea. It's not explicitly stated about its effectiveness for younger and less mature English learners such as elementary school students in Korea. Moreover, preceding works that deals with prosody conversion methods on elementary schools have not been discovered at the moment.

1.2. Research Goal

The primary research goal of this research is to provide children from Korean elementary school with a tool that they can take home and practice prosody, even when they're away from typical classroom environments and when they don't have any help from teachers.

The previous works on prosody conversions, the prosody conversions were conducted mostly by the researchers, going through the complicated steps that can ' t be done by learners alone, especially for those who are not comfortable with computers or phonetics. Moreover, since the process requires voice samples from

native speakers speaking the same sentences the learners record, human presence is still required in the existing methods.

This thesis argues that iTonation would be a meaningful advancement, because it enables the existing methods of practicing prosody through prosody conversion accessible for much wider ranges of users without any external assistance. Using speech recognition, children can record any arbitrary sentences the'd like to speak and convert them to have native prosody. With speech synthesis, iTonation makes it possible to provide the model voice samples for prosody conversion without external help.

Rather than passively learning prosody at school, Children can use iTonation themselves and will have an opportunity to practice prosody at home without any help from school teachers or native speakers.

1.3. Thesis Outline

This thesis is comprised of five main chapters.

- **Chapter 1** discusses current issues related to prosody and prosody training for non-native English learners in elementary schools in Korea, stating the motivation and purpose of this research.
- **Chapter 2** gives more details on previous works and technology related to the issues mentioned in Chapter 1, including prosody conversion, speech recognition, and speech synthesis.
- **Chapter 3** describes the main concept and design process that leads to various design iterations.
- **Chapter 4** describes the implementation and the user test process for this research.
- **Chapter 5** summarizes the entire thesis and gives overall conclusions.

Chapter 2
Related Works and Technology

This chapter introduces several works and findings that led to the topic of this research, as well as technologies that comprises various components of iTonation.

2.1. Prosody

2.1.1 Defining Prosody

In linguistics, prosody is defined as elements of speech that don't belong in the category of individual phonetic segments, more commonly known as vowels and consonants, but are properties of larger units of speech. Prosodic elements include intonation, tone, stress, and rhythm. Those elements are also called as suprasegments.

2.1.2 Difference in Prosody Between English and Korean

According to Yoon (2011) [6], When native speakers speak English sentences, each sentence stress is isochronic. This means that the time between each stresses is always equal no matter how many words and syllables a stress is comprised of. In other words, English sentences have continuous and rhythmic beats defined by stress, and length of syllables can be different. On the other hand, when a Korean speaks his or her language, sentence rhythms are defined not by stress, but syllables. Korean speak every syllables in same lengths.

2.1.3 Difficulties For Non-native Learners

Because of this fundamental difference, it's often challenging for non-native English learners to acquire its prosody. One of the problems is that there's no established symbolic systems for dedicated to depicting prosodic systems. Traditionally, suprasegmental elements were approximately represented by drawings, such as intonation curves, but there hadn't been a comprehensive systems for accurately displaying all elements of prosody including prosody, duration, and stress. [6] Although, there have been a couple of efforts for consistent depiction of symbols for prosody, including ToBI [7], many of the new systems leave room for debates and not yet universally accepted.

Another problem non-native learners face is that, due to the vast prosodic difference between English and Korean language, and because prosodic elements are not always explicitly identified, there's been no comprehensive method of feedback while they try to learn English prosody. Traditionally, as mentioned in Chapter 1, non-native learners in Korea learned prosody from native teachers pointing out prosodic patterns in sentences.

2.2. The Situation of English Education in Korea

There are a number of research and articles that deals with the situation of English education in Korean elementary schools.

First, Lee (2018) [8] has conducted a research regarding difficulties of teaching English in Korean elementary school. According to Lee, considerable number of English teachers have difficulties regarding the subject they're teaching. According to the research, many teachers feel the level of English competency is different among themselves and they feel the need of assistance from native teachers from foreign countries.

However, it has been reported that Korean schools recently began to decrease the number of native speakers from foreign countries. [9] The first reason is the limitation in school budgets. According to the article, 48 million Korean Won (About 48 thousands in US Dollars) is spent annually per one native teacher, which is a burdening amount, especially for schools from small rural communities. The second reason is that the Korean educational bureau claims that many young

teachers now go through extensive English training courses before starting their career, negating the need for foreign teachers.

On the other hand, Lee (2018) points out several findings that contradict this claim. According to her research, English teachers having difficulties with their subject include those who went through the English training and have more than 3 years of field experience.

Also, According to Lee (2020), [10] a research on the perception of teacher and students about English speaking education and evaluation has indicated that native teachers overall did not show much discomfort in teaching their language, while non-native teachers felt they often had difficulties and burden on teaching how to speak in English.

2.3. Computer's Role in Prosody Education

2.3.1 CALL & ICT

CALL(Computer-Assisted Language Learning) refers to any process in which a learner uses a computer and, as a result, improves his or her language. [11] Computers are becoming an essential element in education and acquirement of foreign languages. [12]

Also, Yokokawa (2010) states about the impact of utilizing various forms of digital technology and media toward foreign language education in elementary schools, also called Information and Communications Technology (ICT) [13]. According to the research, incorporating ICT enables children to have greater interest and motivation toward learning foreign languages. ICT also help alleviating burdens from many teachers who have trouble or lack confidence regarding English in their class. However, Hirokawa also states that a considerable number of teachers dealing with ITC, even more so compared to teaching English.

2.3.2 The Benefits of Computers

According to Abusa'aleek (2012), [12] with the help from computers, classroom teaching become more effective. Language teaching in the past was conducted mainly in classroom environments with students' passively learning from teachers,

using blackboards and later with some video materials. Computers can present abstract ideas in ways more easier to understand. Many students who get bored and in traditional English classes become more motivated in learning the challenging language.

Also, computer enables for more individualized education. One of the problems in typical classroom educations is that in many cases, interests and skills toward English vary among students. Computers are able to offer different students different learning materials and methods, and allow students work at their own pace. Thus, they become the center of their learning, leading to more autonomous learning compared to classroom enviroments.

Computer also gives learners the option to study anytime and anywhere. Traditionally, learnings were limited in a fixed time and in a fixed environments (i.e. classrooms).

2.3.3 Prosody Conversion

A number of research including Felps et al. (2009), have argued that foreign language learners can benefit from listening their own voices that have prosodic features from the languages they are learning [14].

Computer Software For Prosody Conversions

Recently, a few CAPT(Computer-Assisted Pronunciation Training) systems have incorporated functions of converting prosody [14]. One software is WinPitch LTL by Martin (2020) [15], which allows users to re-synthesize their voices by manually editing the prosodic features. Other systems such as SpeedLingua [16] goes through computer algorithms to process learner's utterances. The particular method this thesis intends to focus on is one which involves a software named Praat.

Using Praat

Praat is an open-source software package written by Paul Boersma and David Weenink of the University of Amsterdam [17]. The software is primarily used by linguists for analyzing speech, and supports scripting for more efficient workflows

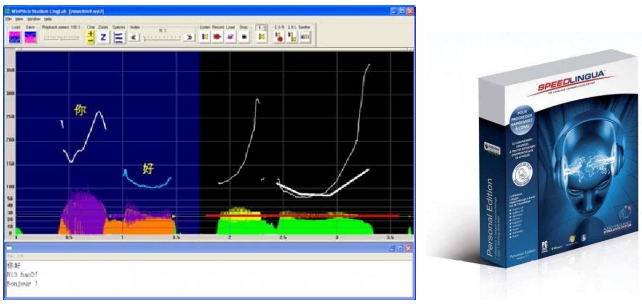


Figure 2.1 WinPitch LTL II & SpeedLingua [1]

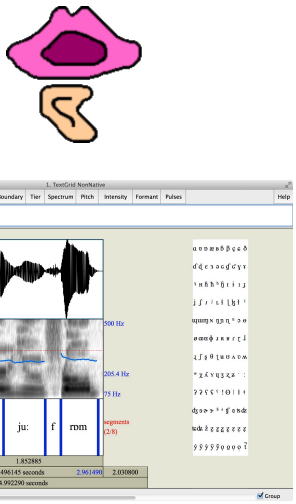


Figure 2.2 Praat

[18]. In addition, Praat is available as a library that can be used in programming languages, such as Parselmouth for Python developed by Jadoul et al. (2018) [19].

Yoon (2007), combined with a custom script, implemented an algorithm to handle prosody conversion [2]. According to Yoon (2007), through the algorhythm, if two different people record their voice speaking a same sentence, intonation and rhythm of a person's recorded voice can be changed to that of the other person. The process involves PSOLA(Pitch Synchronous Overlap and Add) technique [20], a feature is built in to Praat. This allows the modification of the prosody of natural speech while retaining a high level of naturalness.

The prosody conversion is comprised of three main steps. The sample phrase

"came in" will be used as the sample to illustrate each steps. In this situation, we'll assume a native speaker and a non-native learner have said the phrase in their own prosody. The three steps are depicted in Figure 2.3.

The first step is the alignment step. This process is very important, because this alignment data would be the basis of the two proceeding steps. The segments of voice samples of non-native speakers are manipulated either by stretching or shrinking each segments, so the position and length of the segments match those from native speakers.

In the second step, after the segments have been properly aligned together, the fundamental frequency[1] (often abbreviated as F0) from native utterances is imposed on non-native utterances. Changing F0 is possible because the segments of the two samples have been aligned together by the first step.

In the third and final step, The sound intensity contour of non-native utterances is changed to the contour from native utterances.

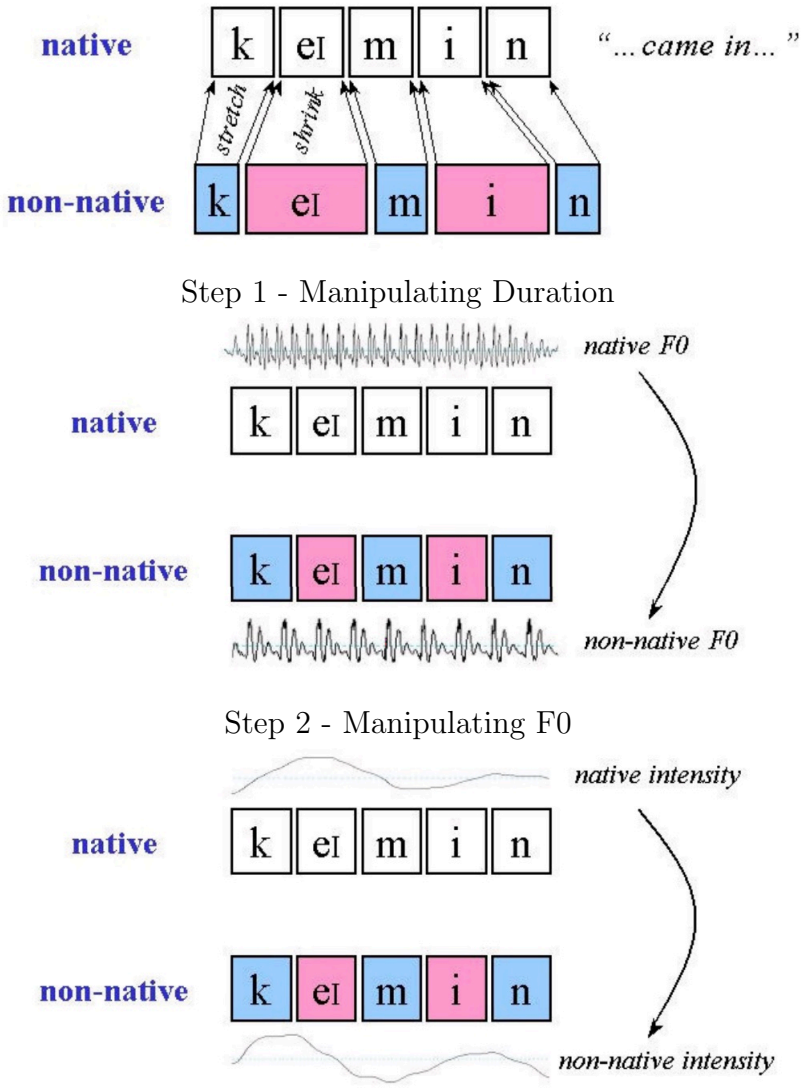
The method proposed by Yoon (2007) is well-known among researchers who are focused on prosody training methods for foreign language learners, such as Meo at al, (2013) [21] and Pellegrino at al (2015) [22].

2.3.4 Actual Applications on Prosody Education

Yoon et al. (2010) conducted a research on application of the prosody conversion method by Yoon (2007) on university students in Korea. The researchers had one group of English learners in Korean train prosody mimicking utterances from an English speaker, whereas another group was trained with their own voice [4]. Utterances from learners were evaluated both from before and after the experiments. The experiment has shown that learners trained listening to their voices were rated as having prosody more native than the other group. Another benefit Yoon et al. (2010) has stated is that by listening to their own converted voice, non-native learners are able to gain more motivation and confidence on learning English and its prosody.

Similar experiments have been conducted outside the scope Korea and English as well. Nagano and Ozawa (1990) [23] conducted a similar experiment on English

1 Fundamental frequency (F0) is defined as the lowest frequency of a periodic waveform.



Step 1 - Manipulating Duration

Step 2 - Manipulating F0

Step 3 - Manipulating Intensity

Figure 2.3 Steps of Prosody Conversion [2]

learners in Japan, with the results that are consistent with those by Yoon et al. Also, Bissiri et al. (2006), while investigating non-native learners studying German, has argued that, learning from learner's own voice (with correct prosody) is more effective form of feedback than utterances from a German native speaker. [24]

Limitations

However, as mentioned in the introduction chapter, after re-creating prosody conversion process suggested by Yoon (2007) and Yoon et al. (2010), this research has concluded that this method has some limitations to be addressed.

First, in order to deal with the conversion, recorded voices had to be manually broken down to segments and transcribed by hand. Therefore, although the prosody conversion by the Praat algorithm is much simpler process compared to other manual systems including WinPitch LTL, the process would be still difficult for ordinary English learners who do not have expertise in computers or phonetics.

Also, this prosody conversion process still required a human native speaker to provide voice to be used as the model for conversion.

2.3.5 Speech Recognition and Synthesis

The idea of machines emulating the way human behaves, especially in the area of synthesizing natural speech and responding to spoken language, has always caused a huge interest among scientists and engineers for a very long time [25]. The idea of speech analysis has been proposed since the 1930s by H. Dudley of Bell Laboratories.

In recent years, computer became more and more advanced in recognizing inputs from human voices. Voice assistants like Siri, Alexa, or Google Assistant are able to interpret commands and questions and make responses. Also, in recent years, many speech APIs are now able to get timestamps for individual words spoken by users. [26] Also, computers became sophisticated in recognizing inputs made by human voice. Especially in the last few years, machine learning and deep neural networks made speech synthesis much more advanced, such as WaveNet by Oord, et al. (2016) [27]

Utilizing speech recognition and synthesized speech for practicing intonation and prosody has been conceived since the early 2000's in the context of CALL(Computer-Assisted Language Learning). The use of TTS (Text-To-Speech) enabled teachers to create pronunciation exercises simply by typing in the orthographic transcription of the pronunciation models to be presented [28]. Likewise, Neri et al. (2003) discuss how speech recognition can be employed to develop reliable tools for CALL [29].

Notes

Chapter 3
Design and Prototypes

This chapter describes steps and process that led to the design of iTonation.

3.1. Pilot Research - Prior to Desigining iTonation

Prior to designing iTonation itself, the initial purpose of this research was to apply the prosody conversion method on singing English songs. As mentioned in Chapter 2, it has been established that listening to learners' own voice converted to have native prosody can be a benefit toward their prosody training. However, almost every existing research about prosody conversion focused on plain spoken words. Meanwhile, according to Hong (2004) [30], music and spoken language share similarity. Also, many learners enjoy songs, and they are memorable and easy to repeat, making it an effective material for studying prosody and pronunciation. Other works such as Matsuzawa (2009) [31] and Goto (2018) [32] are also emphasizing song as an effective method for learning prosody.

Base on the above, this research initially intended to apply the prosody conversion by the Praat script developed Yoon (2007) on singing voice from English learners instead of plain sentences.

3.1.1 Pilot Research on an Elementary School in Korea

Regarding this direction, a preliminary research has been conducted in an elementary school in Daegu, Korea on October 2018. The school had an English music festival, but many students and teachers who were supposed to sing at the event had trouble signing in English. The teacher in charge suggested to try this

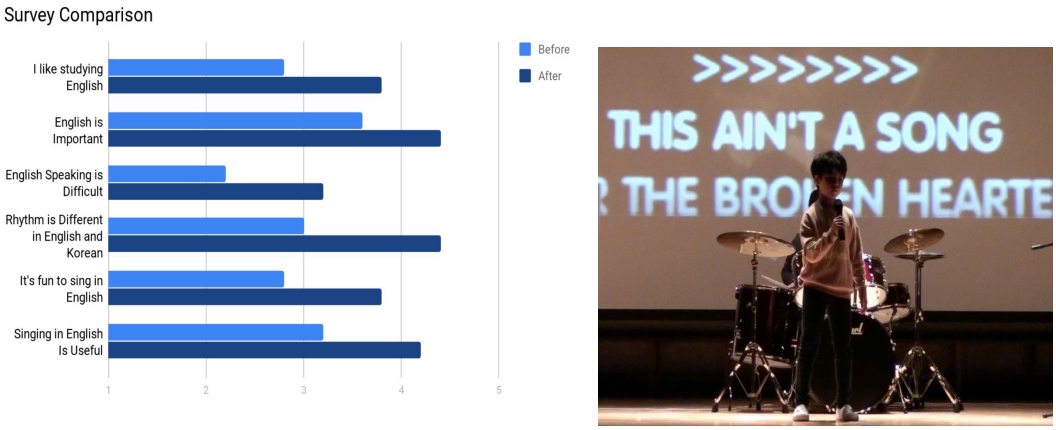


Figure 3.1 Survey & Children Singing At the Music Festival

method to convert their voices and let them hear it. Two students, along with the teacher herself volunteered and recorded their initial attempts to sing their songs. After the voices had been recorded, voice samples were processed by the Praat script to convert the pitch, stress, rhythm, etc. The songs that each of the participants wanted to sing were: 'It's My Life' by Bon Jovi, 'Lemon Tree' by Fool's Garden, and 'Rolling In The Deep' by Adele. As the model for prosody conversion, acapella versions for each of the songs were used since any other exterior noise would make the process impossible. For 'Rolling In The Deep', the researcher acquired a voice sample from the acapella version performed by Adele, the original performer. For 'It's My Life' and 'Lemon Tree' however, there were no acapella version available. The researcher created a synthesized versions of the song using VOCALOID software.

After the voice has been processed, the participants later heard their converted voice and continued practicing singing for the event. As result, the students were amazed to hear their own singing voice in perfect pitch and rhythm, and showed more motivation toward practicing singing in English.

A survey was also conducted on the students participating to the event which showed general increase in awareness on English songs and prosody.

However, due to the limited time frame, the research was unable to conduct

a comprehensive comparison of actual increase in proficiency in English prosody between before and after the event. One of the main reasons for this predicament was that the conversion process took considerable time, as it required manual transcription and alignment of voice samples, as stated in Chapter 1 and 2. Also, procuring model voice samples for converting learner's voices proved to be difficult than anticipated, requiring to use synthetic voices as the model in some of the songs.

As result, the main focus has been shifted toward creating more simpler, automated, and autonomous process for prosody conversion with speech recognition and synthesis, putting the topic of musical application on hold for future research. Nevertheless, the pilot experiment did show some degree of positive changes regarding motivation to speak in English and recognition in prosodic features in the language.

3.2. The Main Concept

Based on the preceding findings, the idea of creating more simpler process of converting learner's prosody by utilizing voice recognition has been conceived, as well as using synthesized voices as model for prosody conversion.

iTonation is an application intended to aid non-native English learners who do not have a clue how to properly speak in fluent prosody. For instance, a learner can take his favorite phrases from sources such as a movie or a book and read them aloud (as much as possible). The app will record the phrases and speak them back to the learner but with native prosody added. The learner can listen their converted voices as reference for practicing speaking in English prosody. The idea is that listening their own voice can be a better feedback or guidance

iTonation is based on following key concepts.

- **Practice Prosody Outside Typical Classroom Environments**

  iTonation aims for learners to able to practice English prosody regardless of time and place. A large number of English learners in Korea study the language in classroom environments, either through schools for students, private cram schools for others. However, in a classroom time is limited

for learners to learn English. An elementary school in Korea, for example, officially offers only two class hours per week for English. Moreover, learners are hardly exposed to English-speaking environments on their regular lives, making it difficult to practice outside of their classrooms. iTonation gives opportunity to practice prosody at home, to supplement what they learn in class.

- **No Complicated Procedures**

  iTonation intends to utilize speech recognition to automate the process required to change the prosody of recorded speeches. Existing methods for prosody conversion required careful transcription and alignment of voice samples from the learner and the model speaker. This may lead to some complications for learners who are not comfortables with computers and acoustics. For end users, this means they just need to press few buttons and iTonation would do the trick.

- **Requiring Less Help From Native Speakers**

  Traditionally, English learners in Korea practiced prosody using words spoken by native speakers as reference or model. Although it's been discovered learner's own voice can be a better reference and feedback, the current algorithm for prosody conversion still requires native speakers to provide voice samples to function properly. iTonation intends to eliminate this requirement altogether by opting to go for synthesized speech as the model for prosody conversion process.

- **More Motivation for Learners**

  iTonation can also bring positive changes to motivational aspects toward learning English. Being able to hear their voice in fluent English can be a boost for morale and motivation toward learning the new language.

3.2.1 Target User

This thesis intends to focus on measuring the application's effect on elementary school students in Korea. As mentioned already, a large number of English learners in Korea are school students, and in many cases, elementary school is where

they begin to learn English for the first time. Additionally, in Korea, there's an extremely high interest among parents regarding their children's English education, making young English learners and their family a huge potential user base.

3.3. Usage Scenario

This research suggests using iTonation as a supplementary mean along with regular English classes (especially speaking classes) taught at elementary schools in Korea. In Korea, English classes are mostly taught by teaching key expressions and sentences which are selected by the national curriculum as necessary for improving children's communication skills. Accordingly, the names of chapters in English textbooks are those key expressions. For instance, in an English textbook for 5th graders, the title of chapter 1 is "Where are you from?". This research presents the following scenario for children to practice speaking those key phrases at home.

- **Stage 1 - At school**

  In classes, students learn to speak the key phrase from the chapter they're on. After a class has finished, their teacher may give them a homework. The teacher might instruct them to practice speaking phrases they've learnt during their class.

  This research suggests that it may be possible to utilize iTonation in this situation. The teacher may install the app on devices like laptops or tablets and lend them to the children. The teacher may tell them to practice speaking the phrases, while listening to their own voice converted with native intonation.

- **Stage 2 - At Home**

  First of all, when children return home after school, they can open iTonation app they received from their teacher, and they record the sentences they've learned in school. After the voices have been recorded, iTonation will convert the voices, making them sound more like native speakers with

native prosody. Children can keep practicing speaking those phrases, trying to imitate the change in intonation and rhythms presented in converted voices.

- **Stage 3 - Back at school**

  The next week, when children are back at their English classes, and their teacher may review the previous chapter before moving onto the next. While doing a review, the teacher may let children speak the phrases they've learned at the previous lesson and practiced at home. The teacher may evaluate utterances by children and may give a prize for the student who've been much improved, which may further motivate other children to practice using iTonation.

3.4. Design Iterations

3.4.1 UI Prototype

Designing UI

The primary idea behind designing user interface of iTonation is making it simple to operate. The idea was to split the user interface into multiple sections. Since iTonation is comprised of multiple components, jamming all those elements into one window would confuse users, especially for children. Furthermore, those key elements has to be executed sequentially to function properly, so the idea of breaking down into multiple sequential sections fits the overall UI design of iTonation.

Mockup UI Prototype

Figure 3.2 represents the initial mock-up prototype presented at the interim presentation held internally in Keio University at end of January 2020. As mentioned earlier, the UI had been designed to be based on multiple sequential pages similar to Kamishibai. Initially, the app's name was 'Morphsody' as it can be shown in the figures. However, the name later changed to 'iTonation' to give

clearer idea to users that the app is heavily related to modifying intonation of recorded voices.

When a user opens iTonation, in the first page, the app briefly explains what does the app do, and prompts the user to proceed. The next page is the recording section, where the user would press the record button to record his or her voices. As the user move to the next page, the app would process the recorded voice samples, recognizing words spoken, as well as timestamps for each of them. Then, the recognized texts would be displayed on the next page, and the user can check if the recognized texts are correct, as well as checking his or her recorded sound. In the next page, the user can choose which specific English intonation he or her wants her voice to be converted into. The app then would synthesize a voice sample with native prosody to use as the model for prosody conversion, based on texts that have been recognized from the user's voice. Finally, the app displays the original voice along with the converted one for the user to compare them.

3.4.2 Functional Prototype

Components

- **Python**

  For writing iTonation, Python has been chosen as the primary programming language. This is primary due to its versatility and short time required to write applications. However, in the future, the researcher may consider porting the app to other languages like Swift or Java to be used in mobile systems such as iOS or Android.

- **Voice Recording**

  In order to handle recording voice from users, iTonation utilizes the PyAudio library to handle voice inputs. The original intent was to fully implement the UI with a button for toggling recording mode on or off, along with some other buttons for pausing and playback as well. However, due to the time constraints, by the time of the first user test, interface became more primitive. During the first day, there's only one input field for deciding how long the audio would be, and a record button to record the voices for

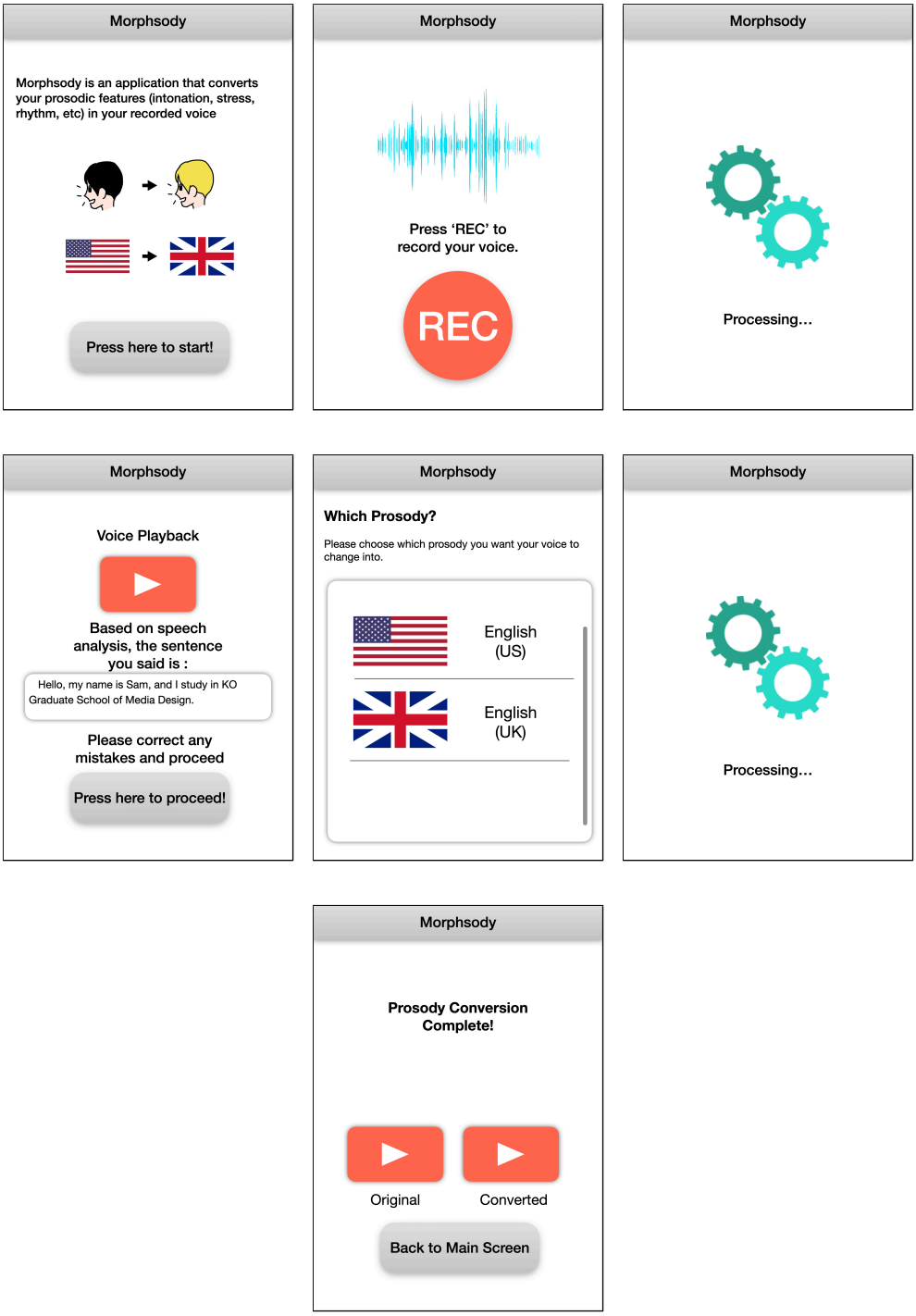


Figure 3.2 Mockup Prototype

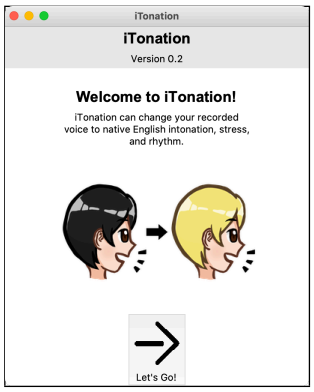


Figure 3.3 Functional Implementation Using Python

the duration set by the user. However, in the second day of user test, the prototype was able to feature a toggle button to initiate and stop recording voices. An image was also placed inside the button to better illustrate the function of the record button for children. After children finish recording, The file is saved as WAV format, and the data is handed over to IBM Watson for speech recognition.

```
FORMAT = pyaudio.paInt16
CHANNELS = 1
RATE = 44100
CHUNK = 1024
RECORD_SECONDS = int(input("Please determine the duration of your recording\n(in seconds (natural number)) :"))
WAVE_OUTPUT_FILENAME = audio_path = path.join(path.dirname(path.realpath(__file__)), "nonnative/NonNative.wav")

audio = pyaudio.PyAudio()

# start Recording

stream = audio.open(format=FORMAT, channels=CHANNELS, rate=RATE, input=True, frames_per_buffer=CHUNK)
print ("recording...")
frames = []

for i in range(0, int(RATE / CHUNK * RECORD_SECONDS)):
    data = stream.read(CHUNK)
    frames.append(data)
print ("finished recording")
```

Figure 3.4 Code For Voice Recording

- **Speech Recognition Engine**

  For handling speech recognition, IBM Watson has been decided as the primary engine. IBM Watson has several key advantages over competing speech recognition libraries like Google or CMU Sphinx. First, its speech recognition has a high accuracy, not only for detecting words, but also timestamp for each words, which is crucial for iTonation. Although other libraries were

able to detect words accurately, while testing, timestamp detection was substantially limited compared to IBM, making the resulting audio highly unsatisfactory to listen. Additionally, IBM Watson was a perfect choice for the experiment because, although with limitations, it offered a free tier that did not require additional charge or limitations of a trial period.

Figure 3.5 repesents the Python code written to handle the authentication toward IBM Watson for speech recognition. After the credentials have been authorized, IBM Watson would load the audio files generated by the voice recording section of iTonation, and recognize the words along with timestamps.

```
#IBM Authentication

authenticator = IAMAuthenticator([redacted])
speech_to_text = SpeechToTextV1(
    authenticator=authenticator
)
speech_to_text.set_service_url([redacted])

def watson_nonnative() :
    p002_label_pleasewait.grid(row=5)
    audio_path = path.join(path.dirname(path.realpath(__file__)), "nonnative/NonNative.wav")

    with open(audio_path, 'rb') as audio_file:
        result_nonnative = speech_to_text.recognize(
            audio=audio_file,
            content_type='audio/wav',
            timestamps=True
        ).get_result()
    str_nonnative = str(result_nonnative)
```

Figure 3.5 Code For IBM Watson's Speech Recognition

If IBM Watson successfully analyzes the voice samples, it would return the value as a series of text strings. However Praat could not understand the the raw output generated by IBM Waston and has to be run through a couple of filters written by the researcher, mostly based on text replacing features built into the Python language.

One of the concerns during writing code was the fact the authentication code has been input directly to the main Python code. Should the source code leak or reverse-engineered, it'd be a substantial risk from the perspective of security and IBM Watson being misused. Another concern is that IBM tend to revise their output format for speech recognition time to time, which

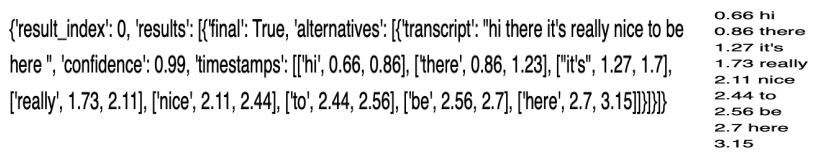


Figure 3.6 IBM Watson's Output (Raw & Converted)

unfortunately occurred during the development of iTonation prototypes, and the codes for iTonation had to be modified accordingly.

- **Speech Synthesis**

  This research choose Pyttsx3, the text-to-speech engine for Python, for handling synthesizing speech for model prosody. Pyttsx3 allows Python to tap into the speech synthesis modules that are embedded inside operating systems. In this case, since iTonation has been installed on the researcher's MacBook, Pyttsx3 used macOS's speech functionality.

  For the purpose of prosody conversion, the researcher had Pyttsx3 utilize two different voices from the local system speech library installed on the researcher's laptop (a MacBook) used during the research. The reason for importing two voices is that every person makes speech in different ways. The difference is especially distinctive between male voices and female voices. For instance, when the researcher (male) tested prosody with his own voice while using female voices as the conversion model, the pitch from the converted voices was raised too high that the they sounded unnatural.

  Therefore, for the research purposes, one male voice engine and one female voice engine have been selected.

- **Praat - Prosody Conversion**

  The prosody conversion, which is the essential component of the app, is handled by Praat. As mentioned in Chapter 2, a script written by Yoon (2007) converts the prosody of spoken words by nonnative learners, based on the data inputs generated from the previous two components.

  Unfortunately, prosody conversion part (Praat) had to be called externally

from the software itself to handle the conversion, rather than properly integrating into the main software, making the process less straightforward than this research has intended. There was a library called 'Parselmouth', which is basically Praat that have been integrated as a Python API. However, due to the limited time frame, this research was not able to secure enough time to port the existing scripts to handle prosody conversion to Parselmouth. As result, additional app page had to be created to create button to call Praat functionalities, potentially increasing the complexity toward users, especially children.

Based on the preceding elements, the researcher was able to write a functional prototype. First, as mentioned earlier, iTonation gathers inputs from a user (i.e. voice samples) and goes through the speech analysis process by IBM Watson. IBM Watson outputs the recognition data with timestamps as text strings, which are converted by a Praat script developed by Mietta (2017) into a form that the application can understand to conduct prosody conversions. [33] Then the Pyttsx3 library combined with macOS's speech capabilities produce synthesized voice samples speaking the same phrase as the user. Synthesized voices also go through the same process of speech recognition and data conversion process. Finally, the data from the two voice samples are compared and Yoon (2007)'s script finally handles the prosody conversion process.

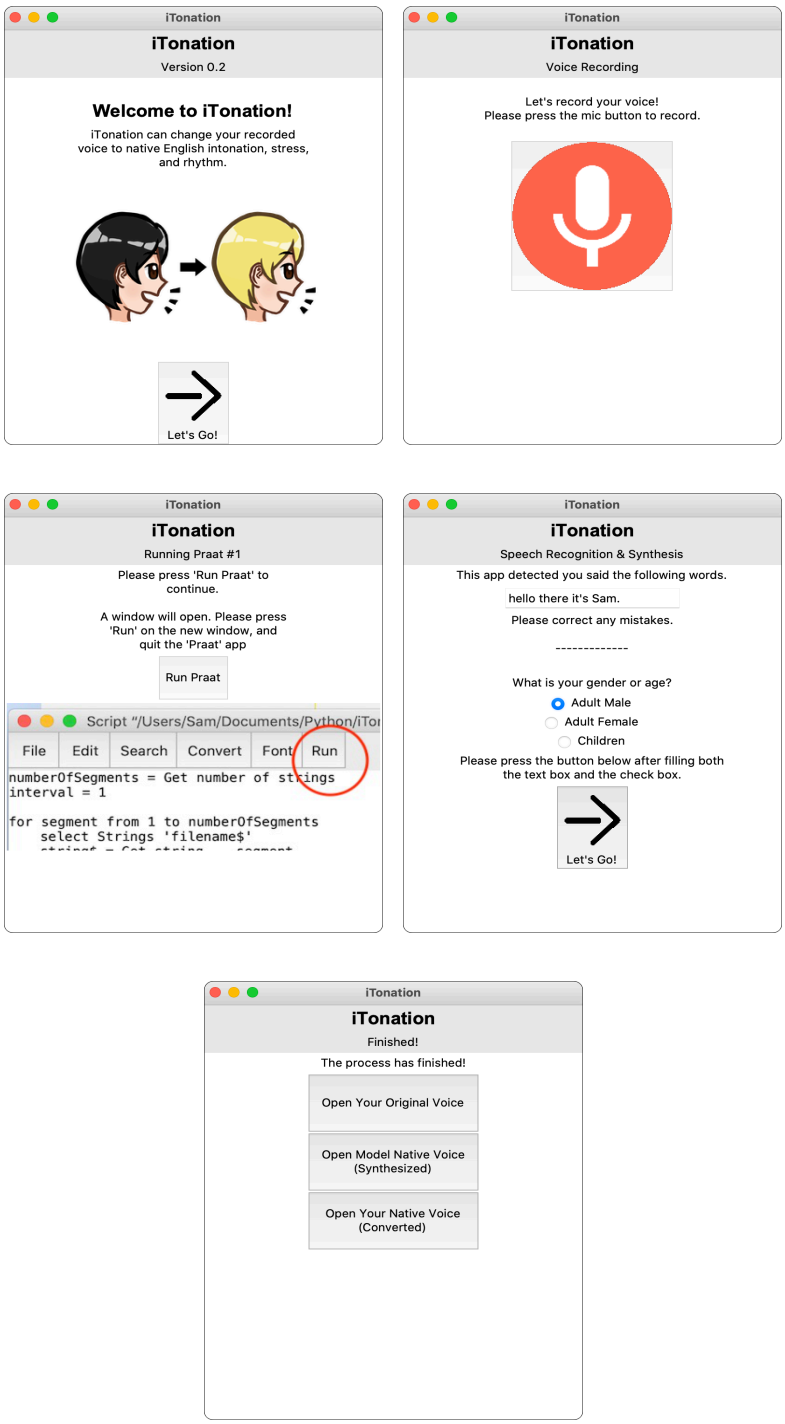


Figure 3.7 Screenshot of the Functional Prototype

Chapter 4
User Tests & Evaluation

This chapter describes the user tests designed to evaluate the concept of iTonation and the analysis of its results. The user test last from June 25th to 26th, 2020.

4.1. User Tests Overview

4.1.1 Focus

Regarding the effectiveness of the prosody conversion method, Yoon (2010) has already shown that listening to their voice converted with native English prosody can provide better method of practicing prosody. Therefore, the focus of this user test is to evaluate whether children would be able to conduct the prosody conversion and practice autonomously, through the use of iTonation, combining the speech recognition and synthesis.

4.1.2 Participants

The researcher had an opportunity to meet four children from a local elementary school along with a teacher from the same school. This thesis will refer to each children as Child A, Child B, Child C, and Child D from now on.

All of the children were female, and they were 11 years old who belonged in the fifth grade of the school. The teacher used to be their homeroom teacher in 2016. The user test took place at the teacher's house where the participants were invited to visit.

During the user tests, the following characteristics have been observed from the four children.

- **Child A** : She seemed to have a pessimistic kind of personality. Although she complied every instructions during the user tests, she showed the relative lack of enthusiasm compared to the other three. While she did agree to join the test, she commented she's not good at English and have aversion to it.
- **Child B** : She showed an introverted characteristic, but not as pessimistic as Child A.
- **Child C** : She showed a highly optimistic and outgoing personality. She participated in the user test with the most enthusiasm among the four children.
- **Child D** : According to the teacher, she is a quiet and thoughtful type of person.

4.1.3 Methods

The user tests were generally comprised of two major steps. First, certain English phrases were provided to the participants. Children read the phrases and briefly practiced speaking the phrases until they were ready to record the sentences. The researcher told the participants that they don't have to be able to speak the sentences perfectly. Their speech only needed to be audible for the speech recognition system, as any phrases spoken too fast or slow would be corrected by the prosody conversion system.

Second, after they have practiced enough, they actually ran the iTonation prototype installed on the researcher's laptop. In most cases, the participants were told to try running the app by themselves, unless they were really stuck on what to do. While running the application, they recorded the phrases they were to speak and clicked the buttons to initiate speech recognition, synthesis, and the prosody conversion process. Finally, the application generated the converted voices, and the participants compared to their original recordings.

After the activity were concluded, a series of rating and interviews on the experience during the user tests had been asked to the participants to fill in.

4.1.4 Materials

The phrases provided were based on the actual English textbook they study in their school. [34] The phrases were: "Where are you from?", "What do you do on weekends?", May I take a picture?", and "Whose book is this?". These are the main phrases learned during chapter 1, chapter 2, chapter 3, and chapter 4 from the textbook. The specific dialogue sheet presented to the children is included in the appendix.

Figure 4.1 English Textbook

4.2. User Tests Results

This section explains about the actual activities the participants have went trough, along with the observations the researcher was able to gather from the participants.

4.2.1 June 25th - Day 1

On the first day of the test, the teacher gave a brief explanation what the researcher was about to do for the day and asked the children whether they'd like to join. The researcher had explained that the research has been about how to make English more easy and fun to practice using computers, and they agreed to participate.

As mentioned in the methods section, before running the prototype of iTonation, the researcher handed out the actual sentences they were about to record. The

children practiced speaking the phrases for some time before moving on. This stage is closely related to the first step in the proposed usage scenario from Chapter 3, which is about children studying about a certain phrase in class before trying to record their voice at home.

After practicing the phrases, the children sat behind a computer with the iTonation prototype installed. The recording and conversion session was done for one child at a time. While one participant was doing the process, the others were watching from behind.

As for the second step, after they've had enough practice, they sat behind a computer and recorded their voices speaking the phrases. After they finished recording their sentences, iTonation had converted the voices samples, based on the synthesized speech speaking the same words. Later, iTonation played the converted voices back to children.

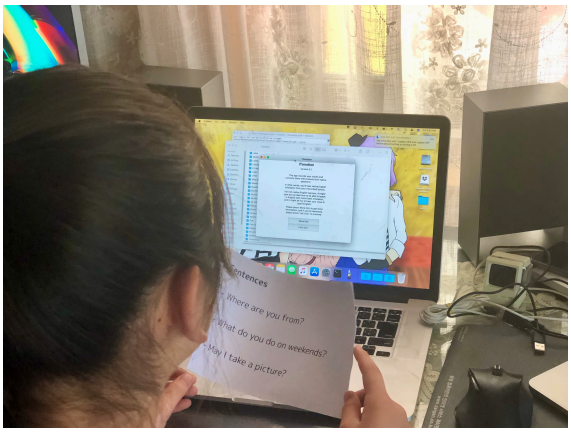


Figure 4.2 Children Running iTonation - Day 1

During the the first day of the user test, iTonation was able to successfully recognize the sentences spoken by the participating children. Speech synthesis was successfully processed as well, producing the speech samples for the same sentence the children have spoken, based on the speech recognition system. As result, based on the two components, the Praat component was able to successfully manipulate rhythm, stress, and intonation contours of voices recorded by the children, just as it would under the conventional prosody conversion methods without speech recognition and synthesis. The graph in Figure 4.3 shows the sound waves and

corresponding words from one of the phrases spoken by one of the students (Child C) during the first day.

While the converted voices were played back one by one (as the app had been run by one child at a time), All of the four children were listening in when each child played back the converted voices. Upon hearing their converted voices, the room was filled with laughter and giggles. Some children (Child B and D) expressed a minor embarrassment when hearing her own voice, which was understandable and expected. Nevertheless, they did join the giggles when listening to the voice of others, indicating their response may be a positive one nonetheless.

On the other hand, upon hearing the children's converted voices, the teacher has commented that the voices are too fast for children to properly listen to. This was caused due to the speaking rate of the synthesized voices too fast, therefore children's voices have ended up being converted based on that fast rhythm. The workaround was to make the synthesized voice slower by modifying the parameter from the iTonation's source code. After the synthesized voice has been slowed down, participants felt the converted voice's speaking rate went down to the acceptable level. The teacher further pointed out that rather than the voice's prosody being changed, she felt voice itself had been swapped to the synthesized voice, noting that the resulting voice sound rather robotic. One of the children (Child A) made a similar comment; while it might be a fun idea to hear own voice sounding dramatically different way, she expressed concerns that the converted sound didn't sound like their's than they anticipated.

The researcher speculated that the possible reason for this is that the inappropriate voice synthesis library had been used for the conversion. As mentioned in Chapter 3, iTonation included one male voice and one female voice, in order to be able to use for both of the genders. The researcher assumed that female voice samples could be the model for prosody conversions for little girls as well as women. However, feedbacks from the teacher indicated that this is not the case, creating the necessity for preparing a synthesized voice engine designed specifically for little children.

Regarding the UI of iTonation, due to the application window broken down to multiple and sequential sections, the participants were able to proceed with most of the procedures without much trouble. However, during the first day, the UI was

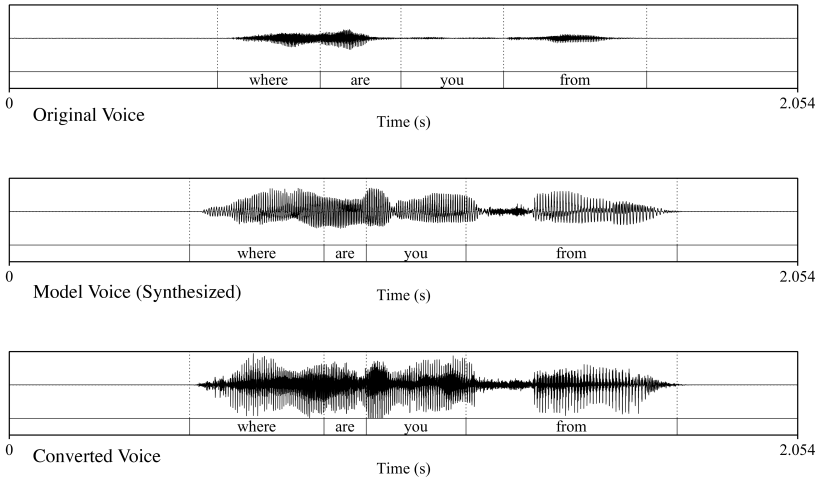


Figure 4.3 iTonation's Result (Child C - June 25th)

not completely finished by the time of the first day of the tests; necessary image data was not imported in the UI, and the app only managed to show white screens with some texts and buttons. Understandably, the participants have unanimously remarked that the UI looked bland. The researcher explained that UI was still work in progress, and the UI element would be introduced in the next day of the user test.

After the first day's user test has been ended, the children and the teacher agreed to conduct an another test the next day.

4.2.2 June 26th - Day 2

After the test on the first day had been finished, the code of iTonation had been modified. Regarding the problem of the converted voice not sounding 100% like the children's own voice, the researcher attempted to acquire a new voice library to be used with children.

However, the challenges was that there are few voice system specifically designed to sound like children. Most of the voice synthesizer engine were available either as adult male for female. Therefore, a makeshift fix had been implemented. The pitch of the existing female voice from the speech synthesis engine was raised so

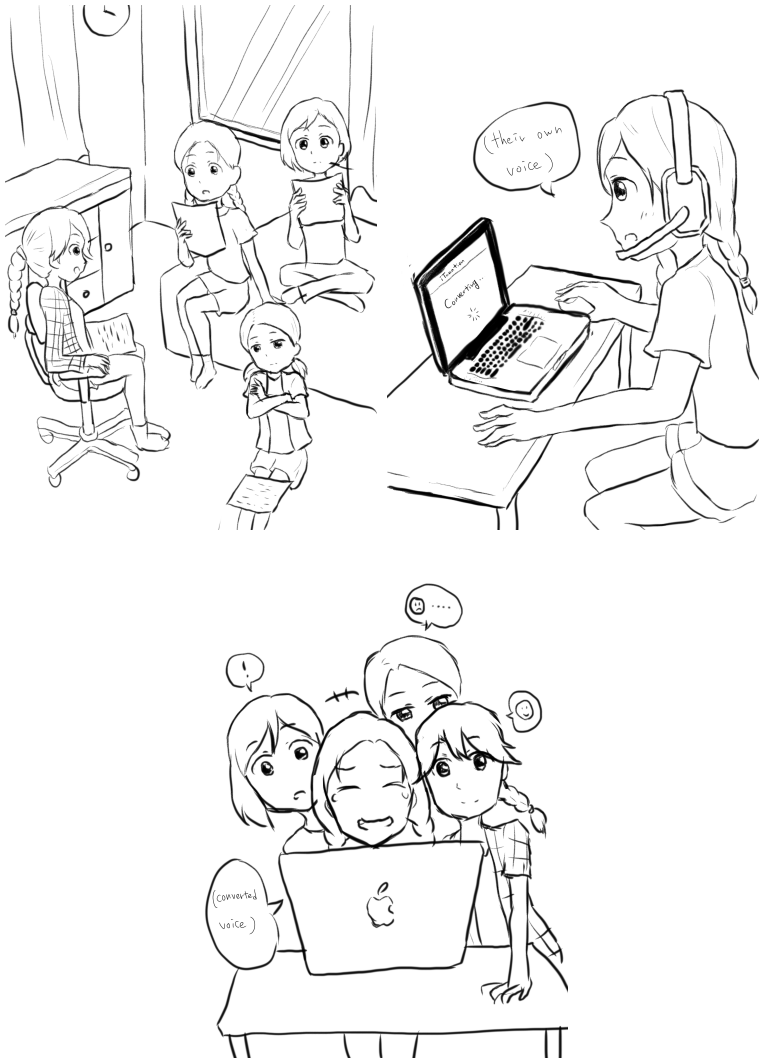


Figure 4.4 Illustrations of the Children During the User Tests (※ Note - Children insisted against posting actual photos of their faces.)

the voice would sound more like children.

Based on these fixes, the user test has been resumed with the same children and the teacher. After the second day of the user test has been concluded, students heard the voices as they did on the first day. While the response upon hearing the voices were more positive than the previous day. However, while they remarked that the converted voices became closer to their own voice this time, Child A remarked that the voice sounded still too robotic and she's not too pleased.

For the UI elements, the improved version has been presented to the participants, and the response was much favorable compared to the first day. The biggest improvement is some of the text being removed, replaced with simpler images. As result, the participants were able to proceed with most of the procedures without too much trouble. The time required to go through one process for a child has been reduced as well. It took around two minutes to finish the conversion with the crude prototype in the first day. The second day's prototype took around 1 1:30 to finish the conversion.

However in some sections, children were still mildly confused how to proceed. The part in question was related to invoking the Praat app. (The third image in Figure 3.4 Screenshot of the Functional Prototype) In order to use the raw data created by IBM Watson ' s speech recognition for prosody conversions, the data had to be converted into a format of which the conversion component (Praat) could understand. As mentioned in Chapter 3, this is done by a script created by Mietta (2017). Therefore, when the children reached the section, a Praat screen had appeared prompting to click the run button. Children were a little confused at the first time, but I told them to click the button and they were able to proceed. This process had to be repeated twice, one for the non-native speech by children, and the model speech generated by the application.

4.2.3 Rating By The Children

After the user test in the second day, the researcher asked the children to rate their experiences. The rating sheet handed out to the children comprised of 6 questions, and each question were answered in a scale of 1 to 5.

- **Q1**. Do you usually enjoy studying English?

- **Q2**. Do you think studying English is important.
- **Q3**. Is it pretty hard to be good at speaking English.
- **Q4**. Do you think intonation and rhythm are very different between English and Korean?
- **Q5**. Was it interesting to hear your voice in a different way?
- **Q6**. Would you like to see similar methods on actual English studies?

The children responded to the question in the following way:

Children	Q1	Q2	Q3	Q4	Q5	Q6
Child A	3	4	2	3	2	1
Child B	3	3	2	3	5	5
Child C	3	3	3	3	4	4
Child D	3	5	4	3	4	3

Table 4.1 Rating Score by Each Children

- **Child A** : For the questions about the English language itself (Q1,Q2,Q3,Q4), she mostly gave the average ratings (2∼3 out of 5), and for the Q3 (English Difficulty), she surprisingly marked 2 out of 5. However, as mentioned above, she commented about her aversion to speaking in English. This may indicate that English they learn at school may be easy, but she may lack the confidence when it comes to using the language in real life. This may be the reason for the lowest rating among the four children regarding Q5 and Q6.
- **Child B** : She gave the highest rating for Q5 and Q6, both scoring 5 scores out of 5.
- **Child C** : Contrary to being the most enthusiastic during the user test itself, Child C may have been the most haphazard when responding to the rating sheet. She first answered the every questions '3' in less than 5 seconds, and after the teacher remarked she answered too carelessly, she got a new sheet

and marked it again. Nevertheless, Q5 and Q6 was scored 4 out of 5, and combined with the fact she actively participated in the user test, she seemed to be positive about the prototype.

- **Child D** : Among the children, she got the highest score on Q2 (The Importance of English). While she gave the score 4 in Q5 (Impression on Converted Voices), in Q6 (Desire to Use iTonation on Actual English Practice), she scored 3, relatively lower than other children but still higher than Child A.

4.2.4 Interview

Aside from the rating scores measured by the children, a series of interviews have been conducted on the participants. The purpose of the interviews were to gain in-depth and specific comments on the experience during the user tests.

Interview With Children

The primary questions asked during the interview was as following:

- **Quesion 1** - Upon listening to your own converted voice after recording English sentences, what was your feeling like?
- **Question 2** - Do you think it's a nice idea to use this app on studying English? What are your thoughts?

The first and second interview items share the same questions from the Question 5 and Question 6 from the rating sheet handed out earlier, albeit inquiring for more in-depth respones on the two questions.

During the interview, the children provided the following responses.

- **Child A**

  Question 1 - "The idea of converting voices did sound interesting. It was a mystery how such things is even possible. But the converted voice sounded different than my own and felt awkward."

Question 2 - "The cram schools I used to go taught English centered around exams, and did not do speaking much, so it might be good to have a tool or an app that puts more emphasis on speaking."

- **Child B**

  Question 1 - "It was really interesting. English classes have a lot of singing sessions, but I'm not really good at singing. So I wish I had a similar app which converts my singing voice to a better version."

  Question 2 - "Rather than getting better at exams, I really want to be better at speaking English. I really wish I had more change at practicing speaking. So it'd be really nice to have an app for practicing English songs."

- **Child C**

  Question 1 - "It was really shocking to hear my voice speaking like a native speaker. I could not help laughing to hear my friends' voice in that way. I wish I could save friends' converted voices on my phones.

  Question 2 - "I really think that this app should be available on mobile phones. We all carry smartphones 24/7, so it'll be nice if this app can be installed on our phones. I think I can use this during when I'm studying English."

- **Child D**

  Question 1 - "First, listening to my own voice like that was a bit embarrassing. Still, it was interesting to hear my voice sound like a native person. Also, it was funny to hear my friend's voice."

  Question 2 - "I do think that it's important to learn English, but it's always hard to memorize English phrases because it's not used often in daily life. Right now, if I see a foreigner, I would just run away. So it'd be really nice to have a tool I can use to practice English that I can use anytime and not get bored."

Interview With Teacher

In addition to the interview toward children on the experience of iTonation, an another interview has been conducted on the teacher as well. The teacher ' s interview is centered around the quality of the voices generated by iTonation.

The questions in the interview are : "Overall quality of the voice", "Speaking Rate", "Resemblance compared to the original voices by children.", "Pronunciation Clarity", "Audio parts that are uncomfortable to hear", "Degree of interest in prosody of children's English", and "Potential of using iTonation from teacher's point of view"

The interview questions are originally based on the mean opinion scale (MOS) survey method suggested by Mahesh Viswanathan et al. in 2005. The survey is intended to take survey from multiple participants and gather the mean score to be used as a quality index of a voice synthesis system. However, since this research is conducted on an extremely limited number of participants, it is not going through quantitative studies, and the questions from the survey by Viswanathan et al. has been converted as the interview questions for teachers.

- **The overall voice quality**

  As mentioned during the user tests, although the teacher acknowledged that the prosodic features like pitch, rhythm or stress from recorded voices has been changed through the use of iTonation, due to the voices sounded less like the children's own voice, she noted that voices generated by iTonation may not be 100% perfect at the moment.

- **Speaking rate**

  During the first day, she initially felt the converted voice were too fast to be properly heard by children. However, after slowing down speaking rate of the synthesized voices, the converted voice also managed to slow down as well. After the voice has been slowed down, she said the speed might be adequate and made no particular comment about the speaking rate afterward.

- **Voice Resemblance**

  This is perhaps the biggest criticism during the user tests; she noted the voices did not sound like the children's own than they had hoped.

- **Clarity of Pronunciation**

  The teacher did not make any particular points with pronunciation clarity. This is because iTonation do not interfere with pronunciation made by children. Although it's true that computer generates the model voice for prosody conversion, only prosodic features are applied to the modification of children's voices, therefore any unclarity with pronunciation is caused by children themselves.

  However, the teacher did note that iTonation should be utilized alongside with teaching pronunciation, since prosody would not have much a meaning if children are unable to properly speak the English words in the first place.

- **Audio parts that are uncomfortable to hear**

  This may be a similar issue with the first and third question; the teacher mentioned that some parts of the converted song sounded considerably robotic, making them somewhat uncomfortable to listen to.

- **Interest in prosody of children's English**

  As a long-time teacher, she've always been interested in making learning experiences more fun for children. As for English, she mentioned that elementary school teachers are always having trouble with teaching children English.

- **The Potential of iTonation**

  Although she pointed out several flaws with iTonation's prototype, she mentioned it may have its potential use. One particular area she insisted on applying iTonation is music. She was present at the pilot research mentioned in Chapter 3 that involved an English music festival in October 2018. Out of the experiences from the event, she strongly suggested to keep working on the research from the perspective of musical application.

4.3. Summary

From the user tests, this research was able to gain the following findings.

First, the application was able successfully recognize the words spoken by the children and converted the prosody, using the speech samples from the built-in speech synthesis engine.

Because of this, the children were able to operate the prototype by themselves without the assistance, except for some sections which had technical difficulties (Mostly Praat-related). During the second day, the improved UI element has been presented to the participants, which made it more faster for the children to use the app compared to the previous day.

During the interview and user rating, while three children responded positively upon listening to their converted voices, one of the child responded with lesser enthusiasm. The negative response from one of the children may stem from her aversion to English studies. Also, when asked about whether they'd like to use this app on their English studies, the response had been greatly varied.

Thus, while iTonation made it possible for children to practice prosody only by themselves, unlike Yoon (2007) and Yoon et al. (2010)'s claims, using the prosody conversion method does not automatically enhance the motivation for learners to practice more prosody as well, at least for young children. It may be desirable to adopt additional incentives in addition to using iTonation to boost children's desire to know more about prosody. One suggestion came up by one of the participants was to apply this technique to songs rather than plain spoken words.

Chapter 5
Conclusion & Future Works

5.1. Conclusions

Based on the findings during the user test, the research was able to deduce the following conclusions and answers to the primary research goal described in Chapter 1.

The ultimate goal of creating the application iTonation is to provide a tool that enables young children to practice prosody even when they're not with their teachers or native speakers, by listening to their own voice converted with native English prosody.

The conventional methods for prosody conversion were possible by a careful and manual analysis of voice samples by an experienced individual, as well as requiring a human native speaker to provide the model voice sample for prosody conversion. However, using speech recognition and synthesis modules, iTonation has accomplished the prosody conversion without the need for external help, allowing children to practice prosody autonomously.

For the development of the prototype, the following components have been used. First, Python is used as the primary programming language. Next, IBM Watson has been chosen as the module for handling speech recognition. Also, for the speech synthesis module, Pyttsx3 wrapper For Python along with the native speech engine installed on the laptop used for the prototype (an Apple MacBook in this case). Lastly, Tkinter has been used to handle the UI programming.

During the user tests, the participating children were able to use the app start to finish with only minor technical issues, and the voices were successfully converted through the two aforementioned speech modules. Furthermore, when the children heard the converted voices, they reacted positively. However, one of the children did not respond to the voices with much enthusiasm. The reason may be related

to her apparent dislike with English. Also, when asked about the question on whether the children would like to use the actual app on their English studies, the responses were rather varied.

Thus, while iTonation did provide the basis for autonomous training of English prosody for children, this may indicate that learners listening listening to their own voice with native prosody alone may not automatically improve the motivation, at least for children. Therefore, it may be necessary for schools and teachers to adopt a supplementary incentive along with utilizing the application.

5.2. Limitations

Although this research claims that iTonation may potentially be an effective app for practicing English prosody, this research also left a substantial numbers of limitations to be solved in future works.

First, due to the limited time frame, Python was selected as the mean to write the most part of iTonation. However, while Python enabled the researcher to write programs quickly, it had its limitations when it came to mobile systems. A large number of digital devices used today are mobile devices like smartphones or tablets, and children are no exception: many elementary school students in Korea have smartphones of their own. Unfortunately, through python, it's difficult to write native apps optimized for mobile usages. In the future, it may be necessary to port the Python codes to other languages to continue on with this research.

Another issue is that at the current stage, iTonation is little more than a non-standalone wrapper that does not work on its own. Moreover, two Praat scripts, Yoon (2007) and Mietta (2017), had to be called externally, causing a minor hiccups during the user tests. Since it is possible to integrate the Praat functionalities into Python language using a package such as Parselmouth by Jadoul et al. (2018), so the Praat scripts used for the current prototype may have to be reverse-engineered to be integrated into Python or any other kinds of programming language used in future development.

5.3. Future Works

Also, should the development of iTonation reach the fruition in the future, the researcher intends to expand to the other areas like music. For instance, as described in Chapter 2, it has been discovered that the prosody conversion method can be applied to converting the pitch and rhythm from recorded songs. The teacher and one of the participants from the user test also pointed out users listening to their own singing voice but in much better pitch would certainly motivate users to use iTonation. Since most of the elementary English textbooks in Korea has a singing session in every chapters, she mentioned combining prosody conversion and music is highly beneficial English classes in schools. Therefore, should iTonation be successful in the future, it'd be an honor to collaborate with other researchers on the subject of prosody conversion or music.

References

[1] Amazon UK. Speedlingua personal edition (pc). http://www.amazon.co.uk/SpeedLingua-5453002600009-Personal-Edition-PC/dp/B002HDL340, 2010.

[2] Kyuchul Yoon. Imposing native speakers' prosody on non-native speakers' utterances: The technique of cloning prosody. *The Journal of Modern British & American Language & Literature*, 25(4):197–215, 2007.

[3] Mahesh Viswanathan and Madhubalan Viswanathan. Measuring speech quality for text-to-speech systems: Development and assessment of a modified mean opinion score (mos) scale. *Computer Speech and Language*, 19(1):55–83, January 2005. doi:10.1016/j.csl.2003.12.001.

[4] Kyuchul Yoon, Anna Heo, and Sangcheol Ahn. Teaching english intonation through learner utterances with cloned native intonation. *Modern Studies in English Language & Literature*, 54(1):141–171, February 2019. URL: http://www.kci.go.kr/kciportal/ci/sereArticleSearch/ciSereArtiView.kci?sereArticleSearchBean.artiId=ART001424250.

[5] KyungYim Lee, WonKey Lee, and Kyungja Ahn. The present situation analysis of teaching of the speaking in primary elt and some strategies for improvement. *Primary English Education*, 2015.

[6] Kyuchul Yoon. Using the prosody cloning technique in teaching english prosody. *Studies in English Language & Literature*, 37(1):245–271, 2011.

[7] Kim Silverman, Mary Beckman, John Pitrelli, Mari Ostendorf, Colin Wightman, Patti Price, Janet Pierrehumbert, and Julia Hirschberg. Tobi: A standard for labeling english prosody. 01 1992.

[8] Suhyeon Lee. *A Study on Awareness of ' Teaching Difficulties in English Class ' and Ways of Dealing with Them by Elementary School Teachers.* PhD thesis, Ewha Womans University, year = 2018, address = Seoul, South Korea,.

[9] Jaechun Park. English teaching roles are reverting back to korean teachers...numbers of foreign teachers diminishing. *Yonhap News.* Korean News Article. URL: `http://www.yna.co.kr/view/AKR20170227079200064`.

[10] Iljee Lee. The study of perception of teachers and students about english speaking education and evaluation. *The Journal of Humanities and Social Sciences 21*, 11(2):373–387, 2020.

[11] Ken Beatty. *Teaching and Researching: Computer-Assisted Language Learning.* Routledge, 2nd edition, June 2010.

[12] Ao Abusa'aleek. Computer assisted language learning: Merits and demerits. *Language in India : Strength for Today and Bright Hope for Tomorrow*, 12, April 2012.

[13] Hirokazu Yokokawa, Kayoko Fukuchi, Yuko Ikuma, and Katsuhiko Masaki. Significance of ict materials for foreign language activities in elementary schools and its effective use. *Computer & Education*, 29:36–41, 2010. `doi:10.14949/konpyutariyoukyouiku.29.36`.

[14] Daniel Felps, Heather Bortfeld, and Ricardo Gutierrez-Osuna. Foreign accent conversion in computer assisted pronunciation training. *Speech Communication*, 51(10):920–932, October 2009. URL: `http://doi.org/10.1016/j.specom.2008.11.004`.

[15] Philippe Martin. Winpitch ltl ii, a multimodal pronunciation software. May 2020.

[16] SpeedLingua-SA. Speedlingua. `http://home.speedlingua.com/en/`, 2018.

[17] Paul Boersma and David Weenink. Praat: doing phonetics by computer (version 6.1.16)[computer program]. retrieved june 28, 2020, 2020.

[18] Will Styler. Using praat for linguistic research. http://wstyler.ucsd.edu/praat/UsingPraatforLinguisticResearchLatest.pdf, 2017.

[19] Yannick Jadoul, Bill Thompson, and Bart de Boer. Introducing parselmouth: A python interface to praat. *Journal of Phonetics*, 91:1–15, November 2018. doi:10.1016/j.wocn.2018.07.001.

[20] Eric Moulines and Francis Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Commun.*, 9(5–6):453–467, December 1990. URL: https://doi.org/10.1016/0167-6393(90)90021-Z, doi:10.1016/0167-6393(90)90021-Z.

[21] Anna De Meo, Marilisa Vitale, Massimo Pettorino, Antonio Origlio, and Francesco Cutugno. Imitation/self-imitation in computer-assisted prosody training for chinese learners of l2 italian. January 2013.

[22] Elisa Pellegrino and Debora Vigliano. Self imitation in prosody training:a study on japanese learners of italian. September 2015.

[23] Keiko Nagano and Kazunori Ozawa. English speech training using voice conversion. In *ICSLP*, 1990.

[24] Maria Paola Bissiri, Hartmut Ptzinger, and Hans Tillmann. Lexical stress training of german compounds for italian speakers by means of resynthesis and emphasis. January 2006.

[25] B. Juang and Lawrence Rabiner. Automatic speech recognition - a brief history of the technology development. 01 2005.

[26] Google. Getting words timestamps — cloud speech-to-text documentation. https://cloud.google.com/speech-to-text/docs/async-time-offsets.

[27] Aaron van den Oord, Sander Dieleman, Karen Zen, Heiga abd Simonyan, Karen Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, September 2016.

[28] Zoe Handley. Towards establishing a methodology for benchmarking speech synthesis for computer-assisted language learning (call). October 2011.

[29] Ambra Neri, Catia Cucchiarini, and Helmer Strik. Automatic speech recognition for second language learning: How and why it actually works. *Speech Communication*, January 2003.

[30] KyungSuk Hong. *a Study of the role of songs in teaching English prosody in elementary school*. PhD thesis, Korea National University of Education, North Chungcheong, South Korea, February 2004. Korean.

[31] Kiyoshi Matsuzawa. *Eigo Mimi Drill*. ASCII Media Works, Tokyo, Japan, 2009.

[32] Yuta Goto. Prosody master club : Expansion prosody learning environments to other situations, 2018. Japanese.

[33] Mietta Lennes. SpeCT - Speech Corpus Toolkit for Praat (v1.0.0). First release on GitHub, March 2017. URL: `https://doi.org/10.5281/zenodo.375923`, `doi:10.5281/zenodo.375923`.

[34] Sunae Ham, Yangsoon Lee, Hyuna Kim, Sukyung Park, Jangwoong Park, Eun Junghua Ahn, Soyeon, Jungmin Lee, Namhee Lim, Shinwoo Jung, and Heather L. Reichmuth. *Elementary School English 5*. Chunjae Education, Korea, 2nd edition, March 2020.

Appendices

A. User Test Materials

Dialogue 1:

Q: Where are you from?
A: I'm from _____
(예시 : Korea, the U.S., Mexico)

Dialogue 2:

Q: What do you do on weekends?
A: I ○○○○ on weekends.
(예시 : play soccer, watch a movie, go to the cooking club. ⋯)

Dialogue 3:

Q: May I take a picture?
A: Yes you may take a picture / No, you many not take a picture.

Dialogue 4:

Q: Whose book is this?
A: it's ○○'s. / It's mine.

Figure A.1 The Dialogue Sheet During The User Test

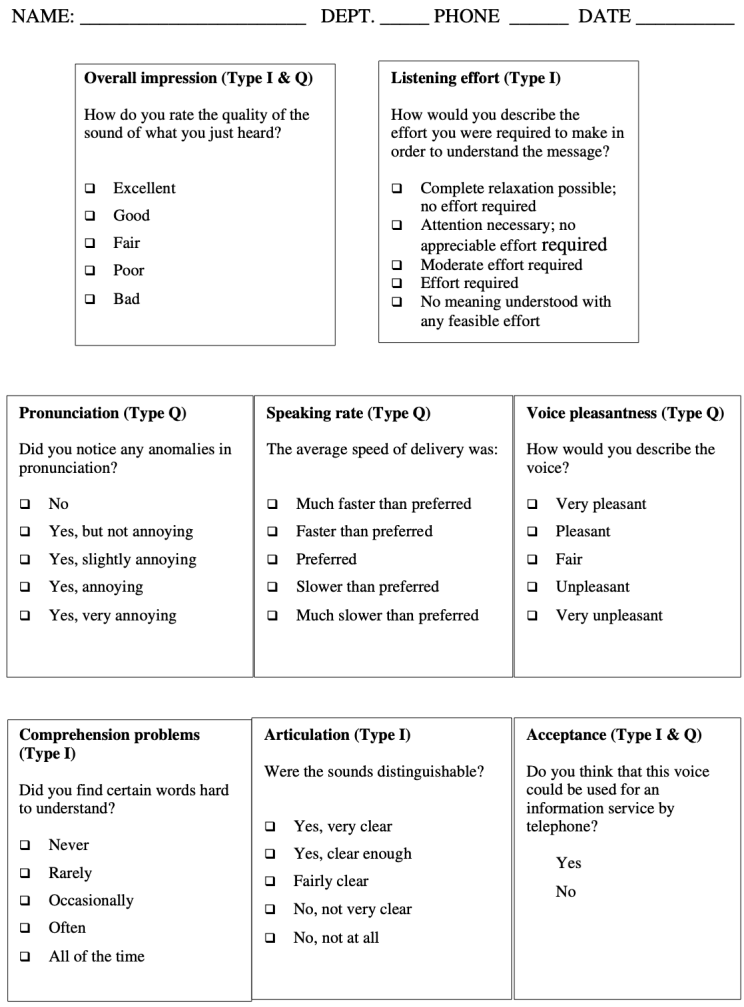


Figure A.2 The Original MOS Survey by Viswanathan et al. (2005) [3]

설문지

	질문	전혀 그렇지 않다	그렇지 않다	보통	그렇다	매우 그렇다
Q1	나는 영어 공부가 재미있다					
Q2	나는 영어 공부가 중요하다고 생각한다					
Q3	나는 영어 말하기가 어렵다고 생각한다					
Q4	영어와 한국어는 말하는 리듬과 높낮이가 다르다고 생각한다.					
Q5	자신의 목소리가 원어민 억양처럼 들려서 재미있었다.					
Q6	오늘 한 것과 같은 방법을 영어 학습에 적용해 보고 싶다.					

	Question	Not at all	Not so much	So-so	Pretty Much	Very Much
Q1	I enjoy studying English					
Q2	I think studying English is important					
Q3	It's pretty hard to be good at speaking English					
Q4	I think intonation and rhythm are very different between English and Korean					
Q5	I practice speaking English outside school					
Q6	I'd like to try similar methods on actual English studies.					

Figure A.3 The Rating Sheet (The Original & Translated)