Voice Recognition for English Pronunciation Practice: A Case Study of Siri in iPad

Saranya Pathanasin [1]
Maurice Blackford [2]

Abstract

Voice recognition has proven potential in the teaching of English pronunciation. Many tools have been developed specifically for this purpose; however, such tools are neither available in normal language classrooms nor students' self-practice. The present study tested the possibility of using the Siri voice recognition in the iPad as a tool for pronunciation practice for the reason that Apple products are owned by many Thai University students, who are our target group. We compared the accuracy of voice recognition when used by a native and a non-native speakers of English. The corpus data was on three levels: word, phrase, and sentence. The Chi-square test was used to test differences of accuracy between the native and the non-native speakers in the corpus. The differences between three assessments of voice recognition were measured between the two speakers. Results showed that Siri could recognize utterances spoken by the native speaker significantly differently to those spoken by non-native speaker for word ($p<0.05$), phrase ($p<0.05$), and sentence ($p<0.05$). No significant difference between the three assessments was found within both speakers. Among all corpus tested, the least correctness was detected at sentence level, which differed significantly to word and phrase levels (49.7% of native speaker ($p<0.001$) and 16% of non-native speaker ($p<0.001$)). This

[1] Lecturer,Dr., International Language Center, Faculty of International Studies, Prince of Songkla University, saranya.p@phuket.psu.ac.th

[2] Lecturer, International Language Center, Faculty of International Studies, Prince of Songkla University

proved that correct pronunciation was essential to be recognized by the application. We conclude that the voice recognition application in Apple products has the potential to be used for pronunciation practice in students' self-study scheme.

Keywords: Voice Recognition in EFL, Pronunciation Practice, Siri

การใช้โปรแกรมรู้จำเสียงพูดในการฝึกออกเสียงภาษาอังกฤษ: กรณีศึกษา Siri ใน iPad

สรัญญา พัฒนศิลป์ [1]
Maurice Blackford [2]

บทคัดย่อ

โปรแกรมรู้จำเสียงพูดได้ถูกนำมาใช้เป็นประโยชน์ในการสอนการออกเสียงภาษาอังกฤษได้ จึงได้มีผู้คิดพัฒนาโปรแกรมรู้จำเสียงพูดเพื่อวัตถุประสงค์นี้โดยเฉพาะ แต่เนื่องจากโปรแกรมรู้จำเสียงพูดเพื่อการเรียนการสอนภาษาอังกฤษยังไม่แพร่หลาย ทั้งในห้องเรียนและการศึกษาด้วยตนเอง การศึกษานี้ทำการทดสอบความเป็นไปได้ที่จะใช้โปรแกรมรู้จำเสียงพูด Siri ใน iPad เป็นเครื่องมือในการฝึกออกเสียงด้วยเหตุผลว่า นักศึกษามหาวิทยาลัยซึ่งเป็นกลุ่มเป้าหมายของการศึกษาครั้งนี้เป็นเจ้าของผลิตภัณฑ์ APPLE เป็นจำนวนมาก การศึกษานี้ทำการเปรียบเทียบความแม่นยำในการรู้จำเสียงพูดของโปรแกรมเมื่อทดสอบโดยเจ้าของภาษาและผู้ที่ไม่ใช่เจ้าของภาษา คลังข้อมูลที่ใช้แบ่งเป็น 3 ระดับคือ ระดับคำ ระดับวลี และระดับประโยค ใช้วิธี Chi-square วัดความแตกต่างระหว่างความถูกต้องของการรู้จำเสียงพูดของเจ้าของภาษาและผู้ที่ไม่ใช่เจ้าของภาษา ผลการศึกษาพบว่า SiRi สามารถรู้จำเสียงพูดของเจ้าของภาษาได้ดีกว่าผู้ที่ไม่ใช่เจ้าของภาษาอย่างมีนัยสำคัญทางสถิติ ในระดับคำ ($P<0.05$) ระดับวลี ($P<0.05$) และระดับประโยค ผลการวิเคราะห์ชี้ว่าไม่มีความแตกต่างอย่างมีนัยสำคัญในการทดสอบทั้งสามครั้งด้วยผู้พูดคนเดียวกัน นอกจากนี้ยังพบว่าโปรแกรมรู้จำเสียงพูดมีประสิทธิภาพด้อยที่สุดในการรู้จำเสียงพูดระดับประโยค โดยมีความแตกต่างอย่างมีนัยสำคัญเมื่อเทียบกับระดับคำและวลี 49.7% ในเจ้าของภาษา ($p<0.001$) และ 16% ในผู้พูดที่ไม่ใช่เจ้าของภาษา ($p<0.001$) ผลการศึกษาชี้ว่าการออกเสียงที่ถูกต้องมีความจำเป็นในการทำงานของโปรแกรม ข้อสรุปของการศึกษาคือโปรแกรมรู้จำเสียงพูดของผลิตภัณฑ์ APPLE มีประสิทธิภาพที่นำมาใช้ประโยชน์ในการฝึกการออกเสียงด้วยตนเองของผู้เรียนภาษาได้

คำสำคัญ: โปรแกรมรู้จำเสียงพูดในการเรียนการสอนภาษาอังกฤษ การฝึกการออกเสียง Siri

[1] อาจารย์ ดร. ศูนย์ภาษานานาชาติ คณะวิเทศศึกษา มหาวิทยาลัยสงขลานครินทร์ saranya.p@phuket.psu.ac.th

[2] อาจารย์ ศูนย์ภาษานานาชาติ คณะวิเทศศึกษา มหาวิทยาลัยสงขลานครินทร์

Introduction

Technology, which plays a critical role in our modern lives, is becoming fruitful in the field of English as a Foreign Language (EFL) in both teaching and learning. Many efforts have been made in trying to exploit technology both inside and outside classrooms, from audio to visual. Some are successful while some need further development. One area in which technology has been proven viable in EFL is pronunciation teaching, in which voice recognition must surely be one of the technological tools with the highest potential within the scheme. According to Chandra et al. (2011), speech recognition or voice recognition is the process of converting acoustic signals, captured by microphone or telephone, to a set of words. Previous studies reported the attempts at using the tool to enhance pronunciation in second language learners. For example, Junqua (1999) invented a speech recognition tool by using an adaptation system. Based on the principle that other tools were trained to recognize corpus produced by native speakers, the adapted system tool was designed specifically to recognize the speech of second language learners. The inventor claimed that the adapted system was useful in pronunciation practice. In addition, Witt and Young (1998) employed a computer-assisted tool in pronunciation teaching. This system worked by scoring the 'goodness of pronunciation'. With this method, the machine was able to detach mispronunciation and identify the sounds that tended to be problems in students.

Voice recognition has been exploited not only in EFL but also in other languages as well. In 2000, Kawai and Hirose employed Double-Mora Phonemes, a computer-aided language learning system (CALL), in teaching pronunciation to second language learners of Japanese. The purpose of the 'Double-Mora Phonemes' was to solve learners' problems in pronouncing vowels which sound very

similar. Students were assigned to read minimal pairs and the machine would evaluate how likely it was that Japanese native speakers would be able to understand the differences in the pairs. Previous studies which focused on using voice recognition technology in language teaching and learning, posed questions on how this technology could be really viable in normal classrooms, since many tools which had been created are not available in normal language classrooms, let alone with students' self-study outside their classes. Although we realize that voice recognition systems have not been fully developed at the present time and there are several limitations, we seek possibilities to make use of this technology in pronunciation practice. Our intention is to apply voice recognition in smart phones to enhance Thai students' English pronunciation. For this intention, Siri, which is a voice recognition application in iPad produced by the Apple Company, has been selected. Although Siri was not designed specifically for language teaching and learning purposes, the iPad is owned by a large number of university students in Thailand, who are our main target.

It should be noted here that the present study is a part of an umbrella project consisting of voice recognition testing in six languages, both Asian and European. In this article, we focus only on English language testing. Our main aim was to test the possibility of using Siri applications in English pronunciation practice and to compare the pronunciation percentages by using a statistical method. We hypothesized that Siri would be able to recognize utterances if the speakers pronounced them clearly and correctly. It is hoped that results of this study could be extended to further development in using voice recognition for students' self-study as well as being a guideline for adopting voice recognition applications in other EFL aspects. In the sections that follow, overviews of English pronunciation problems in Thai EFL learners are reviewed

and brief information on Siri application system is presented in the second section. In section III, methodology of the study is described in steps. Then, results and discussions are in section IV. The final section presents a conclusion and the limitations of the study.

Literature Review

This section has two sections. Firstly, phonological differences between English and Thai are described, accompanied with reviews of previous studies concerning pronunciation problems in Thai learners. Secondly, a brief introduction of the Siri application is presented in order to provide basic information on its purpose and working process.

1. Phonological differences between English and Thai

English pronunciation has proven challenging for Thai learners of all levels. Difficulties in English pronunciation derive from phonological differences between the two languages. In this paper, we summarized three main phonological differences between English and Thai, which are: speech sounds, intonation, and prosodic features. Significantly, they are problematic to Thai students.

1.1. Speech sounds

English and Thai have different sound systems and phonological rules. Table 1 and Table 2 demonstrate the distinctive consonant sound systems between the two languages.

Table 1: English Consonant Sounds

	Bilabial		Labio-dental		Dental		Alveolar		Post-alveolar		Palatal	Velar		Glottal
Plosive	p	b					t	d				k	g	
Nasal		m						n					ŋ	
Fricative			f	v	θ	ð	s	z	ʃ	ʒ				h
Affricate									tʃ	dʒ				
Lateral								l						
Approximant		(w)						ɹ			j		w	

Adapted from Kanokpermpoon (2007)

Table 2: Thai Consonant Sounds

	Bilabial	Labio-dental	Alveolar	Lamio-prepalatal	Palatal	Velar	Glottal
Plosive	p b p^h		t d t^h			k k^h	ʔ
Nasal	m		n			ŋ	
Fricative		f	s				h
Affricate				tɕ $tɕ^h$			
Tap			ɾ				
Lateral			l				
Semivowel	(w)				j	w	

Adapted from Kanokpermpoon (2007)

As can be seen from the above tables, there are some English consonant sounds which do not exist for Thai such as /z/, and /v/. According to Kanokpermpoon (2007), these sounds posed problems for Thai students. Furthermore, phonology rules which are different in the two languages were another obstacle to Thai students when pronouncing English words. For example, English phonology allows some consonant sounds i.e. /f/ /v/ to occur at the word-final position. This cannot be found in the same position in Thai. Kanokpermpoon (2007) pointed out that when facing such difficulties, Thai students tended to substitute Thai sounds for English sounds which do not exist in Thai, for example, the word 'leave' /li:v/ is often pronounced /li:b/ by Thai students because the voiced labiodental fricatives /v/ does not exist in the word-final position in Thai. Thai students therefore substitute the voiced bilabial stop /b/ to /v/, and it resulted in incorrect pronunciation.

1.2 Intonation

English is not a tonal language, but Thai is. The high and low tones in English do not change word meanings, for example, if a speaker says 'car' /ka:r/ either in low or high tone, the word 'car' still means a vehicle. In contrast, Thai words have specific tones. For example, the word คา /ka:0/ with low tone means 'hang on', but ค่า /ka:3/ with high tone means 'price'. Intonation was found to be a

problem in Thai learners' pronunciation as reported in a study by Pongprairat (2011) indicated that Thai university students with low English proficiency were scored by native speakers of English lower than those with high proficiency both in intelligibility and comprehensibility dimensions with regard to intonation.

1.3 Prosodic Features

Prosodic features include length, pitch, and stress. Pitch and stress were found to be critical problems in the English pronunciation of Thai learners. In English words, one syllable receives prominent stress whereas others are lesser or unstressed syllables. The schwa /ə/ represents the unstressed vowel. For example, the word 'about' is pronounced /əba´wt/. Different stress can change word meanings. The word 'digest' /´dɪʤest/ as a noun means summation of articles, but as a verb 'digest' /dɪ´ʤest/ means to absorb food, for example.

Native speakers of English have knowledge of English prosodic features, but such features surely challenge L2 learners especially when their mother tongue languages are not stressed languages such as Thai. A study by Kamkhien (2010); which focused on word stress problem in Thai students, pointed out that students' performances on word stress assignments were unsatisfying. Results of the study also showed that students scored lowest in pronouncing five-syllable words, and had least difficulty with two-syllable words. In addition, female students could perform better than male students, while disparities in students' faculties and years of studying English made no significant difference to levels of success carrying out the same tasks.

As presented above, the three different features between English and Thai (sound systems, intonation, and stress) cause difficulties in the English pronunciation of Thai learners as proven in several studies. To help Thai learners overcome the problem, researchers in the field offered pedagogical suggestions in order to

improve students' pronunciation skill. More importance should be given to pronunciation practice in classrooms, and self-study practice outside classrooms is beneficial.

Concerning this matter, the present study attempts to seek possible self-study tools for students. Voice recognition is seen as a viable technology. Among several voice recognition applications, Siri in iPad was chosen in the present study. Therefore, basic information of the application is presented in the next part as background knowledge of the study.

2. SIRI: Voice Recognition

According to Geller (2012), Siri is an artificial intelligence application which works on iOS operating system, which is integrated in products of the Apple Company. Siri derives from a combination of several technologies: voice recognition, information management, artificial intelligence, task fulfillment, and user interface. It works by two main applications, namely: grammar-based command recognition and language-based dictation. The former is a conversational system which principally does not need to understand large amounts of vocabulary. The latter system recognizes and transcribes every word it hears by analyzing an amount of vocabulary which is too large to be stored in a portable phone. The collection of data is in Cloud storage and can be retrieved by a high-speed bandwidth. This enables Siri to interact with users.

The current purpose is to receive spoken pieces as commands and then respond to the commands. For example, if the user says 'Lunch with Miss Melanee this Sunday', the command will be added in calendar appointment. However, Siri is still ongoing development. The researchers' goal is to develop Siri to be able to interact with the users in higher capacity. For example, it can call the user by name, or if the user says 'I'm drunk', it can offer to call a taxi cab

for her.

Undoubtedly, Siri is a promising technology which can be useful in many different ways. However, we realize that Siri is not invented specifically for teaching and learning purposes. The success of the Apple products means they have potential for university students' self-study many of whom own personal iPads.

Methodology

This is a quantitative sample study. Our methodology consisted of four steps, and this section gives a brief details of each step.

The first step was compiling a corpus which was the data of the present study. As mentioned earlier, this sample study was a part of a larger project consisting of six pieces of language data. Our corpus therefore was a collection of utterances from six languages. By following the same criteria, the research team provided data in three sub categories: 20 pieces at the word level, 20 pieces at the phrase level, and 20 pieces at the sentence level. In total, our corpus consisted of 120 words, 120 phrases, and 120 sentences. They were all in English. All utterances were inputted in an excel file by separating one category for each excel sheet. At this stage, we had three excel sheets containing 120 words, 120 phrases, and 120 sentences, respectively.

The second step was the voice recognition testing by a native speaker of English. A British subject was assigned to read all utterances in the data. Each utterance was read three times to the Siri application in the iPad. If Siri could recognize the utterance correctly, the subject would mark ✓ and ✘, for the incorrect ones. The marks were recorded in separated columns next to the utterances in the same excel sheet.

Basically, the third step of analysis repeated the previous step all over again, but the voice recognition testing at this stage was

done by a non-native speaker of English. The subject was a Thai female who uses English as her second language.

The last step was statistical analysis. A statistical test called 'Chi-square' was used to test differences between the accuracy of the two speakers. The results of voice recognition testing by the two subjects were analyzed using Chi-square test in three different angles. Firstly, the correctness of voice recognition between native and non-native speakers was compared on word, phrase, and sentence levels. This analysis was to find out how different percentages of correctness were between native and non-native speakers. Secondly, a comparison was done between the three trials of each category and of each speaker. This was to investigate whether Siri recognized the same utterance differently when repeated three times by the same speaker. Lastly, overall results of each data category were compared separately with respect to the speakers. This analysis shows the difference in percentages of correctness between categories tested by the same speaker.

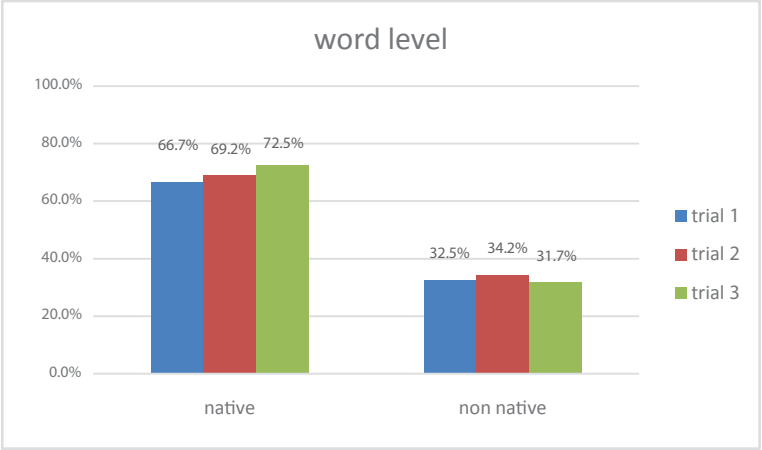
Results and Discussion

This section presents results of the correctness of voice recognition and comparisons as described above in the methodology section. Our data was in three categories: words, phrases, and sentences, so the results are presented in three sub-sections, along with the statistical comparison between the two subjects.

1. Word Level

The data in our first category was on word level. There were 120 items. These words were randomly selected from various fields and were in different parts of speech, both nouns and verbs. Figure 1 presents the result of the comparison between the two subjects.

Figure 1: Word level



It can be seen from Figure 1 that Siri application in iPad machine could recognize the voice of the native speaker approximately twice as accurately as that of the non-native speaker. The results showed that Siri could recognize utterances spoken by a native speaker significantly differently from those spoken by the non-native speaker at the word level ($p<0.05$).

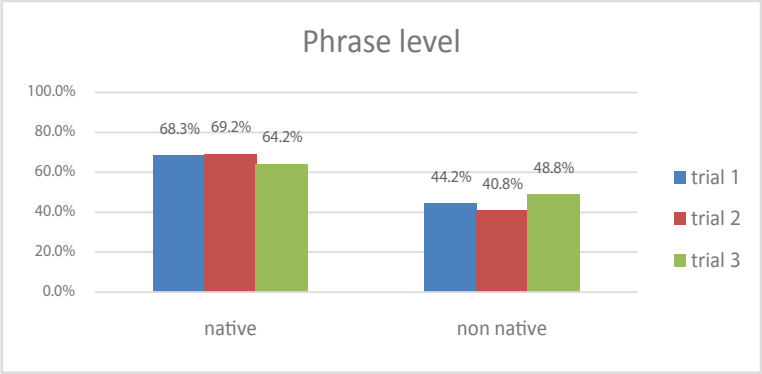
In addition, the result of the Chi-square test revealed that with native voice recognition, there was no significant difference of correctness between three trials. This finding was similar to that of the non-native speaker.

A deeper analysis on length of words revealed that the voice recognition application tended to have problems recognizing monosyllabic words such as sand, high, breeze, serve, etc. spoken by both subjects. It tended to perform better in multisyllabic words such as website, drama, environment, homework, computer, etc., even if the speakers stressed the wrong syllable. This finding was not different between the two subjects.

2. Phrase Level

The data on phrase level was analyzed similarly to the data in the first group. The overall result is shown in Figure 2.

Figure 2: Phrase level



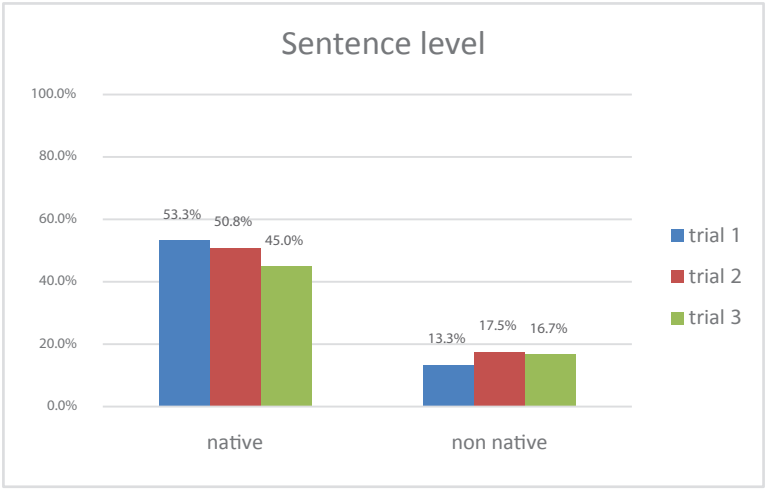
Overall, it was found that the percentage of correctness in the native speaker was relatively similar to the percentages at the word level, but percentages of correctness in the non-native speaker were higher than at word level. The results showed the same trend as the results of the first category. That is to say, the voice recognition could recognize the voice of a native speaker differently from a non-native speaker (phrase; $p<0.05$).

As we consider its performance by trials, statistical results showed that the application could perform relatively similarly in all three trials, and this trend was similar in both subjects.

3. Sentence Level

The last category of data was on sentence level. They were simple sentences, compound sentences, and complex sentences. The analysis method was similar to the other two sets of data above. The results of the analysis are as follows:

Figure 3: Sentence Level



Similar to the other two subsets presented earlier, it was found that Siri application could recognize the voice of a native speaker better than that of a non-native speaker. As indicated in Figure 3 indicated, the highest percentage of correctness in native speaker data was 53.3%, whereas the highest correctness in non-native speaker data was only 17.5%. The analysis also revealed that Siri could recognize short sentences such as 'You are right', 'I like you', 'What is the passport number?' more effectively than long and complex sentences such as 'It can be seen from the table that the number of population increases rapidly', 'The reason I don't have a credit card is that my salary is low', etc. Statistical analysis revealed that the percentages of accurate dictations by the Siri application between native and non-native speakers were significantly different in the sentences measured ($p<0.05$). In addition, the three trials were compared and analyzed statistically. It was found that there was not significant difference between trials for both native and non-native speakers.

In the last step of the analysis we compared percentages of accuracy between the three categories of each speaker. It was found that the lowest percentages of correctness fell into the sentence category in both speakers and was significantly different to word and phrase correctness percentages, i.e. 49.7% for native ($p<0.001$) and 16% for non-native ($p<0.001$).

Conclusion

In conclusion, the results of the analysis clearly prove that Siri application could recognize utterances spoken by a native speaker of English better than those spoken by a non-native speaker in all categories. It can be claimed that correct pronunciation is essential for the machine to recognize an utterance. Interestingly, the results proved that the application could recognize words with more than one syllable and longer pieces of utterances (multisyllabic words and phrases) obviously better than monosyllabic words. The performances of the two subjects, although significantly different, were in the same trend. That is to say, the Siri application could recognize utterances on word and phrase levels better than utterances on sentence level.

We can thus come to the suggestion that Siri can be used in EFL teaching. We can apply this principle in English pronunciation teaching, for example, students may be assigned to practice pronunciation vocabulary using voice recognition. However, we should be aware of its limitations in recognizing monosyllabic words as well as long utterances such as compound and complex sentences. We also recommend that EFL teachers design pronunciation tasks to be suitable for the voice recognition's capacity.

Lastly, it should be noted here that this study has at least two limitations. Firstly, the present study was conducted using only the Siri application on an iPad only. Therefore, the results might not be fully applicable to other voice recognition applications which are

available on the market. Teachers of EFL who would like to apply the results of our study in their teaching might use our analysis method in order to understand the particular limitations of their desired machines. In addition, the number of subjects was only two and the size of corpus was not large. Therefore, it is recommended that a larger sample sizes should be tested by more subjects to confirm the results of the present study.

Acknowledgement

We would like to express our sincere thanks to the Faculty of International Studies, Prince of Songkla University, for supporting this project. Thanks are also extended to Dr.Sombat Khruathong for including us in the project. Special thank is also offered to Dr. Sarika Pattanasin for assisting in statistical analysis.

References

Chandra et al. (2011). Automatic Speech Recognition: Architecture, Methodologies and Challenges-A Review. *International Journal of Advance Research in Computer Science*. 2(6), 326-331.

Fromkin et al. (2007). *Introduction to Language*. Thomson Higher Education. Boston.

Geller, T. (2012). Talking to Machines. *Communications of the ACM*. 55(4), 14-16.

Junqua, J. (1999). Speech recognition and teaching apparatus able to rapidly adapt to difficult speech of children and foreign speakers. *U.S. Patent* No. 6, 253,181. 26 Jun. 2001.

Kamkhien, A. (2010). Thai Learners' English Pronunciation Competence: Lesson Learned from Word Stress Assignment. *Journal of Language Teaching and Research*. 1(6), 757-764.

Kanokpermpoon, M. (2007). Thai and English Consonantal Sounds: A Problem or a Potential for EFL Learning?. *ABAC Journal.* 27(1), 57-66.

Kawai,G. and Hirose, K. (2000). Teaching the pronunciation of Japanese double-mora phonemes using speech recognition technology. *Speech Communication*. pp.131-143.

Pongprairat, R. (2011). *A Study of Interlanguage English Intonation in Thai Learners, and the Degree of Intelligibility and Comprehen sibility in Native Speakers' Judgments*. (Doctoral dissertation, English as an International Program). Chulalongkorn University. Bangkok.

Witt, S. and Young, S. (1998). Performance Measures for Phone-Level Pronunciation Teaching in CALL. *Proc. of the Workshop on Speech Technology in Language Learning,* pp. 99-102, Marholmen, Sweden.