문학석사 학위논문

A Study of the Relationships Among ASR Accuracy, Human Transcription, and Comprehensibility Judgment

영어 비원어민 발화에 대한 음성 인식기의 전사 정확도와 인간 청자의 전사 정확도 및 이해가능도 평가 간의 연관성 연구

2021년 8월

서울대학교 대학원

영어영문학과 어학 전공

강 지 민

A Study of the Relationships Among ASR Accuracy, Human Transcription, and Comprehensibility Judgment

지도 교수 이 용 원
공동 지도 교수 정 민 화

이 논문을 문학석사 학위논문으로 제출함

2021년 8월

서울대학교 대학원
영어영문학과 어학 전공
강 지 민

강지민의 문학석사 학위논문을 인준함
2021년 8월

위 원 장 _____ 김 선 희 _____

부위원장 _____ 이 용 원 _____

위 원 _____ 정 민 화 _____

Abstract

A Study of the Relationships Among ASR Accuracy, Human Transcription, and Comprehensibility Judgment

Kang, Jeemin
Department of English Language and Literature
The Graduate School
Seoul National University

This paper investigates the relationships among measures of ASR accuracy, human transcription accuracy, and human comprehensibility judgment of non-native speech, which can potentially be utilized for computer-assisted pronunciation training (CAPT). Native and non-native human listeners were asked to transcribe 1,505 short fragments of non-native read speech and subsequently rate the comprehensibility of each of the fragments on a 5-point scale. The recognition accuracy rates of two different ASR systems (Google, ETRI) were compared, one of which was for general use and the other was optimized for recognizing non-native speech. These two ASR systems' accuracy rates were compared with the transcription accuracy of human transcribers, and the correlations between ASR

accuracy and the two kinds of human measures (i.e., the intelligibility (or transcription accuracy) score and comprehensibility rating) were obtained and closely examined. Both ASR systems showed a significantly lower accuracy rate compared to human listeners in transcribing the non-native speech, but the ASR system whose recognition model was built based on non-native speech data showed a significant enhancement in recognizing non-native speech, almost approaching the accuracy rate of human listeners. In terms of correlations, a moderate positive correlation was obtained between ASR accuracy and human recognition accuracy and comprehensibility scores. Of the two ASR systems used in this study, it was found that the ASR that modeled non-native speech showed higher correlation with human intelligibility. These results suggest the potential of using ASR systems optimized for non-native speech in providing pronunciation feedback to L2 learners.

Keywords: computer-assisted pronunciation training, automatic speech recognition, speech transcription, L2 speech, comprehensibility, intelligibility

Student Number: 2017-26486

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

1.1. Background and Motivation

When it comes to pronunciation acquisition, L2 learners are spread out on a wide spectrum of learning needs and L1 backgrounds, and thus each learner requires different types of training and feedback. However, in typical English as a Second or Foreign Language (ESL/EFL) classrooms, teachers find it difficult to integrate pronunciation training into the curriculum not only due to the lack of time and confidence in teaching pronunciation (Baker, 2011) but also the inherent difficulties in providing individualized feedback. As a promising way to solve this problem, some scholars take note of computer-assisted pronunciation training (CAPT) systems for their capabilities to provide language learners with round-the-clock access, an anxiety-free practice environment, a sense of learner autonomy, and personalized feedback (Guskaroska, 2019).

Although there are many available CAPT tools that provide learners with a score or feedback on their speech production, such as Rosetta Stone and Duolingo, they have limited flexibility in that learners should follow the prescribed plans built into the CAPT program and can only practice preprogrammed utterances presented by the system. Automatic speech recognition (ASR) systems, on the other hand, provide more flexibility in choosing the topics or language items to practice for language learners (McCrocklin et al., 2019). They are simple but can be used as effective tools for providing pronunciation feedback, and, for these reasons, their use for language learning has significantly increased in recent years

(Ahn & Lee, 2016; Wang & Young, 2015; Witt & Young, 2000). Automatic speech recognition (ASR), as shown in Figure 1.1, is the process of converting speech signal input into a string of words using a number of information sources (Cucchiarini & Strik, 2018). These information sources are obtained by training the speech recognizer with a large amount of audio data and their corresponding text transcriptions. Through this process, acoustic models and language models are derived. The acoustic model is used to identify the individual speech sounds, while the language model contains information on which words are likely to follow each other. With these two models and a lexicon (dictionary), which contains all the words that can be recognized along with their pronunciations, the decoder (search algorithm) searches for the most probable string of words that corresponds to the input speech. Non-native speech is challenging to recognize because L2 learners deviate from native speakers in terms of pronunciation, vocabulary, grammar, and also fluency that are associated with all three information sources mentioned above.

While enhancing the ASR's recognition accuracy of non-native speech is also an area of interest to many researchers, it is also interesting how some researchers are focusing on the ASR's recognition errors to identify which parts of the non-native speaker's utterance is difficult to understand for a potential listener. If a learner speaks in English, the ASR converts the speech into text, which can in turn give feedback to the learner on which parts of their utterance were less clear or needs improvement. Although speaking to a human listener could have many advantages, it is not always possible for language learners to find a listener who can help them patiently and consistently with their pronunciation practice. ASR

systems, on the other hand, are tireless and consistent in assisting learners, and with their recognition accuracy approaching that of human listeners in recent years, there is a hope that ASR systems could act as human-like listeners in L2 pronunciation training contexts.

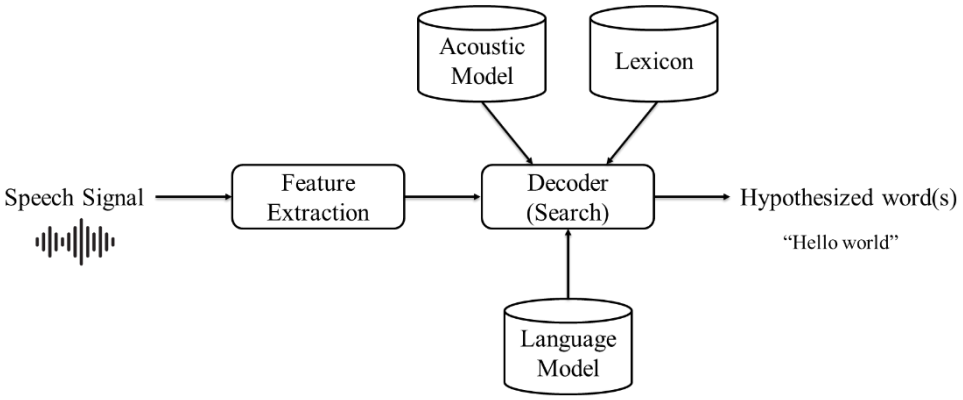

Figure 1.1 Automatic Speech Recognition

In the field of L2 pronunciation teaching and assessment, comprehensibility and intelligibility of non-native speakers have become a very important concept (Chalhoub-Deville, 1995; Derwing & Munro, 1997; Isaacs & Trofimovich, 2012; Munro & Derwing, 1995a, 1995b). For L2 learners, comprehensible and intelligible pronunciation, as opposed to native-like pronunciation, has been advocated as a more realistic goal of learning in recent years. Although native-like pronunciation could be a desirable goal to pursue for some learners, it can be very challenging or pedagogically unnecessary to make all L2 learners pronounce words exactly as native-speakers do in many real L2 learning contexts. One can argue that a certain range of deviations from the native

speaker norm should be tolerated as long as the L2 learner's pronunciation does not severely interfere with the listeners' comprehension of the overall message.

Then, one intriguing question is whether we can incorporate this key concept of comprehensibility and intelligibility into ASR-based pronunciation training. In this sense, it is important to first understand the current state of ASR accuracy of non-native speech by comparing it with human transcription accuracy done by human listeners. By investigating the correlation between ASR and human transcription accuracy rates, we can gain insight into how reliable ASR systems are in modelling human intelligibility, and how teachers and students can utilize these ASR systems for pronunciation instruction or practice purposes. Another essential topic of investigation in this study is the relationship between human comprehensibility ratings and ASR transcription accuracy of non-native speech. In summary, we want to investigate to what extent state-of-the-art ASR systems model human listeners' intelligibility and comprehensibility in processing utterances from non-native speakers of English.

Results of such a line of studies can encourage L2 teachers and students to utilize pre-existing ASR systems for pronunciation practice purposes and contribute to the development of ASR-based L2 speech instruction, assessment, and feedback systems. ASR to be utilized for pronunciation instruction should be built and trained to recognize the speech similarly to how human listeners would recognize it. ASR errors that occur where human listeners also experience difficulty in recognizing or understanding could provide useful feedback to learners

by pinpointing the parts that caused communication difficulties (or breakdowns). On the other hand, if the ASR errors occur too frequently in parts where human listeners had no difficulty understanding, then the errors can cause confusion and frustration to the L2 learners and exert a harmful effect on their self-motivation and language development in general. In this sense, an ASR system that closely models human listeners' transcription behavior and reflects human comprehensibility judgments can effectively assist learners with assessing their weaknesses as well as strengths in L2 pronunciation.

Then, how can we evaluate the existing ASR systems in terms of their accuracy and usefulness in L2 instruction and assessment? In this regard, Derwing et al. (2000) proposed two important criteria for evaluating the usefulness of ASR in providing corrective feedback to EFL speakers. First, the ASR must recognize the L2 speech at an acceptable level. Second, the ASR's identification of L2 speech must resemble that of native listeners in order to enhance its capability to identify areas where L2 learners have production difficulties and to provide feedback on how to improve their pronunciation. With these as a background, this study aims to examine the relationship among human comprehensibility judgment, human transcription accuracy, and the accuracy of two kinds of ASR systems generally in accordance with the two criteria proposed in Derwing et al. (2000). Acknowledging the fact that the accuracy of ASR systems continues to improve, the two selected ASR systems are evaluated in this study by capturing a snapshot of their performance on ESL learners' speech samples.

1.2 Research Questions

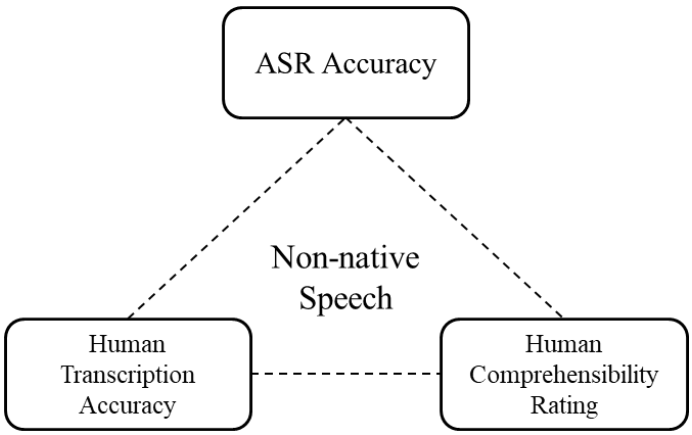

Figure 1.2 Scope of the Current Research

Figure 1.2 presents the scope of the current study. The goal of the current study is to investigate the usefulness of two state-of-the-art ASR systems in providing pronunciation feedback to Korean learners of English. The ASR systems' usefulness is evaluated in terms of the ASR's accuracy rate compared to human transcribers, and the strength of correlation between the ASR accuracy rate and human measures of intelligibility and perceived comprehensibility. This will provide useful insight into how much modern ASR systems' recognition errors correspond to that of human listeners when it comes to recognizing non-native speech. Furthermore, the current study compares two publicly available ASR systems that have different characteristics. One ASR is widely used for recognizing native speakers' speech, while the other ASR is optimized for recognizing both native and non-native (L1 Korean) speakers' English speech. Some selected examples illustrating the differences between the two ASR systems' recognition

results are discussed.

The current study examines the following main research questions:

1. Do ASR and human listeners achieve a comparable level of transcription accuracy for Korean EFL learners' speech?
2. To what extent does the ASR transcription accuracy of non-native speech correlate with human listener recognition accuracy (or intelligibility score) and comprehensibility judgement of utterances?
3. How do the two ASR systems used in this study differ in terms of recognition accuracy and correlations with human measures? What might possibly have contributed to such performance differences between the two ASR systems?

1.3 Organization of the Thesis

The current thesis is organized as follows. Chapter 2 provides a general review of previous research on ASR used for L2 pronunciation instruction. Chapter 3 discusses the methodology and data analysis for the present study. Chapter 4 reports the results of the current study, followed by Chapter 5 discussing the major findings of the study in relation to the research questions posed in this study. Lastly, Chapter 6 concludes the thesis with a summary of the findings, limitations, and suggestions for future studies.

Chapter 2. Literature Review

This chapter provides a review of previous literature on L2 pronunciation instruction and the use of ASR technology for the provision of pronunciation feedback.

2.1 Theoretical Framework

2.1.1 Nativeness vs. Intelligibility Principle

According to Levis (2005), two contradictory principles, namely the Nativeness and the Intelligibility Principles, have influenced pronunciation research and teaching. The Nativeness Principle argues that "it is both possible and desirable to achieve native-like pronunciation in a foreign language," while the Intelligibility Principle argues that "learners simply need to be understandable" (p. 370). The latter recognizes that having a foreign accent does not necessarily impair successful communication (Munro & Derwing, 1995a; Derwing & Munro, 2015), and underscores that it is more important to focus on features that have a big impact on understanding (Brown, 1988).

Levis (2020) clarifies that the Nativeness Principle relates to the concept of *accentedness*, while the Intelligibility Principle actually includes the two concepts of *intelligibility* (actual understanding) and *comprehensibility* (the ease of understanding) proposed by Munro and Derwing (1995a). He strongly argues that the Intelligibility Principle is a superior way to think about pronunciation teaching and learning due to its implications. In particular, it recognizes the great strengths

that non-native teachers can bring to the teaching of L2 pronunciation, as opposed to the Nativeness Principle that views non-native teachers as deficient models of L2 speech. While the Nativeness Principle implies that only certain native accents (such as General American or Standard Southern British) are truly acceptable, the Intelligibility Principle encourages learners to use or develop their own accents, adjusting them, when necessary, in different contexts to achieve intelligibility. In other words, if pronunciation is intelligible and comprehensible, then the Intelligibility Principle says that it does not need to be taught.

2.1.2 Definition and Operationalization of Accentedness, Comprehensibility, and Intelligibility

There have been many L2 pronunciation studies exploring the constructs of accentedness (also called foreign accent or native-likeness), comprehensibility, and intelligibility. Munro and Derwing (1995a) first introduced the interrelated but partially independent dimensions of pronunciation in their influential paper *Foreign Accent, Comprehensibility, and Intelligibility in the Speech of Second Language Learners*. These three concepts have become the basis of not only Munro and Derwing's subsequent research (Derwing & Munro, 1997; Derwing, Munro, & Wiebe, 1998; Derwing & Munro, 2005; Derwing & Munro, 2009; Munro & Derwing, 1999; Munro & Derwing, 2006), but also have had a massive impact on the field as a whole.

Table 2.1 presents a summary of the three constructs and how they are operationalized in related literature. *Accentedness* refers to a listener's perception of

"how closely the pronunciation of an utterance approaches that of a native speaker" (Munro & Derwing, 1995a, as cited in Kennedy & Trofimovich, 2008, p.461). It is typically evaluated by human raters using a 9-point Likert scale, where 1 = no foreign accent, and 9 = very strong foreign accent (Munro & Derwing, 1995a, p.79).

While foreign accent is quite straightforward in terms of definition and operationalization, the two related concepts, comprehensibility and intelligibility, are often confused for one another. *Comprehensibility* is defined as a listener's perception of how easy or difficult it is to understand a given L2 speech (Derwing & Munro, 2009). Listeners assign comprehensibility judgments using a 9-point Likert scale, where 1 = very easy to understand and 9 = very difficult to understand (Derwing & Munro, 2005; Kennedy & Trofimovich, 2008; Munro & Derwing, 1995a; O'Brien, 2014), or vice versa, 1 = very difficult/hard to understand and 9 = very easy to understand (Trofimovich & Isaacs, 2012; Isaacs & Trofimovich, 2012; Isaacs & Thomson, 2013). The current study chooses the latter type of scale because it corresponds with the intuition that high comprehensibility means that an utterance is relatively easier to understand.

Intelligibility refers to "the extent to which a speaker's message is actually understood by a listener" (Munro & Derwing, 1995a, p.76). Although there is no universally accepted method of assessing it, Munro and Derwing used a listener transcription technique to measure intelligibility. They asked native listeners to transcribe in standard orthography exactly what they had heard, word for word.

The intelligibility score is calculated as the percentage of words correctly transcribed per utterance.

Table 2.1 Definition and Measure of Accentedness, Comprehensibility, and Intelligibility

Term	Definition	Measure
Accentedness	A listener's perception of how closely the pronunciation of an utterance approaches that of a native speaker	Scalar judgment task 1 = no foreign accent, 9 = very strong foreign accent
Comprehensibility	A listener's perception of how easy or difficult it is to understand an utterance	Scalar judgment task 1 = very difficult to understand 9 = very easy to understand
Intelligibility	The extent to which a speaker's message is actually understood by a listener	Transcription task % of words correct per utterance

Although we presented a 9-point scale as the typical operationalization of accentedness and comprehensibility, the issue of the optimum number of points and score range for rating scales have been controversial (Bendig, 1953; Matell & Jacoby, 1971; McKelvie, 1978). We can roughly summarize this controversy by Miller's (1956) dictum, "the magical number seven, plus or minus two" which represents limits on human capacity to process information. Therefore, most studies used a 5, 7 or 9-point scale to measure perceived accentedness and comprehensibility.

2.2 Automatic Speech Recognition for L2 Pronunciation Feedback

2.2.1 Using ASR in L2 Classroom Pronunciation Exercises

As ASR systems have reached very high levels of recognition accuracy, there is a growing body of research on the actual use of ASR systems in L2 classrooms for pronunciation exercises. Wallace (2016) used transcriptions of the Google Web Speech API to transcribe students' speech and asked the students to correct and mark the transcript as a way of reflecting on their own pronunciation. From this process, students could gain an understanding of where it is possible to improve their delivery, including words they might not be pronouncing with high intelligibility. These exercises embrace the errors of an ASR system as an opportunity for students to check their own pronunciation and find what parts of their speech they must improve. For example, the system recognized the speech as "the find" when the intended speech was "define." A likely reason for this error was the learner's dentalization of the initial /d/. With the instructor's guidance, students could make inferences as to why the speech was not recognized correctly and receive feedback on what could be done to improve recognition.

2.2.2 Evaluating the Usefulness of ASR Systems for Pronunciation Feedback

As the use of ASR systems for pronunciation practice and feedback is increasing, there is also a growing need to evaluate the usefulness of state-of-the-art ASR systems for pronunciation practices. Many pronunciation studies explored the

potential of using ASR dictation systems to provide L2 learners with pronunciation feedback (Ashwell & Elam, 2017; Coniam, 1999; Derwing et al., 2000; McCrocklin, 2016, 2019; McCrocklin et al., 2019; Mroz, 2018; Wallace, 2016).

Coniam (1999) assessed the accuracy of Dragon System's Naturally Speaking for 10 Cantonese speakers who read a passage consisting of 1,000 words to the computer. He assessed the software's accuracy of recognizing Cantonese-accented speech by counting the number of words, phrases, and other speech units recognized correctly by the computer, and concluded that the system was considerably less effective in recognizing Cantonese speakers' English compared to that of native speakers. Although the software was not yet usable by ESL learners, the study pointed out important pedagogical implications of the system for providing corrective feedback. It has been suggested that if a more highly developed ASR system is used, learners may view the system's errors as an indication of their mispronunciation needing correction. The software used in Coniam's (1999) study was a readily available ASR package that was not specifically designed for ESL learners but were nevertheless purchased by some ESL learners for the purpose of practicing English.

Following up on this study, Derwing et al. (2000) proposed two criteria for assessing the usefulness of ASR software in providing corrective feedback to ESL learners. First, the software should be able to recognize ESL speech at an acceptable level. Second, the software's recognition errors must result from pronunciation errors that also reduce human listener intelligibility. With these two

criteria, Derwing et al. (2000) assessed the accuracy of Dragon System's Naturally Speaking for 20 high-proficiency English learners whose native language was either Cantonese or Spanish. The speakers read 60 true/false sentences which were carefully designed sentences extensively used in other studies of the intelligibility of accented speech (Munro & Derwing, 1995b). False sentences were included to make sure listeners could not rely on world knowledge to recognize the unclear parts of the utterance. For the listening experiment, two sentences were selected (one true, one false) from each of the 30 speakers (20 non-native speakers, 10 native speakers) resulting in a total of 60 sentences. The human listeners transcribed these 60 sentences and rated them for comprehensibility and accentedness on a 9-point scale. The results showed that the software's recognition score for the non-native speech was 24% to 26% lower than that of human listeners and 9% lower for recognizing native speech. In this study, the human listeners rated the Cantonese speakers as significantly less comprehensible compared to the Spanish speakers, but they did not show any significant difference in the intelligibility scores of the two groups. The Pearson correlations between the software's recognition score and the human listeners' intelligibility score and ratings of comprehensibility and accentedness were not significant and close to zero, indicating that the software's recognition of non-native speech was not related to human judgment. Expert raters also marked each sentence for segmental errors, but results showed no significant relationships between the software's recognition accuracy and percentage of segmental errors. Derwing et al. (2000) states that for a L2 learner to learn from computer feedback, the software should be as humanlike

as possible producing errors in places where human listeners are likely to misunderstand. This study further suggests that speech recognition software be carefully evaluated using the two criteria proposed in this study to be of benefit to ESL learners.

After these two studies, unfortunately, research into ASR dictation programs in providing pronunciation feedback largely halted, with attention shifting to CAPT programs that integrate ASR (Cucchiarini & Strik, 2018). Recently, however, interest in the use of ASR dictation programs as a pronunciation feedback tool for L2 learners has resurged (McCrocklin & Edalatishams, 2020). ASR dictation programs, such as Google Voice Typing, are free, accessible to anyone, and flexible in that students can choose their own text for dictation. Furthermore, research has shown that the use of ASR-based dictation programs for L2 pronunciation practice provides a range of benefits. The transcripts make it possible to assess human intelligibility (Mroz, 2018), locate individual word and sound errors (McCrocklin, 2019b), and detect error patterns across words (McCrocklin, 2019c; Wallace, 2016), which can help learners improve not only segmental accuracy (McCrocklin, 2019a) but also their overall intelligibility (Mroz, 2020). Daniels and Iwago (2017) investigated the accuracy of two prominent cloud-based speech recognition engines, Apple's Siri and Google Speech Recognition, to determine which engine was more accurate at transcribing Japanese learners' English speech. The results of the study revealed that Google's accuracy (82%) was significantly higher than that of Siri (66.9%) for recognizing L2 English speech. Despite the renewed interest, little research has examined the

issues of how much the ASR accuracy levels improved for non-native speech and how much the ASR errors resemble misunderstandings of human listeners. Table 2.2 summarizes the previous studies along with the current study.

Table 2.2 Comparison of the Current Study with Previous Studies

	Derwing et al. (2000)	McCrocklin & Edalatishams (2020)	Current Study (2021)
ASR	Dragon Naturally Speaking	Google Voice Typing	Google Web Speech API & ETRI Open API
Speakers	10 Cantonese, 10 Spanish	10 Chinese, 10 Spanish	151 Korean
Data	read speech (60 true/false sentences)	read speech (60 true/false sentences)	read speech (1505 sentences)
Listeners	41 Native listeners (Canadian)	37 Native listeners (USA)	4 Native listeners 4 Non-native listeners
Human measures	accentedness, intelligibility, comprehensibility	accentedness, intelligibility, comprehensibility	intelligibility, comprehensibility
Correlation with ASR accuracy	no significant correlation	significant correlation with L1 Chinese, no significant correlation with L1 Spanish	

2.3 The Current Research

While there are many previous studies that examined the usefulness of ASR systems for L2 pronunciation feedback, most studies used ASR systems that are built to generally recognize native speech such as Windows Speech Recognition (McCrocklin et al., 2019), Google Voice Typing (McCrocklin et al., 2019), and Google Web Speech (Wallace, 2016; Ashwell & Elam, 2017). To go one step further, in the current study, we used an ASR that models the characteristics of non-native speech as well as Google Web Speech to compare the results between the two ASR systems and human listeners. We would like to see whether the ASR that models non-native speech shows a significant enhancement in recognition accuracy compared to the ASR that does not specifically model non-native speech. We would also like to examine whether using the ASR optimized for non-native speech produces higher correlations with human intelligibility and perceived comprehensibility.

The current research used a 5-point scale to rate comprehensibility as opposed to the 9-point scale used in previous studies. According to Isaacs & Thomson (2013), there were no significant differences in mean comprehensibility scores obtained using 5- versus 9-point scales. Moreover, raters who used the 9-point scale reported more difficulty in meaningfully differentiating between "so many numbers," particularly in the mid-scale range. Therefore, in this study, we used a 5-point scale to alleviate the processing burden of our raters. Moreover, although the previous studies examined the ASR's correlations with accentedness,

intelligibility and comprehensibility, in the current study we only examined intelligibility and comprehensibility since accentedness was not our point of interest.

This study provides insight into how much modern ASR systems' accuracy rates correlate with human listeners' intelligibility score and comprehensibility ratings of Korean learners' English speech. This can be helpful in evaluating the usefulness of these ASR system as a pronunciation practice tool for Korean EFL learners. In addition, this study takes one step further from the previous studies in the sense that an ASR optimized for recognizing non-native speech was used and compared with a general ASR system. This is expected to enrich our understanding of the potential effect that different types of ASR systems can have on their capabilities to provide useful feedback to non-native learners. Finally, this study also includes examples of the ASR results in comparison to human listeners, which could suggest how we could utilize or improve the ASR systems for L2 pronunciation feedback.

Chapter 3. Methods

This chapter describes the research methods used to collect and analyze data for the current study. Detailed descriptions of the corpus used in this study are provided, which are followed by background information about listeners (transcribers/raters), the data collection procedure, the two ASR systems, and data analysis methods.

3.1 Data

We used the ETRI (Electronic and Telecommunications Research Institute) English Read Speech Corpus, which consists of read-aloud English speech produced by Korean learners of English. The original corpus has a total of 30,200 utterances (261,720 words; mean length of 8.67 words per utterance) with a total duration of 31.1 hours of speech. A total of 151 learners (79 males and 72 females) each read 200 sentences. Generally, the sentences each learner read did not overlap, except for 3,876 sentences repeated across learners. This corpus is actually divided into Set 1 (101 learners' data) and Set 2 (50 learners' data) with no other big difference than the time we acquired the data. For the 101 learners in Set 1 whom we have information on, the age ranges from 19 to 54 years, with a mean age of 26.81 years (SD = 6.97) at the time of recording. All of the learners were native speakers of Korean who had learned English as a second/foreign language no later than at the age of 13.

In this study, we used a subset of the original corpus for ASR/human transcription and comprehensibility rating purposes. This subset ("assessment" set)

consists of a total of 1,505 utterances (12,100 words) with a total duration of 1.45 hours of speech from the 151 learners mentioned above. The assessment set was selected randomly, solely making sure that the number of words in the utterance was approximately 8 words, so that we can perform various analyses on the data with less effect of sentence length. We selected 5 utterances from each of the 101 speakers in Set 1, and 20 utterances from each of the 50 speakers in Set 2. This was because the speakers in Set 1 were mostly advanced speakers, while the speakers in Set 2 were intermediate to advanced speakers who exhibited more pronunciation variations that are commonly observed in non-native speech. The mean length of the utterances in the assessment set was 8.45 words (SD = 1.12). The following is an example sentence: "You should find what you're looking for there."

3.2 Listeners (Transcribers/Raters)

In this study, native and non-native human listeners performed the work of transcription and comprehensibility rating for the learner assessment set. We recruited a total of 8 participants, carefully selected from applicants. Four participants were native speakers of English (NS) currently residing in Korea who had stayed in Korea for 1 to 3 years. Ideally, we wanted native speakers with little or no exposure to Koreans' English. However, due to the shortage of applicants meeting such requirements, we selected those who had the least length of residence in Korea, and who reported low proficiency in Korean listening and speaking. Three out of the four native speakers had experience in teaching Korean learners of English, which is very common for English native speakers living in Korea. The

other four participants were native speakers of Korean (non-native speakers of English; NNS) with an advanced level of English proficiency. Of the four Korean participants, two had experience living abroad in an English-speaking country for an extended period of time (longer than three years), but the other two had no experience living abroad. Nevertheless, the two non-native participants with no experience living abroad reported that they used English very frequently in their daily lives. All of the non-native participants self-rated their English listening and speaking ability. They tended to rate themselves as high in these areas (i.e., assigning a 4 or a 5 on a 5-point rating). Detailed background information about the participants can be found in Tables 3.1 and 3.2.

Table 3.1 Background Information of Native Participants

	NS_1	**NS_2**	**NS_3**	**NS_4**
Gender	female	male	female	male
Age	25	30	26	28
Nationality	Australia	USA	South Africa	Canada
Residence in Korea	1.5 years	1 year	3 years	3 years
Experience teaching Koreans	Full time (1.5 years)	Part-time (online)	Full time (3 years)	None
Korean listening*	1	2	1	3
Korean speaking*	1	2	1	2

* Self-assessment on a 5-point scale

Table 3.2 Background Information of Non-Native Participants

	NNS_1	NNS_2	NNS_3	NNS_4
Gender	male	female	female	male
Age	28	28	26	30
Residence in English-speaking country	None	None	7 years[1]	3 years[2]
Frequency of English use[3]	5	5	4	2
English listening[3]	4	5	5	4
English speaking[3]	5	4	5	4

[1] 2 years in New Zealand (age 7) & 5 years in USA (age 17)
[2] 3 years in Canada (age 10)
[3] Self-assessment on a 5-point scale

All 8 participants were trained regarding the rules and procedures of the transcription and comprehensibility rating by the researcher. The participants were first invited to an online orientation in which the researcher presented the guidelines of the transcription and comprehensibility rating sessions. After the orientation, the participants were asked to complete transcribing and scoring a sample task of both sessions and were provided feedback about their performance along with additional guidelines if needed. When each of the 8 participants was judged to have met the performance criteria, she (or he) was asked to take part in the main transcription and rating sessions. The reference text for the learners' utterances was provided only after the completion of the transcription task, so transcriptions were purely based on how the participants recognized the speech. Each participant was given a list of audio files in a different, randomized order. The participants carried out the transcription and rating for two weeks. Although they

could carry out the work at their own desired pace, 9 days was the recommended time limit for the completion of transcription, and the remaining 5 days for the comprehensibility rating. Because the participants had to transcribe and rate a total of 1,505 utterances, they were strongly advised to divide the workload to 250~300 utterances per day. These two recommendations were given to prevent the participants from cramming before the deadline and to ensure the high quality of the work. All of the participants generally followed the recommended due date and submitted both types of work by the final deadline.

3.3 Transcription and Rating

3.3.1 Transcription Session

All of the participants were asked to transcribe the speech files in standard English spelling. They were told that there is no right or wrong answer and thus their transcriptions should directly reflect their own understandings. If they could not understand a word, a phrase, or a sentence, they were asked to first spell out the word(s) in a way that best reflected what they had just heard regardless of whether they sounded like a string of real words or non-words. When they did not sound like real words, they were instructed to provide an alternative transcription consisting of real words only. In order to write an alternative transcription, they had to go through a process of matching unintelligible parts to plausible real words in a way similar to how ASR systems work.

3.3.2 Comprehensibility Rating Session

In the rating session, the participants were asked to rate the comprehensibility of each fragment on a 5-point scale. In this study, comprehensibility was defined to be a listener's perception of how easy it is to understand a given L2 speech (Derwing & Munro, 1997). For this study, we adopted the scale used in Isaacs et al.'s (2017) study, which defined comprehensibility ranging from Level 1 ("speech is painstakingly effortful to understand or indecipherable") to Level 5 ("speech is effortless to understand"), as shown in Table 3.3. A higher comprehensibility score means that the speech is easier to understand, and vice versa. The reference text that the learners were asked to read aloud was also provided to the raters in the rating session to help them to compare the original text to what they transcribed based on their understanding.

Table 3.3 Comprehensibility Rating Scale from Isaacs et al. (2017)

Comprehensibility Level	Overall description of comprehensibility
5	*Speech is effortless to understand* Errors are rare and do not interfere with the message
4	*Speech requires little effort to understand* Errors minimally interfere with the message
3	*Speech requires some effort to understand* Errors somewhat interfere with the message
2	*Speech is effortful to understand* Errors are detrimental to the message
1	*Speech is painstakingly effortful to understand or indecipherable* Errors are debilitating to the message

3.4 Automatic Speech Recognition Systems

3.4.1 Google Web Speech API

To obtain ASR results, we used the Google Web Speech API (Application Programming Interface) available in the SpeechRecognition library in Python (Zhang, 2017). Although there were 8 different speech recognition APIs included in the library (e.g., Microsoft Bing Voice Recognition, IBM Speech to Text, etc.), we selected the Google speech recognition as it did not require any API key and thus was more accessible to researchers. The L2 audio files (assessment set) were input into the system, and we obtained as output the speech-to-text transcription results of the Google ASR system.

Word error rate (WER) is a common metric for evaluating the performance of a speech recognition system which is calculated in the following way:

$$WER = \frac{Substitutions(S) + Deletions(D) + Insertions(I)}{Total\ number\ of\ words\ in\ the\ reference(N)} \times 100 \qquad (1)$$

The ASR recognition accuracy used for the subsequent analyses was computed by subtracting the word error rate from 100%. These accuracy rates were obtained because it can better represent the quality of transcription in a conceptual sense. We aligned the Google ASR results with the reference text that the L2 learners read from and calculated the WER and recognition accuracy rate per utterance using 'compute-wer' in Kaldi (Povey et al., 2011). Some textual adjustments had to be made to make these measures as accurate as possible, which included transcription

manipulations, such as expanding common English contractions (e.g., he's → he is), changing all numbers to letters (e.g., 7 → seven), and editing any other cases in which the transcription was clearly right but the style of writing was different from the reference text.

3.4.2 ETRI Open API

The ETRI Open API was created and distributed by the Electronics and Telecommunications Research Institute (ETRI) in South Korea. It employs state-of-the-art artificial intelligence technology and provides a high-performance speech recognition service for ten languages (e.g., English, Korean, Japanese, Chinese, Spanish, French, etc.). The input speech data recorded by a user is passed on to the speech recognition server, which converts the speech into text. The ASR is optimized to recognize the English utterances of Korean learners as well as native speakers (Chung et al., 2014; Lee et al., 2014). To acquire high accuracy in recognizing Korean learners' English speech, a database of sentences uttered by Korean speakers was created and used in the ETRI Open API in such a way that the pronunciation characteristics of Korean learners were reflected in the acoustic model (Chung et al., 2014b). Moreover, common pronunciation errors produced by Korean learners were modeled in order to generate a pronunciation dictionary that was adapted to Koreans' speech. Additional modeling on the common grammatical errors produced by Korean learners (e.g., singular/plural form of nouns) was reflected in the language model to make an ASR that is robust to learners' grammatical errors (Kwon et al., 2015).

3.5 Data Analysis

Before calculating the recognition/transcription accuracy rates of the ASR systems and human listeners, we applied preprocessing techniques using Natural Language Toolkit (NTLK) and python libraries (i.e., pycontractions, spellchecker). First, we had to correct the typos made by human listeners because they did not represent difficulties in understanding. We used a simple spell-checking algorithm called 'pyspellchecker' (Lison & Tiedemann, 2016) to automatically identify typos in the listener transcription data, and the typos were manually corrected by the researcher. Secondly, expressions such as numbers (e.g., 1000 → one thousand) and time (9:00 → nine o'clock) were all converted to words, and British English spellings were converted to American English spellings (e.g., colour → color) for direct comparison. Finally, English contractions were all expanded (e.g., I'll → I will) using a Python library called 'pycontractions' and ambiguous contractions such as "I'd" that could be expanded as either "I would" or "I had" were manually checked and corrected by the researcher.

After data cleaning, we calculated the recognition accuracy rate (intelligibility score) of human transcriptions using the same method as the one used for the ASR results, which was previously described in Subsection 3.4.1. We aligned each of the listener's transcriptions with the reference text and computed a word error rate for each utterance for each listener. The recognition accuracy rate was obtained simply by subtracting WER from 100%. Since 8 listeners transcribed all utterances, the recognition score for each utterance was calculated by averaging

all accuracy rates across the 8 listeners. We also obtained the sub-average of native and non-native listeners per utterance in order to examine differences across the two groups of listeners. For the human comprehensibility ratings, the average of all 8 listeners was calculated per utterance and sub-averages were also obtained for the two groups of native and non-native listeners. For each of the 1,505 utterances from the Korean EFL learners, the accuracy rates of the Google ASR, ETRI ASR, and human listeners' recognition accuracy rates and comprehensibility ratings were inserted into separate columns in a Microsoft Excel spreadsheet. All the following statistical analyses were conducted using IBM SPSS (IBM, 2019) Statistics Version 26.0.

For the statistical analysis, we first obtained descriptive statistics of the accuracy rates of the two ASR systems, the human recognition accuracy rates (intelligibility scores), and the human-assigned comprehensibility ratings. Then the inter-rater reliability coefficients were obtained for both human recognition accuracy rates and comprehensibility ratings by computing intra-class correlation coefficients (ICC; McGraw & Wong, 1996). ICC was selected because it is one of the most commonly-used statistics for assessing inter-rater reliability for two or more raters. ICC estimates and their 95% confidence intervals were calculated using IBM SPSS (IBM, 2019) Statistics Version 26.0 based on a mean-rating (k = 8), consistency, two-way random-effects model. A two-way random-effects model was selected since all subjects were rated by the same set of raters (fully crossed design). We examined consistency instead of absolute agreement of ratings since it was more important that raters should provide scores that were similar in rank

order rather than absolute agreement in value. Also, comprehensibility rating is itself a subjective judgment that is not expected to reach an absolute agreement. We selected the "mean of *k* raters" measurement, instead of single measure, as the unit of analysis since the ICC is meant to quantify the reliability of the ratings based on averages of ratings provided by several raters in this study. Inter-transcriber reliability coefficients were computed for transcription accuracy scores in three different sets of transcription accuracy data: (a) all of the raters, (b) for four native raters only, and (c) for four non-native raters only. Inter-rater reliability coefficients were computed in the same way for the comprehensibility ratings.

To address the first research question as to whether the Google/ETRI ASR and human listeners achieve a similar level of recognition accuracy, we conducted an independent-samples t-test comparing the means of ASR and human recognition accuracy rates. To answer the second research question regarding the relationship among ASR accuracy, human intelligibility, and comprehensibility measures, Pearson correlation coefficients were computed among these measures. This was done to see the direction and magnitude of correlation between the ASR accuracy and human measures. Finally, to address the third research question, some examples of the two ASR and human transcription results were selected and compared to each other.

Chapter 4. Results

This section reports the results of analyses of the collected data, including descriptive statistics about human transcriber/ASR accuracy scores and human listener ratings, inter-rater reliability coefficients, and coefficients of correlation among human accuracy rates, ASR accuracy rates, and human comprehensibility ratings.

4.1 Human Measures

4.1.1 Descriptive Statistics

In this section we report the descriptive statistics of human listeners' transcription accuracy rates and comprehensibility ratings of Korean EFL learners' utterances. Table 4.1 shows the means and standard deviations of human listeners' transcription accuracy rates (or intelligibility score) and comprehensibility ratings. Human listeners showed a mean intelligibility score of 95.29%, which means that they transcribed the non-native speech with a 95.29% accuracy. In other words, they were likely to mis-transcribe approximately 5 out of 100 words. The overall average of comprehensibility ratings given by human listeners was 4.39 on a 5-point scale. In our scale, a score closer to 5 indicates that the speech would be easier to understand.

Figures 4.1 and 4.2 are the histograms showing grouped frequency distributions of transcription accuracy rates and comprehensibility ratings with estimated normal distribution (or bell) curves for these measures. The intelligibility

score distribution in Figure 4.1 is skewed to the left (or negatively-skewed), indicating that most of the sentences (or utterances) produced by the L2 learners in the current study were highly intelligible to the human listeners. The mean transcription accuracy rate ranged from 32.81% to 100%, with almost one third of the sentences transcribed with 100% mean accuracy. The distribution curve of comprehensibility ratings in Figure 4.2 was also skewed to the left, suggesting that most of the non-native utterances used in this study were rated as highly comprehensible (i.e., very easy to understand) by the human listeners. The majority of sentences received a mean comprehensibility rating between 4 and 5 on a 5-point scale. These suggest that the sentences produced by the L2 learners in the current study represent non-native English speech that is both highly intelligible and comprehensible to human listeners.

Table 4.1 Means and Standard Deviations of Human Listeners' Transcription Accuracy (Intelligibility) and Comprehensibility Ratings

	Intelligibility (%)			Comprehensibility Rating		
	NS	NNS	All	NS	NNS	All
M	95.19	95.39	95.29	4.54	4.23	4.39
SD	7.68	7.12	6.79	.37	.49	.38

Note. 100 - word error rate (%) for intelligibility
1-5 rating bands for comprehensibility rating
NS: native listener, NNS: non-native listener

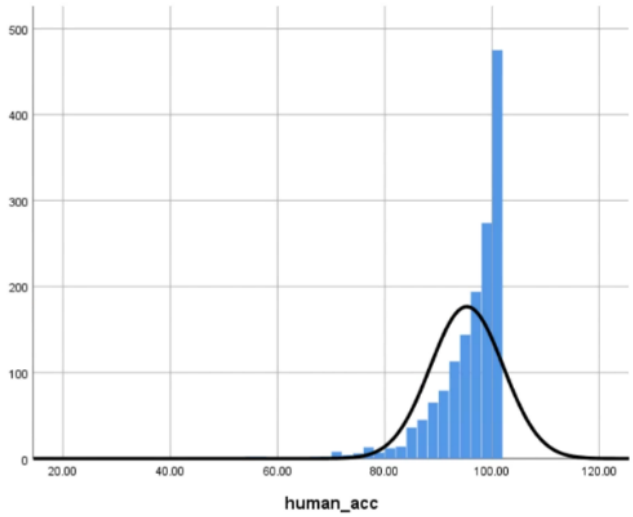

Figure 4.1 Histogram of Human Listeners' Mean Transcription Accuracy (Intelligibility) for Each Sentence

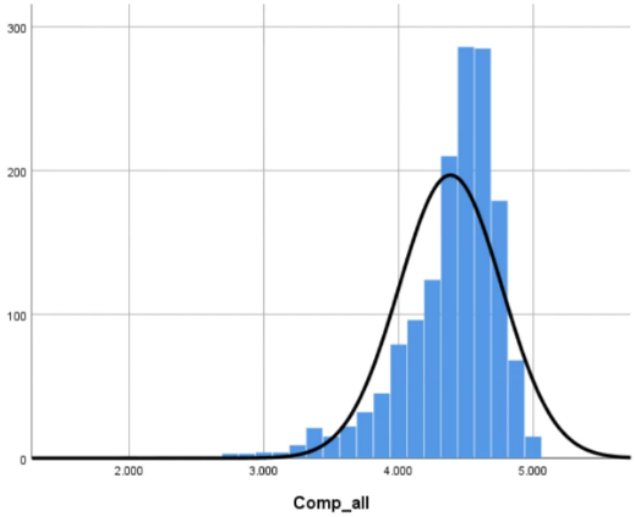

Figure 4.2 Histogram of Human Listeners' Mean Comprehensibility Ratings for Each Sentence

Table 4.2 reports the descriptive statistics of the transcription accuracy of individual human listeners. The transcription accuracy of the individual listeners ranges from 92.12% to 96.76%, showing a 4.64% difference between the listener with the highest and lowest intelligibility. Overall, Native Listener 2 (92.12%) and Non-native Listener 1 (93.52%) show lower accuracy rates compared to the other listeners ranging from 95.78% to 96.76%. This suggests that there exists a certain degree of variation among different listeners in recognizing the non-native English speech.

Table 4.2 Transcription Accuracy of Individual Human Listeners

Transcription Accuracy								
	Native (NS)				Non-native (NNS)			
	ns1	ns2	ns3	ns4	nns1	nns2	nns3	nns4
Mean	95.86	92.12	96.04	96.76	93.52	95.80	95.78	96.47
SD	9.12	13.59	10.30	7.87	11.24	9.78	9.19	8.57

Table 4.3 Comprehensibility Ratings of Individual Human Listeners

Comprehensibility Ratings (5-point scale)								
	Native (NS)				Non-native (NNS)			
	ns1	ns2	ns3	ns4	nns1	nns2	nns3	nns4
Mean	4.80	3.87	4.94	4.57	3.41	4.60	4.01	4.87
SD	.47	.74	.32	.64	.87	.72	.79	.39

Table 4.3 displays the descriptive statistics of the comprehensibility ratings of individual listeners. The mean comprehensibility ratings range from 3.41 to 4.94 across listeners, showing a difference of 1.53 between the most lenient and severe listener. Three of the listeners (NS 1, NS 3, NNS 4) tended to assign higher ratings than the rest of listeners. The mean comprehensibility ratings from these three listeners were 4.8 or above on a 5-point scale, and their standard deviations were relatively smaller. This may be due to the ceiling effect, given that the highest possible score was 5 on the scale. In other words, the non-native speech data used in the current study might have exhibited low variability in terms of comprehensibility ratings, because most of the ratings were clustered towards the higher end of the rating scale.

When we examined each listener's recognition accuracy rates and comprehensibility ratings together, the listeners who gave the lowest comprehensibility ratings (NS 2: 3.87, NNS 1: 3.41) corresponded to the listeners who showed the lowest accuracy in transcribing the speech (NS 2: 92.12%, NNS 1: 93.52%). These two listeners showed both lower intelligibility scores and comprehensibility ratings compared to the rest of the listeners, indicating that they had relatively more difficulty in recognizing and understanding the non-native English speech. These possible outliers are discussed later in this chapter and the following Discussion chapter.

4.1.2 Inter-Rater Reliability

This section reports measures of inter-rater reliability obtained for the transcription accuracy rates (intelligibility score) and comprehensibility ratings of human

listeners. The inter-rater reliability (or inter-rater agreement) was assessed by using intra-class correlation coefficients (ICC; McGraw & Wong, 1996). ICC is designed to assess the degree of consistency across listeners in their transcription and comprehensibility judgement of utterances.

Table 4.4 Intra-Class Correlation (ICC) of Human Recognition Accuracy

Mean *k* raters	Intraclass Correlation	95% Confidence Interval		F Test with True Value 0			
		Lower bound	Upper bound	Value	df1	df2	Sig
All (8)	.83**	.81	.84	5.78	1504	10528	.000
NS (4)	.72**	.69	.74	3.53	1504	4512	.000
NNS (4)	.71**	.68	.73	3.43	1504	4512	.000

Table 4.5 Intra-Class Correlation (ICC) of Human Comprehensibility Ratings

Mean *k* raters	Intra-class Correlation	95% Confidence Interval		F Test with True Value 0			
		Lower bound	Upper bound	Value	df1	df2	Sig
All (8)	.73**	.71	.75	3.76	1504	10528	.000
NS (4)	.57**	.53	.60	2.31	1504	4512	.000
NNS (4)	.62**	.59	.65	2.64	1504	4512	.000

Table 4.4 shows the inter-rater reliability of listeners' transcription accuracy (i.e., intelligibility score) for native, non-native, and all listeners. In general, based on the 95% confidence interval of the ICC estimate, values less than 0.5 indicate poor reliability, between 0.5 and 0.75 indicate moderate reliability, between 0.75 and 0.9 indicate good reliability, and values greater than 0.9 indicate excellent reliability (Koo & Li, 2016). The ICC of all listeners' transcription

accuracies showed a 'good' level of reliability (ICC = 0.83), while the reliability coefficients of native (ICC = 0.72) and non-native listeners (ICC = 0.71) computed separately both showed a 'moderate' level of reliability. The reliability coefficient turned out to be slightly higher for native transcribers than for non-native transcribers. In addition, using both native and non-native listeners increased the level of inter-rater reliability in terms of transcription accuracy.

The ICC results of comprehensibility ratings for native, non-native, and all listeners are shown in Table 4.5. The ICCs of all raters and the native and non-native sub-groups were all in the moderate range (0.5 to 0.75); ICC = 0.73 (all 8 listeners), ICC = 0.57 (4 *native* listeners), and ICC = 0.62 (4 *non-native* listeners). This indicated that listeners had a moderate degree of agreement in rating the comprehensibility of non-native utterances. The reliability coefficient was lower for native raters than for non-native raters, which indicates that non-native raters showed more consistency among them when judging the comprehensibility of a particular utterance. The reliability coefficients of all listeners indicated that the use of both native and non-native listeners for assessing comprehensibility increased inter-rater reliability in this study.

Since the inter-rater reliability coefficients of the transcription accuracy and comprehensibility ratings all revealed a moderate to good level of reliability, further analysis was conducted using the average across all human listeners as well as across the native and non-native subgroups.

4.1.3 Intelligibility and Comprehensibility Scores of Individual Speakers (Learners)

This subsection presents the descriptive statistics of intelligibility scores and comprehensibility ratings obtained for each individual speaker or L2 learner. Figure 4.3 shows the distribution of average intelligibility scores obtained for each of the 151 speakers (learners) in our speech data. A speaker's average intelligibility score was calculated by tallying all the scores they had received from the human listeners and dividing it by the number of listeners and the number of sentences produced by the speaker. This score represents a speaker's overall intelligibility, and it varied from 85.86% to 100% with an average of 95.40%. This means that the sentences (or utterances) produced by the least intelligible speaker were transcribed with an average of 85.86% accuracy by the human listeners, while those produced by the most intelligible speaker were transcribed with an average of 100% accuracy.

A speaker's average comprehensibility rating was calculated in the same way, and its distribution across speakers is presented in Figure 4.4. The average comprehensibility ratings across speakers ranged from 3.85 to 4.93 on a 5-point scale with an average of 4.43. This means that, out of the 151 speakers in our data, the speaker who was rated as the least comprehensible received an overall average score of 3.85 from the human listeners, while the speaker who was rated as the most comprehensible received an average score of 4.93. Although our data showed an overall tendency for speakers to be highly intelligible and comprehensible, human listeners exhibited some variability in the ability to understand utterances

produced by each speaker, possibly due to different English proficiency levels of the speakers (or Korean learners of English).

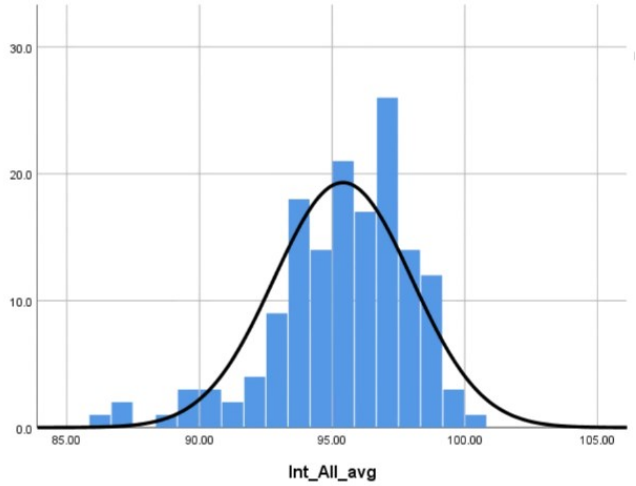

Figure 4.3 Histogram of Human Listeners' Average Intelligibility Scores for Individual Speakers

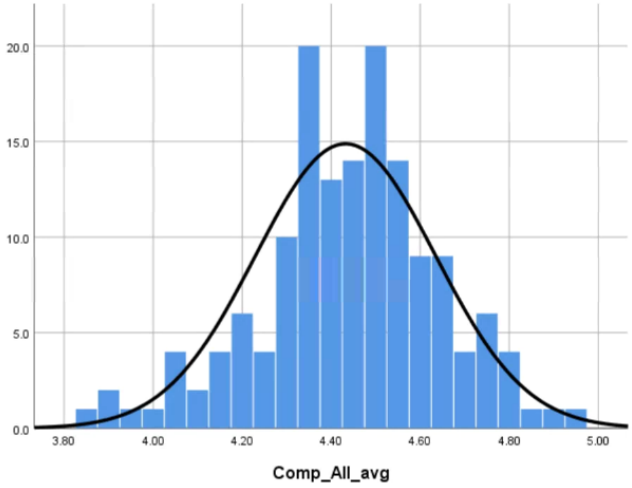

Figure 4.4 Histogram of Human Listeners' Average Comprehensibility Ratings for Individual Speakers

4.2 ASR Accuracy

In this section we report the descriptive statistics of the two ASR systems used in the current study. Table 4.6 shows the means and standard deviations of Google and ETRI ASR systems' recognition accuracy rates. The Google ASR achieved a mean accuracy rate of 85.27%, which shows that it was likely to mis-transcribe approximately 15 out of 100 words spoken by the non-native learners. The ETRI ASR exhibited a 93.77% mean accuracy rate in transcribing the non-native speech, indicating that it was likely to transcribe approximately 6 to 7 words incorrectly out of a total of 100 words.

Table 4.6 Means and Standard Deviations of Google and ETRI ASR Systems' Recognition Accuracy

	Google ASR (%)	ETRI ASR (%)
M	85.27	93.77
SD	20.81	11.46

Figure 4.5 presents the histogram of Google ASR's accuracy in recognizing each sentence produced by the non-native learners. Google's accuracy ranged from -33.33% to 100%. Although it was not common, there existed three cases in which the Google ASR showed a negative accuracy rate. This happened when the ASR recognized all the words in the reference text incorrectly and additionally produced insertions errors, which resulted in exceeding the number of words in the original text. Figure 4.6 displays the distribution of ETRI ASR's

accuracy in recognizing the non-native speech. ETRI's accuracy ranged from 22.22 to 100% across the L2 utterances. Compared to the Google ASR, the ETRI ASR's accuracy showed a smaller variation in accuracy of recognizing the L2 learners' speech, as shown by the smaller standard deviation and the distribution of the accuracy rates.

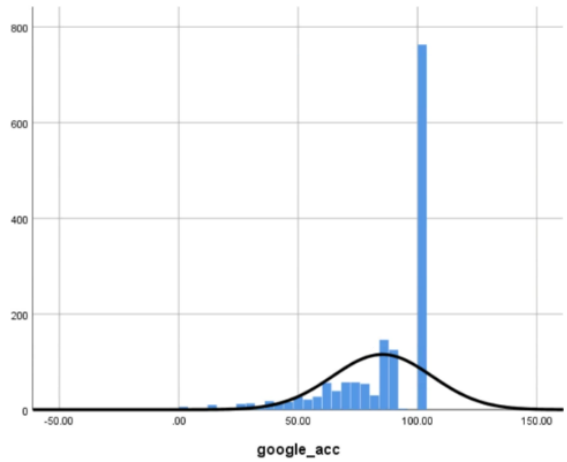

Figure 4.5 Histogram of Google ASR's Mean Transcription Accuracy for Each Sentence

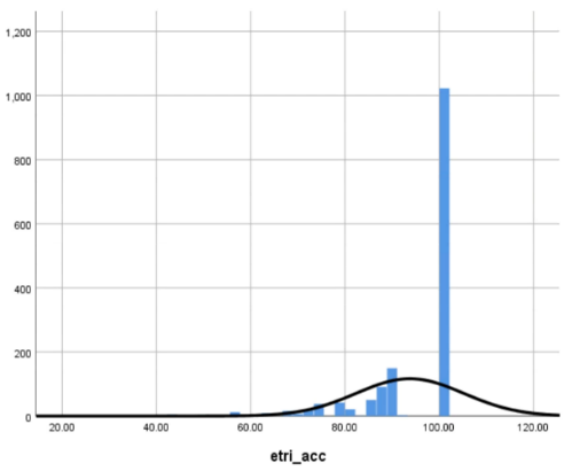

Figure 4.6 Histogram of ETRI ASR's Mean Transcription Accuracy for Each Sentence

Figures 4.7 and 4.8 are the histograms of the two ASR systems' mean accuracy rates computed for individual speakers. The average recognition accuracy for individual learners ranged from 50.67% to 100% for the Google ASR, with a mean accuracy of 87.05% (*SD* = 9.23). This means that Google transcribed the utterances of the least intelligible speaker with a 50.67% mean accuracy. The recognition accuracy of the ETRI ASR for each speaker ranged from 75.56% to 100%, with a mean accuracy of 93.62% (*SD* = 4.70). This means that the ETRI ASR transcribed the speaker who was the least intelligible with a mean accuracy of 75.56%. Overall, the Google ASR showed a bigger variability in terms of accuracy of recognizing the speech of individual L2 learners, compared to the ETRI ASR.

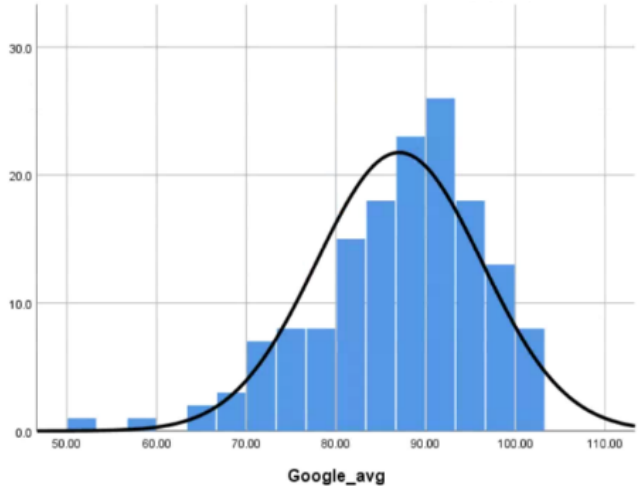

Figure 4.7 Histogram of Google's Intelligibility Scores for Individual *Speakers*

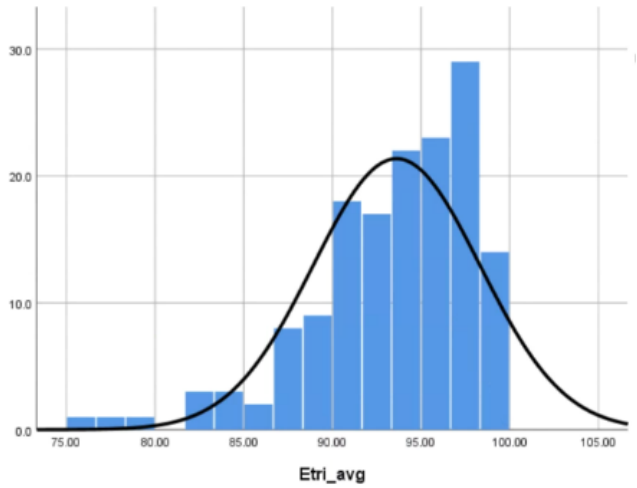

Figure 4.8 Histogram of ETRI's Intelligibility Scores for Individual *Speakers*

4.3 Comparison between Human Listeners and ASR Systems

Based on the descriptive statistics in Subsections 4.1 and 4.2, this subsection examines the mean differences between the ASR systems and average human listeners as well as among the native and non-native subgroups within human listeners. Table 4.7 summarizes the accuracy rates of the two ASR systems, human transcription accuracy (intelligibility score), and human comprehensibility ratings obtained for the current study. To investigate whether the mean differences were statistically significant, an independent-samples t-test was conducted for each pair of interest.

Table 4.7 Summary of ASR Accuracy and Human Measures of Intelligibility and Comprehensibility Rating

	ASR Systems		Human Listeners					
	Accuracy (%)		Intelligibility (%)			Comprehensibility Rating		
	Google	ETRI	NS	NNS	All	NS	NNS	All
M	85.27	93.77	95.19	95.39	95.29	4.54	4.23	4.39
SD	20.81	11.46	7.68	7.12	6.79	.37	.49	.38

First, a statistically significant difference was exhibited between the mean recognition accuracy rates of the two ASR systems used in the study. An independent-samples t-test revealed that there was a significant difference in the recognition accuracy for the Google (M = 85.27, SD = 20.81) and ETRI (M = 93.77, SD = 11.46) ASR systems, $t(2339.024) = -13.879$, $p < .001$. These results show that the ETRI ASR, which was optimized for recognizing non-native speech, showed a significantly higher level of accuracy, compared to the Google ASR system. To be more specific, the ETRI ASR outperformed the Google ASR by 8.5% for the task of recognizing the non-native data used in this study.

Secondly, there were statistically significant differences in the recognition accuracy among the two ASR systems and human listeners. An independent-samples t-test was conducted to compare the transcription accuracy of the human listeners and Google ASR system. There was a significant difference in the transcription accuracy for the human listeners (M = 95.29, SD = 6.79) and the Google ASR (M = 85.27, SD = 20.81) conditions, $t(1820.891) = 17.757$, $p < .001$. These results suggest that the Google ASR's accuracy was significantly lower than

that of human listeners when recognizing the L2 learners' speech. Specifically, our results indicated that human listeners were better than the Google ASR in transcribing the non-native speech by 10.02%. Likewise, the mean difference of the transcription accuracy in the two conditions of human listeners (M = 95.29, SD = 6.79) and the ETRI ASR (M = 93.77, SD = 11.46) was also statistically significant, $t(2445.071) = 4.429$, $p < .001$. These results suggest that the ETRI ASR's accuracy was significantly lower than that of human listeners in terms of transcribing non-native speech, with human listeners outperforming the ETRI ASR by 1.52%. Overall, the results suggest that human listeners outperformed both ASR systems in terms of correctly transcribing (recognizing) the non-native speech. However, the ETRI ASR (1.52%) had a smaller difference from human listeners in terms of recognition accuracy, as compared with the Google ASR (10.02%).

Third, the mean difference between native and non-native listeners was not significant for transcription accuracy (intelligibility) but turned out to be significant for comprehensibility ratings. The results of an independent-samples t-test indicated that there was no significant difference between native (M = 95.19, SD = 7.68) and non-native (M = 95.39, SD = 7.12) transcribers' accuracy, $t(3008) = -0.732$, $p = .464$. These results show that native and non-native transcribers showed a similar level of accuracy when transcribing the non-native speech, although non-native transcribers had a slightly higher mean intelligibility score than native listeners by 0.2%. In terms of comprehensibility ratings, an independent-samples t-test verified that there was a significant difference between native (M = 4.54, SD = 0.37) and non-native (M = 4.23, SD = 0.49) raters, $t(2809.595) = 20.091$, $p < .001$.

The mean comprehensibility rating of native listeners was 0.31 higher than that of non-native listeners. These results suggest that native listeners generally gave higher comprehensibility ratings to the non-native speech and thus were more lenient than the non-native listeners were. This is interesting because native and non-native listeners did not show any significant difference in terms of objective intelligibility scores but only showed a difference in the subjective comprehensibility ratings.

Figure 4.9 Bar Chart of the Percentage of Correctly Transcribed Sentences by Listener

Lastly, Figure 4.9 presents the percentage of the number of sentences correctly recognized by each human listener, and the two ASR systems. "Correctly recognized" means that the transcribed (human) or recognized (ASR) sentence and the reference sentence matched for every single word. From a total of 1,505 spoken

sentences, an average of 1,130 sentences were correctly recognized by human listeners. In other words, human listeners correctly transcribed 75% of the sentences without any insertion, deletion, or substitution errors. However, even among human listeners there existed quite a variation from 63.19% to 80.27%. More specifically, Native speaker 2 (63.19%) and Non-native speaker 1 (65.18%) had lower percentages of correctly transcribed sentences than the other listeners (ranging from 76.68% to 80.27%). Although both ASR systems had a lower percentage of correctly recognized sentences compared to the average of human listeners, the ETRI ASR (67.91%) showed a slightly higher percentage compared to the two human listeners with the lowest accuracy. The Google ASR, on the other hand, correctly recognized 50% of the sentences, suggesting that half of the sentences contained transcription errors.

4.4 Correlations among ASR Accuracy, Human Intelligibility, and Comprehensibility

Pearson product-moment correlation coefficients were computed to assess if there were any meaningful relationships among the two ASR systems' accuracy rates, human listeners' transcription accuracy rates (intelligibility), and ratings of comprehensibility (Tables 4.9, 4.10). For both Google and ETRI ASR systems, there was a statistically significant correlation between ASR recognition and human listener intelligibility as well as between these measures and ratings of comprehensibility. A statistically significant correlation obtained between two

variables means that the null hypothesis that there is no relationship between them was rejected, and therefore we can say that these two variables were related. It also provides information about the magnitude of the correlation (low to high), as well as the direction of the relationship (positive or negative). The proposed guidelines for the interpretation of Pearson correlation coefficients are provided in Table 4.8 shown below.

Table 4.8 Interpretation of Pearson's *r*

Degree of correlation	Coefficient, *r*	
	Positive	Negative
High	$.50 \leq r < 1.00$	$-1.00 < r \leq -.50$
Moderate	$.30 \leq r < .50$	$-.50 < r \leq -.30$
Low	$.10 \leq r < .30$	$-.30 < r \leq -.10$

4.4.1 Google ASR's Correlation with Human Measures

Table 4.9 shows the Pearson correlation results of the Google ASR's accuracy and the two human measures of intelligibility and comprehensibility of utterances. Overall, there was a moderate positive correlation between Google recognition accuracy and human listener intelligibility, $r(1503) = .43$, $p < .001$. This means that Google recognition accuracy was also higher for utterances with high intelligibility to human listeners, and vice versa. Among the human transcribers, the Google ASR had a relatively stronger correlation with native transcribers' accuracy [$r(1503) = .42$, $p < .001$] compared to that of non-native transcribers [$r(1503) = .37$, $p < .001$]. Moreover, the Google ASR accuracy and human listener comprehensibility judgement were also found to be moderately positively correlated, $r(1503) = .45$, p

< .001. This means that Google recognition accuracy was higher for utterances with high comprehensibility ratings from human listeners, and vice versa. Among the human raters, the Google ASR was more strongly correlated with non-native raters' comprehensibility ratings [$r(1503) = .41$, $p < .001$] as compared to that of native raters [$r(1503) = .38$, $p < .001$].

Table 4.9 Correlation of Google Recognition Accuracy with Human Measures

	Human Intelligibility			Human comprehensibility		
	NS	NNS	All	NS	NNS	All
r	.42	.37	.43	.38	.41	.45
p	<.001	<.001	<.001	<.001	<.001	<.001

4.4.2 ETRI ASR's Correlation with Human Measures

The correlation results among the ETRI ASR system's accuracy and human measures of intelligibility and comprehensibility are shown in Table 4.10. There was a moderate positive correlation of ETRI recognition accuracy with human intelligibility and comprehensibility measures. A relatively stronger correlation was observed between the ETRI ASR accuracy and human intelligibility [$r(1503) = .49$, $p < .001$] compared to human comprehensibility ratings [$r(1503) = .41$, $p < .001$]. In light of recognition accuracy rates, the ETRI ASR had similar correlations with both native [$r(1503) = .45$, $p < .001$] and non-native [$r(1503) = .45$, $p < .001$] transcribers. Likewise, it also showed similar correlations with the comprehensibility ratings of native [$r(1503) = .36$, $p < .001$] and non-native [$r(1503) = .36$, $p < .001$] raters. This is different from the Google ASR results,

which showed relatively stronger correlations with native listeners' intelligibility and non-native listeners' comprehensibility.

Table 4.10 Correlation of ETRI Recognition Accuracy with Human Measures

	Human Intelligibility			Human comprehensibility		
	NS	NNS	All	NS	NNS	All
r	.45	.45	.49	.36	.36	.41
p	<.001	<.001	<.001	<.001	<.001	<.001

4.4.3 Correlation between Human Listeners' Intelligibility Score and Comprehensibility Rating

The correlation between human listeners' intelligibility (transcription accuracy) and comprehensibility rating is presented in Table 4.11. Overall, there was a strong positive correlation between the intelligibility score and comprehensibility rating of all human listeners [$r(1503) = .65$, $p < .001$]. Among the native and non-native subgroups, a relatively stronger correlation was observed for the native listeners [($r(1503) = .62$, $p < .001$], compared to the non-native listeners [$r(1503) = .51$, $p < .001$]. This means that native listeners showed a stronger relationship between their transcription accuracy of a certain utterance and their comprehensibility rating of the same utterance.

Table 4.11 Correlation between Human Listeners' Intelligibility (Int) and Comprehensibility (Comp) Rating

	Int-Comp		
	NS	NNS	All
r	.62	.51	.65
p	<.001	<.001	<.001

4.5 The Problem of Outliers

We have previously seen that among the 8 listeners who participated in the current study, there were 2 listeners who showed relatively low intelligibility and comprehensibility ratings compared to the other 6 listeners. Because these raters did not negatively affect the inter-rater reliability in terms of intelligibility or comprehensibility, we included them in the main data analysis presented above. Nevertheless, we also calculated the mean human intelligibility and comprehensibility ratings excluding these two potential outliers and carried out the same data analysis to check whether the exclusion of these outliers changes the overall results of the current study.

Table 4.12 Means and Standard Deviations of Listener Intelligibility and Comprehensibility without Outliers (Compare with Table 4.1)

	%Listener recognition (without outliers)			%Listener comprehensibility (without outliers)		
	NNS	NS	All	NNS	NS	All
M	96.02	96.22	96.12	4.50	4.77	4.63
SD	7.04	7.10	6.38	.47	.36	.36

The mean intelligibility and comprehensibility results without the outliers are shown in Table 4.12. When we excluded the two outliers, the overall transcription accuracy of human listeners was 96.12% which is 0.83% higher than the mean intelligibility of all 8 listeners (95.29%). The mean comprehensibility rating was 4.63 which is 0.24 higher than the comprehensibility of all 8 raters (4.39). When we compare the means of native and non-native listeners without the outliers, native listeners showed a higher mean intelligibility compared to non-native listeners, which is the opposite of our previous results including the outliers. Nevertheless, the difference between the native and non-native listeners' intelligibility was found to be not significant [$t(1504) = -1.301$, $p = .194$], which is in line with our previous results. In terms of comprehensibility ratings, even when we exclude the outliers, the results were the same as our previous results that native listeners gave significantly higher comprehensibility ratings compared to non-native listeners [$t(1504) = -25.374$, $p < .01$].

Table 4.13 Correlation of Google Recognition Accuracy with Human Measures without Outliers (Compare with Table 4.9)

	Human recognition			Human comprehensibility		
	NS	NNS	All	NS	NNS	All
r	.37	.34	.39	.36	.38	.43
p	<.01	<.01	<.01	<.01	<.01	<.01

Table 4.14 Correlation of ETRI Recognition Accuracy with Human Measures without Outliers (Compare with Table 4.10)

	Human recognition			Human comprehensibility		
	NS	NNS	All	NS	NNS	All
r	.42	.39	.45	.37	.34	.40
p	<.01	<.01	<.01	<.01	<.01	<.01

Table 4.15 Correlation of Individual Human Listener's Intelligibility and the Two ASR Systems' Accuracy Rates

	Native				Non-native			
	ns1	ns2	ns3	ns4	nns1	nns2	nns3	nns4
Google	0.29	**0.36**	0.32	0.26	0.30	0.27	0.27	0.24
ETRI	0.36	0.35	0.32	0.31	**0.40**	0.32	0.31	0.28

The correlation between the ASR results and the human intelligibility and comprehensibility ratings without the two outliers are shown in Table 4.13 and 4.14. Overall, the correlations between human listeners and ASR results were lower compared to the results including the two outliers. To explain this unexpected result, we investigated the correlation between the ASR accuracy rates and each human listeners' transcription accuracy rates as shown in Table 4.15. These results show that Native Listener 2 who had the lowest intelligibility among native listeners had the highest correlation with Google ASR's accuracy rates, and Non-native Listener 1 who had the lowest intelligibility among non-native listeners had the highest correlation with ETRI ASR's accuracy rates. This is notable because the two listeners who had relatively low accuracy rates compared to the other listeners, and thus considered outliers in our study, were actually more strongly correlated with ASR systems' accuracy rates.

Chapter 5. Discussion

This chapter discusses the major findings of the current study in terms of the comparison between the transcription accuracy rates of human listeners and the ASR systems, the correlation of the ASR systems' accuracy rates to the intelligibility score and comprehensibility judgment of human listeners, and the difference between the two ASR systems' recognition of non-native speech illustrated by selected examples.

5.1 Comparison of ASR Systems and Human Listeners in Transcribing Non-native Speech

5.1.1 ASR Systems vs. Human Listeners

The first research question posed for the current study has to do with whether ASR systems can achieve a similar level of transcription accuracy with human listeners when transcribing non-native English speech. In this study, two kinds of state-of-the-art ASR systems were compared between themselves and also with human listeners in terms of transcribing read speech produced by Korean learners of English. Human listeners had the highest level of transcription accuracy (95.29%), followed by the ETRI ASR (93.77%) and Google ASR (85.27%). The Google ASR had a significantly lower accuracy rate compared to human listeners when transcribing the non-native speech data. On the other hand, results showed that the ETRI ASR's accuracy closely approached that of human listeners. Although the ETRI ASR's accuracy rate was slightly lower than the overall average of human

listeners, a close inspection of individual listeners' accuracy rates revealed that the ETRI ASR was very similar to, and even outperformed, two out of eight human listeners in terms of transcription accuracy. These results suggest that the ETRI ASR which was specifically trained and optimized for recognizing Korean learners' English could be comparable to human listeners, while the Google ASR which was built to recognize English spoken by native speakers showed a significantly lower accuracy in recognizing English produced by Korean EFL learners.

This implies that EFL learners who wish to practice their pronunciation with an ASR system optimized for English native speakers are likely to receive rather discouraging and frustrating transcription results from the system. If they would like to check how intelligible their speech is to a potential human listener, using an ASR system that is specifically optimized for recognizing their non-native speech could provide transcription results that better resemble those of human listeners.

5.1.2 Native vs. Non-native Listeners

An interesting point to note in the current study is that native and non-native speakers of English were both included as listeners and performed the task of transcription and comprehensibility rating of the EFL learners' English speech data. Such a research design makes the current study unique and differentiated from previous studies that used only native speakers of English as transcribers and raters.

In this study, native and non-native listeners were found to show a statistically significant difference in comprehensibility ratings of the non-native speech, while they showed no significant difference in terms of intelligibility

scores (transcription accuracy). This is surprising, given that human listeners' intelligibility score was rather an objective measure of understanding, whereas comprehensibility rating was a subjective measure of perception in L2 pronunciation research (Thomson, 2018). The results of our study showed that native listeners were more lenient compared to non-native listeners in rating the comprehensibility of the non-native speech even though both groups of listeners had similar levels of intelligibility.

Several factors could have contributed to the lack of difference between native and non-native listeners in terms of transcription accuracy (intelligibility score). First, the native and non-native listeners who participated in this study might have been all accustomed to Korean learners' English. The non-native listeners were all native speakers of Korean with very high English proficiency, so they had a shared L1 background with the non-native speakers of English who provided speech data for the English spoken corpus analyzed in this study. When the current study was undertaken, the native listeners had had extensive exposure to Koreans' English since they had been residing in Korea for 1 to 3 years and also had experience in teaching English to Korean learners for years. Carey et al (2011) and Winke and Gass (2013) found that raters' familiarity with a certain accent ultimately could lead to higher score assignments of non-native speech. Listeners may become familiar with a speaker's accent through the means of: (a) sharing the same L1 with the speaker (Brown, 1995; Kim, 2009; Xi & Mollaun, 2009; Zhang & Elder, 2011); (b) prior experience of studying the speakers' L1 (Bent & Bradlow, 2003; Fayer & Krasinski, 1987; Winke et al., in press); or (c) extended exposure to

the speakers' L2 speech (Chalhoub-Deville, 1995). Furthermore, the third type of familiarization, extended exposure to the speakers' L2 speech, can also be gained in various ways, such as by having: (a) lived in the country where the speaker's L1 is spoken (Carey et al., 2011), (b) worked with or taught speakers of that L1 (Brown, 1995; Carey et al., 2011; Chalhoub-Deville, 1995), or (c) grown up around native speakers of that L1. Since our native raters had resided in Korea, teaching English to Korean ESL learners for some years at the time of participation, it would be reasonable to regard them as having extended exposure to the learners' L1. In other words, both native and non-native listeners in our study might have developed extensive accent familiarity with the Korean EFL learners by the time when this study was undertaken.

Moreover, Jun and Li (2010) pointed out that when rating comprehensibility, non-native raters paid more attention to specific pronunciation features, while native raters focused more on the overall impression of the speech or whether they understood the intended message. This may explain the reason for the relatively lenient ratings of the native listeners in our study.

5.1.3 Outliers

Among the human listeners who participated in the current study, one non-native listener and one native listener had a relatively poor transcription accuracy rate and understanding of the non-native speech compared to the other listeners. The non-native listener had no experience living in an English-speaking country which may account for the relatively poor transcription accuracy and understanding. This may be related to the non-native listener's English proficiency or the lack of knowledge

of certain expressions or phrases that were read aloud by the learners. The native listener who also showed a relatively low accuracy had the least exposure to Koreans' English in terms of length of residence in Korea and teaching experience, which may account for the poor level of understanding. A similarity found between the two listeners was that they made less effort to make sense of the spoken utterance compared to the other listeners. This means that the other listeners tried to make a plausible guess on the less intelligible parts of the utterance while these two listeners were more focused on directly reflecting in the transcription what they had heard.

5.2 Correlation of ASR Results and Human Measures

To answer the second research question, the relationships among ASR accuracy rates and human listeners' intelligibility score and comprehensibility judgment were examined. The results in Subsection 4.4 showed that there was a moderate positive correlation between the ASR systems' recognition accuracy and listeners' transcription accuracy rates (intelligibility score) and ratings of comprehensibility. This means that the accuracy rates of both ASR systems were moderately related to how much human listeners would actually understand the speech, and their perceived difficulty in understanding the speech. It was interesting that the ETRI ASR which modeled the characteristics of Korean learners' English speech had a higher level of correlation with human intelligibility, compared to the Google ASR. While the Google ASR had a stronger correlation with native listeners' intelligibility than that of non-native listeners, the ETRI ASR had a similar level of

correlation for both groups of listeners, showing a big increase in the degree of correlation with non-native listeners.

However, the strength of the relationship was not as strong as the results shown in McCrocklin and Edalatishams (2020). In the previous study, Google's recognition was more strongly correlated with human intelligibility ($r = .78$, $p < .001$) and ratings of comprehensibility ($r = -.71$, $p < .001$) than in the current study (Note that the correlations for comprehensibility were negative since they used an inverse scale to rate comprehensibility). However, in the current study, the correlations between Google and ETRI's recognition and human intelligibility as well as comprehensibility were all in the moderate range.

Although we cannot pinpoint the exact reason for this lower correlation between ASR accuracy and human measures in our study, we could speculate on various factors that might have caused such discrepancies, particularly in terms of the differences in research design and experimental conditions. First, the speech material that the learners read aloud were different. In the previous study by McCrocklin and Edalatishams (2020), 60 true or false sentences taken from Derwing et al. (2000) were used. An example of a true sentence was "Elephants are big animals," whereas a false sentence was "A monkey is a kind of bird." If a speaker pronounced the word "bird" less intelligibly, it would be relatively more difficult to guess the intended word for a false sentence compared to normal sentences. In comparison, in our study the speech material did not include any false sentences and it only consisted of daily expressions such as "Do you have time for

a cup of coffee."

Secondly, in the previous study, the intelligibility score was calculated by counting the number of correctly transcribed words from the reference text. So if a sentence contained 6 words and the program identified 5 correctly, the transcript was counted as 5/6 correct, or 83.33%. However, in our study we used 1 - word error rate (WER) as the metric to calculate the accuracy of the ASR systems as well as human listeners. Therefore, compared to the metric in the previous study, there was also a penalty for insertions applied in the ASR and human listeners' transcriptions. Such differences in operationalization of intelligibility might have partially contributed to the lower correlations obtained for this study.

Third, in terms of comprehensibility ratings, the previous study used a 9-point scale while the current study used a 5-point scale. In the previous study, the overall comprehensibility ratings for L1 Spanish and Chinese speakers were 2.98 and 3.51 on a 9-point scale, where a score closer to 1 meant easy to understand. In the current study, the average comprehensibility rating was 4.39 on a 5-point scale, where a score closer to 5 meant easy to understand. This shows that in both studies the utterances received high comprehensibility ratings and that there is a possibility that the variability in the comprehensibility could have been somewhat obscured or artificially reduced due to the shrunken rating scale with a smaller number of score points used in the current study.

Finally, the resulting transcription accuracy rates of human listeners and ASR systems for the non-native speech data showed different patterns in the previous and current studies. In McCrocklin and Edalatishams (2020), the overall

recognition accuracy rates of ASR and human listeners were 93.09% for the L1 Spanish speakers and 88.95% for the L1 Chinese speakers, respectively, whereas Google's recognition was 92.73% and 90.99% for the same speaker and listener groups.. The overall transcription accuracy was low, compared to our study in which the average human accuracy was 95.29%. Moreover, in the previous study, the recognition accuracy rates of human listeners and Google ASR had a small difference (0.36% for L1 Spanish, 2.04% for L1 Chinese). In terms of recognizing L1 Chinese speakers' speech, the Google ASR outperformed the human listeners by 2.04%. However, in our study, the human listeners outperformed the Google and ETRI ASRs in recognizing the Korean learners' speech with a relatively large gap. Therefore, more complex factors seem to have affected the recognition errors of the ASR systems and in turn impacted the correlations with human measures of intelligibility and comprehensibility.

5.3 Comparison of the Two ASR Systems with Example Transcriptions

The third research question has to do with the difference between the two ASR systems used in this study in terms of recognition accuracy and correlations with human measures. The ETRI ASR results achieved a significantly higher accuracy rate in transcribing the non-native speech compared to the Google ASR and showed the possibility of rivalling human listeners. This enhancement in accuracy rate also contributed to a stronger correlation with human intelligibility scores.

Such enhancement in accuracy and validity of ASR is clearly a positive development. A next step to be taken regarding such positive outcome is to delve deeper into what might have contributed to such performance enhancement of the ETRI ASR, when compared to the Google ASR. A related question is what should possibly be done to further improve the accuracy of ETRI ASR up to the level of human transcribers.

A good starting point for seeking solutions to these important challenges is to examine in an exploratory manner the actual examples of recognition/transcription output that can clearly illustrate the types of different recognition difficulties the ASR systems and human transcribers are struggling with and provide insights into the potential causes of performance differences between Google ASR, ETRI ASR, and human transcribers. Table 5.1 shows some selected examples of the transcription/recognition errors by human listeners and the two ASR systems along with the intended reference text. The first three examples show cases in which human listeners produced no transcription errors, while the two ASR systems showed lower accuracy rates. Compared to the Google ASR's errors which are spread across many words, the ETRI ASR's errors tend to be confined to certain words such as "this" as a substitution error for "teeth," "people" for "before," and "lead" for "lid." In examples 4 to 6, it is interesting how the ASR errors corresponded to the words human listeners also had difficulty recognizing (e.g., "city hospital," "the police caught me," "and awesome"). The last two example sentences show that the results of the ETRI ASR could be more forgiving of learner errors compared to some human listeners.

In particular, the ASR transcription results shown in Example 2 are very impressive since they seem to clearly show in what areas the ETRI ASR does a better job than the Google ASR. The reference text was “She went to the hospital before lunch time,” while the Google and ETRI ASR produced the transcriptions, “Sorrento hospital people run time” (Google) and “She went to the hospital people lunch time” (ETRI), respectively. Even just a quick inspection shows that the ETRI’s transcription is more accurate. Then, a key word that deserves a special attention is “lunch,” which was wrongly recognized as “run” by Google but correctly transcribed as “lunch” by ETRI. It is well-known that Korean learners of English have difficulty discriminating between /l/ and /r/ sounds in perception and production, because an English phoneme /1/ (particularly appearing at the word initial position) does not exist in the consonant inventory of the Korean language. One plausible explanation might be that, although a Korean EFL speaker mispronounced “lunch” as something similar to “runch,” the ETRI might have correctly recognized it as “lunch” since it was extensively trained based on non-native speech data produced by ESL and EFL learners, including Korean learners of English.

Another word that needs to be discussed here is the word “before,” which was incorrectly transcribed as “people” by both the Google and ETRI ASR. The second syllable (-fore) in “before” seems to have made a critical difference in this recognition failure. This clearly shows the limitation of the current versions of both ASR systems. At this moment, it is not possible to make a definitive statement regarding what might have led to such a transcription error, but there seem to be

some possible explanations. First of all, we all know that there is no /f/ sound in the sound inventory of the Korean language and thus Korean EFL learners tend to have hard time differentiating between /f/ and /p/ sounds in perception and production. In this case, an additional source of recognition difficulty may come from the fact that the word stress should be placed on the second syllable in the case of "before." Korean is a syllable-timed language rather than a stress-timed language. Therefore, it is highly likely for Korean EFL learners to assign a primary stress on the first, instead of the second, syllable or pronounce the word with no word stress at all. It was indeed found that the particular Korean EFL learner, who produced the utterance, mispronounced /f/ as /p/ or something similar to it and also pronounced the word with no word stress on either syllable instead of assigning stress on the second syllable. Nevertheless, these pronunciation variations did not affect the transcription accuracy of human listeners, since they all transcribed the utterance correctly despite the segmental and suprasegmental variations in the EFL learner's speech. Such observations point to the exact areas where the current ASR technology needs to improve on and provide insight into how that can be achieved. A more systematic, further investigation is clearly warranted along this line.

This line of research can produce results that can prove very useful and effective for second language instruction and feedback. When the root causes of recognition difficulties for EFL learners' speech and accuracy discrepancies between ASR and human transcribers are identified, such information can be used to enhance the performance of the existing ASR technology and also generate useful instructional feedback for EFL learners. These kinds of feedback can be

useful to learners when they want to know which parts of their speech is less intelligible, and what kind of pronunciation they need further practice in. More research is needed to compare the usefulness of the feedback provided by different kinds of ASR systems.

Table 5.1 Examples of Recognition Errors by the ASR Systems and Human Listeners

No.		Sentence
1	Reference	No just bite it off with your teeth
	Human	-
	Google	No just to buy to eat open this
	ETRI	No just bite it off you with your this
2	Reference	She went to the hospital before lunch time
	Human	-
	Google	Sorrento hospital people run time
	ETRI	She went to the hospital people lunch time
3	Reference	Since the lid was off it spilled everywhere
	Human	-
	Google	Sisterly advice of his spirit everywhere
	ETRI	Since the lead was off it spilled everywhere
4	Reference	It's right next to the city hospital
	Human	It’s right next to the serious pizza
		It’s right next to the siri hospital
		It’s right next to the serious hospital
	Google	It's a lie to that to the serious pizza
	ETRI	It's right next to the serious people
5	Reference	The police caught me running a red light
	Human	The boatess cut me running a red light
	Google	Police academy running or red rights
	ETRI	The bodies cut me running a red light

6	Reference	Look at the gate it’s huge and awesome
	Human	Look at the gate it’s huge compared to some
		Look at the gate it’s huge than the sun
	Google	Look at the gate it's using a son
	ETRI	Look at the gate it's use and us um
7	Reference	I would like your opinion on my new proposal
	Human	I would like your opinion on my new proper shirt
		I would like your opinion on my professor
	Google	I would like your opinion on my new pro pressure
	ETRI	-
8	Reference	Oh really I’d rather go that way
	Human	Oh really I’d let her go that way
	Google	Old lily I’d let her go that way
	ETRI	-

Chapter 6. Conclusion

6.1 Conclusion and Implications

The main objectives of the current study were to investigate the relationship among measures of ASR accuracy, human transcription accuracy, and human comprehensibility judgement of non-native read-aloud speech. Human listeners consisting of both native and non-native English speakers were asked to transcribe more than a thousand short fragments of speech recorded by Korean EFL learners of English. The two ASR systems (e.g., Google ASR for general use and the ETRI ASR optimized for recognizing non-native speech) were used to transcribe these fragments, and the comprehensibility of each of these fragments was rated by human listeners on a 5-point scale. The measures of human transcription and ASR recognition accuracy were obtained for these speech fragment data, and the correlations were computed among these accuracy variables and comprehensibility ratings. The recognition accuracy rates of two different ASR systems were compared with the transcription accuracy of human transcribers, and the correlations between ASR accuracy and the two kinds of human measures (intelligibility score and comprehensibility rating) were examined.

Results of the analyses showed several noteworthy patterns. First, the accuracy rates of the Google Web Speech API and ETRI Open API in recognizing Korean learner's English speech data were generally somewhat lower than those of human transcribers. This may suggest that there is much room for improvement and refinement for the two state-of-the-art ASR systems used in this study. One

promising finding along this line was that the ETRI ASR, which was trained on non-native speech data and modeled the characteristics of Koreans' English, achieved a significantly higher accuracy rate than that of the Google ASR, which was trained on native speaker speech data and mostly adopted for general use .

Second, the ASR accuracy and human measures of intelligibility and comprehensibility had moderate positive correlations. This suggests that the Google and ETRI ASR systems are generally using similar logics and processing mechanisms to human transcribers in recognizing Korean EFL learners' speech data, although the sources of discrepancies between the two ASR systems and between the two ASR systems and human transcribers should be further examined. Another related, notable pattern was that intelligibility and comprehensibility, the two core concepts in describing the quality of non-native speech, turned out to be correlated at a moderate level.

What are some important implications that can be derived from such results and findings? At this point, it should be noted that the ultimate goals of the current study were to: (a) not only evaluate the performance of the existing ASR technology in recognizing the non-native speech data; but also (b) explore the possibility of advancing the current ASR technology in such a way that the accuracy rates of ASR systems are significantly improved for non-native speech data and that the ASR systems are utilized for computer-assisted pronunciation training (CAPT), particularly in terms of generating useful diagnostic, instructional feedback.

A ray of hope can be gleaned from the facts that the ETRI ASR optimized for non-native speech data outperformed the Google ASR developed for general use and approached the transcription accuracy rates of human listeners. The first thing that can be mentioned here is that it would be critical to use non-native speech data as well as native data in terms of model building and training for ASR systems in order to improve the performance of the existing ASR technology. If a particular ASR is intended to be used for Korean learners of English, speech data from this particular L1 group of English must be used in the process of ASR model building and validation.

Another important point worth mentioning is that researchers' efforts to improve the performance of the existing ASR systems should not stop here. Such efforts need to be expanded to investigate how the ASR can be used for CAPT and assessment and feedback. This would eventually require qualitative analyses of the major types of recognition errors committed by ASR systems, and the speech fragments where general-use and non-native optimized ASR systems disagree in recognition.

The last thing worth mentioning is that the golden standard of ASR has been human perception and judgement involved in the recognition and transcription of speech, particularly those of native speakers. However, one should note that there could be a wide variation of perception and judgement among native speakers/listeners associated with their geographical and social backgrounds (e.g., nationality, regions of residence, gender, ethnicity, age, occupation). When it

comes down to the recognition of non-native speech data, the issue of what should be the standard becomes even more complex, particularly in relation to non-native speakers' L1. For this reason, it seems necessary to investigate the perception and transcription differences among native speakers of English, between native and non-native speakers of English, and also among particular L1 groups of English learners. Researchers probably need to find ways to model such variations across native and non-native English speakers in the ASR systems and to utilize both native pronunciation standards and non-native variations for the instructional and feedback purposes in the context of CAPT.

6.2 Limitations and Future Studies

Further research is needed on the similarities and differences of transcription errors produced by human transcribers and ASR systems, and also on the potential causes of the discrepancies between ASR and human transcription. Such more qualitatively oriented analyses of native/nonnative human and ASR errors can shed some important light on the true nature of intelligibility and comprehensibility of speech and on how to further improve the current ASR technology. Another promising line of research is to obtain recognition results from multiple ASR systems built on different recognition models and training data and compare our current results with the ones from these additional ASR systems.

References

Ashwell, T., & Elam, J. R. (2017). How accurately can the Google Web Speech API recognize and transcribe Japanese L2 English learners' oral production? *The JALT CALL Journal 2017*, 13(1), p. 59-76.

Baker, A. (2011). Discourse prosody and teachers' stated beliefs and practices. *TESOL J. 2 (3)*, 263–292. doi:10.5054/tj.2011.259955

Bendig, A. W. (1953). The reliability of self-ratings as a function of the amount of verbal anchoring and of the number of categories on the scale. *Journal of Applied Psychology*, 37, 38–41.

Bent, T., & Bradlow, A. R. (2003). The Interlanguage Speech Intelligibility Benefit. *The Journal of the Acoustical Society of America*, 114(3), 1600-1610.

Bradlow, A. R., & Pisoni, D. B. (1999). Recognition of spoken words by native and non-native listeners: Talker-, listener-, and item-related factors. *The Journal of the Acoustical Society of America 106*, 2074-2085. doi: 10.1121/1.427952

Brown, A. (1988). Functional load and the teaching of pronunciation. *TESOL Quarterly, 22(4),* 593–606. https://doi.org/10.2307/3587258

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing, 12*, 1–15. doi:10.1177/026553229501200101

Carey, M. D., Mannell, R. H., & Dunn, P. K. (2011). Does a rater's familiarity with

a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing, 28*, 201–219. doi:10.1177/0265532210393704

Chalhoub-Deville, M. (1995). A contextualized approach to describing oral language proficiency. *Language Learning, 45*, 251–281.

Chalhoub-Deville, M. (1995). Deriving oral assessment scales across different tests and rater groups. *Language Testing, 12*, 16–33. doi:10.1177/026553229501200102

Chen, N., & Li, H. (2016). Computer-assisted pronunciation training: From pronunciation scoring towards spoken language learning. *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 1–7.

Chung, H., Jeon, H., Park, G., Park, J., Kim, U., Yun, S., & Lee, Y. (2014b). Daewhaeumseong interface gisul mit eungyong service gaebal donghyang [Conversational speech interface technology and application service development trend]. *The Magazine of the IEIE, 41(3)*, 59-78.

Chung, H., Lee, S. J., & Lee, Y. K. (2014a). Weighted finite state transducer-based endpoint detection using probabilistic decision logic. *ETRI Journal, 36(5)*, 714-720. doi:10.4218/etrij.14.2214.0030

Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*. 1994;6(4):284–290.

Coniam, D. (1999). Voice recognition software accuracy with second language speakers of English. System, 27, 49-64.

Cucchiarini, C., & Strik, H. (2018). Automatic speech recognition for second language pronunciation assessment and training. In O. Kang, R. I. Thomson, & M. J. Murphy (Eds.), pp. 556-569. The Routledge handbook of English pronunciation. London: Routledge.

Cucchiarini, C., Strik, H., & Boves, L. (2000a). Different aspects of expert pronunciation quality ratings and their relation to scores produced by speech recognition algorithm. *Speech Communication, 30(2-3)*, 109-119.

Cucchiarini, C., Strik, H., & Boves, L. (2000b). Quantitative assessment of second language learners' fluency. *Journal of the Acoustical Society of America, 107(2)*, 989-999.

Cucchiarini, C., Strik, H., & Boves, L. (2002). Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. *Journal of the Acoustical Society of Amercia, 111(6),* 2862-2873.

Delattre, P. (1947). A Technique of Aural-Oral Approach. *The French Review, 20(3)*, 238-250.

Derwing, T. M., & Munro, M. J. (2005). Second Language Accent and Pronunciation Teaching: A Research-Based Approach. *TESOL Quarterly, 39(3)*, 379-397.

Derwing, T. M., Munro, M. J., & Carbonaro, M. (2000). Does popular speech recognition software work with ESL speech? *TESOL Quarterly, 34*, 592-603.

Derwing, T., & Munro, M. (1997). Accent, intelligibility, and comprehensibility. *Studies in Second Language Acquisition, 19*, 1–16.

Derwing, T.M., & Munro, M.J. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching, 42*, 476–490.

Derwing, T.M., & Munro, M.J. (2015). Pronunciation fundamentals. Evidence-based perspectives for L2 teaching and research. John Benjamins. https://doi.org/10.1075/lllt.42

Doremalen, J., Strik, H., Cucchiarini, C. (2009). Optimizing non-native speech recognition for CALL applications. Speech Communication, 592–595.

Electronics and Telecommunications Research Institute. (n.d.). Speech Recognition Technology. Retrieved June 15, 2021, from https://aiopen.etri.re.kr/guide_recognition.php

Evanini, K., Higgins, D., & Zechner, K. (2010). Using Amazon Mechanical Turk for transcription of non-native speech. *Proceedings of the NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, Association for Computational Linguistics, 53–56.

Fayer, J. M., & Krasinski, E. (1987). Native and nonnative judgments of intelligibility and irritation. *Language Learning, 37*, 313–326.

doi:10.1111/j.1467-1770.1987.tb00573.x

Georgila, K., Leuski, A., Yanov, V., & Traum, D. (2020). Evaluation of off-the-shelf speech recognizers across diverse dialogue domains. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 6469–6476.

Guskaroska, A. (2019). ASR as a tool for providing feedback for vowel pronunciation practice. Graduate Theses and Dissertations. 17020. https://lib.dr.iastate.edu/etd/17020

Hu, W., Qian, Y., Soong, F. K., & Wang, Y. (2015). Improved mispronunciation detection with deep neural network trained acoustic models and tranfser learning based logistic regression classifiers. *Speech Communication, 67*, 154-166.

IBM Corp. Released 2019. IBM SPSS Statistics for Windows, Version 26.0. Armonk, NY: IBM Corp

Isaacs, T. & Thomson, R. I. (2013). Rater Experience, Rating Scale Length, and Judgments of L2 Pronunciation: Revisiting Research Conventions, *Language Assessment Quarterly, 10:2*, 135-159, DOI: 10.1080/15434303.2013.769545

Isaacs, T., & Trofimovich, P. (2012). Deconstructing comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings. *Studies in Second Language Acquisition, 34*, 475–505.

Isaacs, T., Trofimovich, P., & Foote, J. (2017). Developing a user-oriented second language comprehensibility scale for English-medium universities. *Language Testing*.

Jun, H. G. & Li, J. (2010). Factors in raters' perceptions of comprehensibility and accentedness. In J. Levis & K. LeVelle (Eds.), *Proceedings of the 1st Pronunciation in Second Language Learning and Teaching Conference*, Iowa State University, Sept. 2009. (pp. 53-66), Ames, IA: Iowa State University.

Kennedy, S., & Trofimovich, P. (2008). Intelligibility, comprehensibility, and accentedness of L2 speech: The role of listener experience and semantic context. *The Canadian Modern Language Review, 64(3)*, 459-489.

Kim, Y. H. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing, 26*, 187–217. doi:10.1177/0265532208101010

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. Journal of chiropractic medicine, 15(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Kwon, O. W., Lee, K., Roh, Y.-H., Huang, J.-X., Choi, S.-K., Kim, Y.-K., Jeon, H. B., Oh, Y. R., Lee, Y.-L., Kang, B. O., Chung, E., Park, J. G., & Lee, Y. (2015). GenieTutor: a computer assisted second-language learning system based on spoken language understanding. *IWSDS, 2015*.

Lado, R. (1964). Language teaching: A scientific approach. London: MacGraw-Hill.

Lee, S. J., Kang, B. O., Chung, H., & Lee, Y. (2014). Intra- and inter-frame features for automatic speech recognition. *ETRI Journal, 36(3)*, 514-517. doi:10.4218/etrij.14.0213.0181

Levis, J. M. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly, 39(3)*, 369-377.

Levis, J. M. (2006). Pronunciation and the assessment of spoken language. In R. Hughes (Ed.), Spoken English, TESOL and applied linguistics: Challenges for theory and practice (pp. 245–270). New York: Palgrave Macmillan.

Levis, J. M. (2020). Revisiting the intelligibility and nativeness principles. J. Sec. Lang. *Pronunciation 6 (3)*, 310–328. doi:10.1075/jslp.20050.lev

Lison, P., & Tiedemann, J. (2016). OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*

Littlewood, W. (1981). Communicative Language Teaching. New York: Cambridge: Cambridge University Press.

Matell, M. S., & Jacoby, J. (1971). Is there an optimal number of alternatives for Likert scale items? *Educational and Psychological Measurement, 31*, 657–674.

McCrocklin, S. (2016). Pronunciation learner autonomy: The potential of Automatic Speech Recognition. *System, 57*, 25–42

McCrocklin, S. (2019). ASR-based dictation practice for second language pronunciation improvement. *Journal of Second Language Pronunciation, 5(1)*, 98-118.

McCrocklin, S., Edalatishams, I. (2020). Revisiting Popular Speech Recognition Software for ESL Speech. *TESOL Quarterly, 54(4)*, 1086-1097.

McCrocklin, S., Humaidan, A., & Edalatishams, E. (2019). ASR dictation program accuracy: Have current programs improved? In J. Levis, C. Nagle, & E. Todey (Eds.), *Proceedings of the 10th Pronunciation in Second Language Learning and Teaching Conference*, ISSN 2380-9566, Ames, IA, September 2018 (pp. 191-200). Ames, IA: Iowa State University.

McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. *Psychological Methods*. 1996;1(1):30–46.

McKelvie, S. J. (1978). Graphic rating scales: How many categories? British *Journal of Psychology, 69*, 185–202.

Miller, G. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*, 81–97.

Mroz, A. (2018). Seeing how people hear you: French learners experiencing intelligibility through automatic speech recognition. *Foreign Language*

Annals.1-21. https://doi.org/10.1111/flan.12348

Mulholland, M., Lopez, M., Evanini, K., Loukina, A., & Qian, Y. (2016). A comparison of ASR and human errors for transcription of non-native spontaneous speech. *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, 5855–5859.

Munro, M. J., & Derwing, T. M. (1995a). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning, 45(1)*, 73-97.

Munro, M. J., & Derwing, T. M. (1995b). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and Speech, 38(3)*, 289-306.

Munro, M.J., & Derwing, T.M. (2006). The functional load principle in ESL pronunciation instruction: An exploratory study. *System, 34*, 520–531.

Neumeyer, L., Franco, H., Digalakis, V., & Weintraub, M. (2000). Automatic scoring of pronunciation quality. *Speech Communication, 30(2)*, 83-93.

O'brien, M. G., Derwing, T. M., Cucchiarini, C., Hardison, D. M., Mixdorff, H., Thomson, R. I., Strik, H., Levis, J. M., Munro, M. J., Foote, J. A., & Levis, G. M. (2018). Directions for the future of technology in pronunciation research and teaching. *Journal of Second Language Pronunciation, 4(2)*, 182-207.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N.,

Hannemann, M., Motlícek, P., Qian, Y., Schwarz, P., Silovský, J., Stemmer, G., & Veselý, K. (2011). The Kaldi Speech Recognition Toolkit.

Qian, X., Meng, H., Soong, F. (2012). The use of DBN-HMMs for mispronunciation detection and diagnosis in L2 English to support computer-aided pronunciation training. *Proceedings of Interstpeech 2012* (pp. 775-778), Portland, OR.

Raymond, W. D., Pitt, M., Johnson, K., Hume, E., Makashay, M., Dautricourt, R., & Hilts, C. (2002). An analysis of transcription consistency in spontaneous speech from the Buckeye corpus. *Proceedings of the 7th International Conference on Spoken Language Processing ICSLP'02*, 1125–1128.

Richards, J. C. & Rodgers, T. S. (2001). Approaches and methods in language teaching (2nd ed.). Cambridge: Cambridge University Press.

Saito, K., Trofimovich, P., & Isaacs, T. (2017). Using Listener Judgments to Investigate Linguistic Influences on L2 Comprehensibility and Accentedness: A Validation and Generalization Study, *Applied Linguistics, 38(4),* 439–462.

Strik, H., Truong, K., de Wet, F., & Cucchiarini, C. (2009). Comparing different approaches for automatic pronunciation error detection. *Speech Communication, 51(10),* 845-852.

Suzukida, Y., & Saito, K. (2019). Which segmental features matter for successful L2 comprehensibility? Revisiting and generalizing the pedagogical value of the functional load principle. *Language Teaching Research*, 1-20.

Thomson, R. I. (2012). Demystifying pronunciation research to inform practice. In McGarrell, H. M. & Courchêne, R. (Eds.) *Special Research Symposium Issue of CONTACT, 38(2)*, 63-75.

Thomson, R. I. (2018). Measurement of accentedness, intelligibility and comprehensibility. In O. Kang & A. Ginther (Eds.), *Assessment in second language pronunciation* (pp. 11–29). London: Routledge.

Ulasik, K., Hürlimann, M., Germann, F., Gedik, E., Benites, F., & Cieliebak, M. (2020). CEASR: A corpus for evaluating automatic speech recognition. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 6477–6485.

Vaessen, N. (2019). Word error rate for automatic speech recognition. https://pypi.org/project/jiwer/.

Van Doremalen, J., Cucchiarini, C., & Strik, H. (2013). Automatic pronunciation error detection in non-native speech: the case of vowel errors in Dutch. *Journal of the Acoustical Society of America, 134*, 1336-1347.

Wallace, L. (2016). Using Google web speech as a springboard for identifying personal pronunciation problems. In J. Levis, H. Le, I. Lucie, E. Simpson, & S. Vo (Eds). *Proceeding of the 7th Pronunciation in Second Language Learning and Teaching Conference*, ISSN 2380-9566, Dallas, TX, October 2015 (pp. 180-186). Ames, IA: Iowa State University.

Winke, P. and Gass, S. (2013), The Influence of Second Language Experience and

Accent Familiarity on Oral Proficiency Rating: A Qualitative Investigation. TESOL Q, 47: 762-789. https://doi.org/10.1002/tesq.73

Winke, P., Gass, S., & Myford, C. (in press). Raters' L2 background as a potential source of bias in rating oral language. *Language Testing*.

Witt, S., & Young, S. (2000). Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication, 30(2/3)*: 95-108.

Xi, X., & Mollaun, P. (2009). How do raters from India perform in scoring the TOEFL iBT speaking section and what kind of training helps? (TOEFL iBT Research Report RR-09-31). Princeton, NJ: ETS.

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., Yu, D., & Zweig, G. (2017). Achieving human parity in conversational speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(12)*, 2410–2423.

Zechner, K. (2009). What did they actually say? Agreement and disagreement among transcribers of non-native spontaneous speech responses in an English proficiency test. *Proceedings of Speech and Language Technology in Education (SLaTE)*, 3–6.

Zhang, A. (2017). Speech Recognition (Version 3.8) [Software]. Available from https://github.com/Uberi/speech_recognition#readme.

Zhang, B., & Elder, C. (2011). Judgments of oral proficiency by non-native and

native English speaking teacher ratings: Competing or complementary constructs? *Language Testing, 28*, 31–50.

국문 초록

영어 비원어민 발화에 대한 음성 인식기의 전사 정확도와 인간 청자의 전사 정확도 및 이해가능도 평가 간의 연관성 연구

본 연구는 컴퓨터 기반 발음 훈련 (CAPT)에 활용하기 위해 비원어민 발화에 대한 자동음성인식기 정확도, 인간 전사 정확도, 그리고 인간 이해가능도 점수 간의 관계를 조사했다. 원어민 및 비원어민 청자는 비원어민 낭독체 발화 문장 1,505개를 전사하고 각 문장의 이해가능도 (comprehensibility)를 5점 척도로 평가하였다. 본 연구에서는 두 개의 서로 다른 자동음성인식기의 인식 정확도를 비교했는데, 그 중 하나는 일반적인 원어민 발화를 인식하기 위한 시스템이고, 다른 하나는 비원어민 음성 인식에 최적화된 시스템이다. 이 두 음성인식기의 정확도를 인간 청자의 전사 정확도와 비교하였으며, 아울러 이들 음성인식기의 인식 정확도와 인간 전사자 (transcribers)의 명료도 (intelligibility)와 이해가능도 (comprehensibility) 점수 간의 상관관계를 조사하였다. 두 자동 음성인식기는 모두 비원어민 발화를 인식하는 데에 있어서 인간 청자에 비해 낮은 정확도를 보였지만, 비원어민 발화 특성을 모델링한 음성인식기의 경우에는 인간 청자의 정확도에 근접한 정확도를 보였다. 자동음성인식기의 음성인식 정확도와 인간 인식 정확도 (명료도) 및 이해가능도 점수 사이에 중간 수준의 상관관계를 확인할 수 있었다. 본 연구에 사용된 두 음성인식 기 중, 비원어민 발화를 모델링한 음성인식기가 인간의 명료도 점수와 더 높은 상관관계를 보였다. 이러한 결과는 비원어민 발화에 최적화된 자등음성인식기를 활용할 때 제2외국어 학습자에게 유용한 발음 피드백을 제공할 수 있을 것이라는 가능성을 시사한다.

주요어: 컴퓨터 보조 발음 훈련, 자동 음성 인식, 비원어민 음성 인식, 한국 영어 학습자, 이해가능도, 명료도

학번: 2017-26486