12-2014

Designing Acoustics for Linguistically Diverse Classrooms: Effects of Background Noise, Reverberation and Talker Foreign Accent on Speech Comprehension by Native and Non-native English-speaking Listeners

Z. Ellen Peng
Durham School of Architectural Engineering and Construction, University of Nebraska, zpeng@huskers.unl.edu

DESIGNING ACOUSTICS FOR LINGUISTICALLY DIVERSE CLASSROOMS: EFFECTS OF BACKGROUND NOISE, REVERBERATION AND TALKER FOREIGN ACCENT ON SPEECH COMPREHENSION BY NATIVE AND NON-NATIVE ENGLISH-SPEAKING LISTENERS

by

Zhao Ellen Peng

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Doctor of Philosophy

Major: Architectural Engineering

Under the Supervision of Professor Lily M. Wang

Lincoln, Nebraska

December, 2014

DESIGNING ACOUSTICS FOR LINGUISTICALLY DIVERSE CLASSROOMS: EFFECTS OF BACKGROUND NOISE, REVERBERATION AND TALKER FOREIGN ACCENT ON SPEECH COMPREHENSION BY NATIVE AND NON-NATIVE ENGLISH-SPEAKING LISTENERS

Zhao Ellen Peng, Ph.D.

University of Nebraska, 2014

Advisor: Lily M. Wang

The current classroom acoustics standard (ANSI S12.60-2010) recommends core learning spaces not to exceed background noise level (BNL) of 35 dBA and reverberation time (RT) of 0.6 second, based on speech intelligibility performance mainly by the native English-speaking population. Existing literature has not correlated these recommended values well with student learning outcomes. With a growing population of non-native English speakers in American classrooms, the special needs for perceiving degraded speech among non-native listeners, either due to realistic room acoustics or talker foreign accent, have not been addressed in the current standard. This research seeks to investigate the effects of BNL and RT on the comprehension of English speech from native English and native Mandarin Chinese talkers as perceived by native and non-native English listeners, and to provide acoustic design guidelines to supplement the existing standard.

This dissertation presents two studies on the effects of RT and BNL on more realistic classroom learning experiences. How do native and non-native English-speaking listeners perform on speech comprehension tasks under adverse acoustic conditions, if the

English speech is produced by talkers of native English (Study 1) versus native Mandarin Chinese (Study 2)? Speech comprehension materials were played back in a listening chamber to individual listeners: native and non-native English-speaking in Study 1; native English, native Mandarin Chinese, and other non-native English-speaking in Study 2. Each listener was screened for baseline English proficiency level, and completed dual tasks simultaneously involving speech comprehension and adaptive dot-tracing under 15 acoustic conditions, comprised of three BNL conditions (RC-30, 40, and 50) and five RT scenarios (0.4 to 1.2 seconds).

The results show that BNL and RT negatively affect both objective performance and subjective perception of speech comprehension, more severely for non-native listeners than for native listeners. While the presence of foreign accent is generally detrimental, an interlanguage benefit was identified on both speech comprehension and the self-report frustration and perceived performance ratings, specifically for non-native listeners with matched foreign accent as the talker. Suggested design guidelines for BNL and RT are identified for attaining optimal speech comprehension performance to improve classroom acoustics for the non-native English-speaking population.

Copyright

Acknowledgements

This dissertation spanned a majority period of my doctoral training, which has been an extremely rewarding experience. I would like to take this opportunity to express my gratitude to those who have helped me through this journey.

First and foremost, I would like to thank my advisor and mentor, Dr. Lily M. Wang, for her guidance, support and encouragement. I am forever in debt to Lily for my growth into a young professional in the world of acoustics in the past four years. She is a role model who I will forever look up to.

I am very grateful to have an extremely supportive dissertation committee. Thank you *Dr. Thomas Carrell* for spending much time helping me create the dot-tracing task used in this dissertation. Thank you *Dr. Kanae Nishi* for your inspiration and insightful feedbacks on working with non-native English speakers, an area that I feel deeply connected. Thank you *Dr. Josephine Lau* for always sharing your experiences as an early career researcher. Thank you *Dr. Carey Ryan* for providing constructive feedbacks on the statistics in this work.

I greatly appreciate all of the comments from Dr. Ann Bradlow at Northwestern University and her sharing with me the experiences in working with non-native speakers. I would also like to acknowledge Dr. Siu-Kit "Eddie" Lau previously at UNL for his help and guidance on signal processing during the early development of this dissertation.

This project was made possible financially by two external research grants, both from the Paul S. Veneklasen Research Foundation for hiring participants to undergo over 1000 hours of subjective testing. In addition, I would like to thank *Dr. Kenneth Roy, Sean Browne* and anonymous volunteers at Armstrong World Industries, who had provided

support on hardware equipment and the anechoic recordings of speech materials used in this dissertation. I am particularly appreciative of them hosting me during a two-week measurement trip to visit their lab in Lancaster, PA.

Special thanks to the undergraduate research assistants at the University of Nebraska who assisted in this project, including *Brenna Boyd, Kristin Hanna, Adam Steinbach, Laura Brill, and Mary Kleinsasser*. The many hours they spent on proctoring experiments and compiling and analyzing data were essential to this project. And, I also appreciate them giving me a chance to become a peer mentor.

Lastly, I would not have accomplished this dissertation without my extraordinary support system of friends and family. I want to thank my friends in Omaha, NE who shared this journey with me with much joy and laughter: *Hyun Hong, Carl Hart, Ph.D., Matthew Blevins, Joonhee Lee, Jennifer Francis, Chunxiao Su, Xingbin Lin, Shihan Deng, Yinnong Jia, He Zhu,* and *Yifan Shi*. I am deeply influenced and grateful to my parents: Thank you *Mom* and *Dad*, for teaching me passion in science and compassion in humanity, both are the foundation of my research pursuit. Finally, to my wonderful husband *Won Seok Jang*: Thank you – not only for lending your programming expertise to this dissertation, but also for always having faith in me and cheering for me during good and bad times. I could not have done this alone without you!

Grant Information

Durham School of Architectural Engineering and Construction Research Seed Grant

The Paul S. Veneklasen Research Foundation Grant

Table of Contents

List of Figures

List of Tables

Chapter 1 – Introduction

1.1 Background and Motivation

The movement to improve acoustics in classrooms commenced in the 1990's, based on research studies that addressed issues in speech intelligibility performance under adverse acoustic conditions. In 2002, the interdisciplinary collaboration between architectural acoustics and hearing sciences led to the establishment of the ANSI S12.60 American National Standard: Acoustical Performance Criteria, Design Requirements, and Guidelines for Schools (hereafter referred to as the classroom acoustics standard). In the past decade, the performance-driven standard and directives with similar guidelines have been formally adopted by at least 22 entities within the U.S., including local school districts, the Departments of Education in several states, and regional and national building design initiatives (United States Access Board, 2014). Serving as design guidelines for building constructions and major renovations, these notable design initiatives included the Leadership in Energy and Environmental Design (LEED), the High Performance Incentive Program (HPI), the California Collaborative for High-Performance Schools (CHPS), and the Northeast Collaborative for High-Performing Schools (NE-CHPS).

The classroom acoustics standard has primarily remained as a voluntary practice in building design for classrooms. Most recently, the United States Access Board began the legislative process to incorporate the classroom acoustic standard (2010 revision) into the Americans with Disabilities Act (ADA), making the standard compliance mandatory for all buildings funded by the Federal government under the Architectural Barriers Act (ABA).

An excerpt from the U.S. Access Board webpage on classroom acoustics is included below:

> *The Board is undertaking rulemaking to supplement the ADA and ABA Accessibility Guidelines to address acoustics in classrooms… Once these guidelines [ANSI S12.60-2010] are adopted by the Department of Justice, they will become enforceable standards under the ADA.* [Last accessed October, 2014]

In design practice, the classroom acoustics standard provides specific guidelines on maximum background noise level (due to mechanical equipment) of 35 dBA and maximum reverberation time of 0.6 and 0.7 second, depending on the room volume. In comparison to reverberation time, the background noise level requirement was more difficult to satisfy practically due to the capacity of the mechanical equipment and financial budget. This issue is in fact reflected in the frequent revisions on the extra incentives to meet 35 dBA background noise level in the design initiatives (e.g., LEED and HPI).

Research continued to grow in furthering the improvement of classroom acoustics after the ANSI S12.60 establishment. Recent studies using *in situ* data confirmed the negative correlation between background noise level and student academic achievement (Shield and Dockrell, 2008; Ronsse and Wang, 2010). However, from the existing literature reviewed for this dissertation, findings for speech perception performance under excessive reverberation have not been able to provide strong support for the standard guidelines (Bradley *et al.*, 1999; Hodgson and Nosal, 2002; Kennedy *et al.*, 2006;

Bradley, 2011). There has not been sufficient evidence to show a strong link between the compliance of classroom acoustics standard and good learning outcomes.

To further complicate the issue, studies conducted by Klatte *et al.* (2010a) and Valente *et al.* (2012) show that both noise and reverberation are more detrimental for speech comprehension tasks than for speech intelligibility tasks, which are strictly recall tasks and predominantly used in the studies cited by the classroom acoustics standard. The trajectory of these research findings call for a re-examination of the acoustic metrics to provide more solid support on the original goal of performance-driven design, specifically by using a performance measure related to learning outcomes.

The current research, therefore, seeks to determine the design thresholds for background noise level and reverberation time to attain optimal speech comprehension performance. By using the same methodology in experimental design, the effects of background noise level and reverberation time on speech comprehension performance by native and non-native English-speaking listeners are investigated in two studies. The same set of speech comprehension materials were produced by native American English talkers in Study 1 and by native Mandarin Chinese talkers in Study 2. Based on the results of these studies, the recommended design thresholds for background noise level and reverberation time provide supplementary design considerations to the existing classroom acoustics standard, depending on the linguistic background of the talkers and listeners among the classroom occupants.

1.2 Dissertation Outline

The following chapters in this dissertation are arranged as follows. A review of existing literature pertinent to this dissertation is included in Chapter 2. It covers three main topics: 1) the effects of background noise and reverberation, 2) performance measures of speech intelligibility and speech comprehension, and 3) special needs of the acoustic environment in speech perception of non-native English speakers both as talkers and as listeners. The methodology is described in detail in Chapter 3, including the testing facilities and equipment set, the generation of test materials and acoustic conditions, and the testing procedures used in both studies. The procedures of data processing and the statistical techniques used in data analysis for this dissertation are discussed in Chapter 4. The results of analyses are explained in Chapters 5, 6 and 7 for Study 1, Study 2 and the combined study. Finally, conclusions and discussions of the findings, as well as suggestions for future work, are presented in Chapter 8.

1.3 Research Questions

The following research questions are proposed in this dissertation for the investigation of effects of background noise level (BNL), reverberation time (RT) and talker foreign accent on speech comprehension by native and non-native English-speaking listeners. They are outlined, with the hypothesis based on literature review, under the pertinent chapters.

1.3.1 *Research questions in Study 1 (Chapter 5)*

1. What are the effects of BNL and RT, while controlling for English proficiency level? At what is significant performance deficit observed in speech comprehension?

   Hypothesis: Both BNL and RT negatively affect speech comprehension performance. In particular, listeners perform best at the lowest levels of BNL (RC-30) and RT (0.4 second) in comparison with any higher levels in the respective metrics.

2. How do the effects of BNL and RT vary between native and non-native listener groups?

   Hypothesis: The effect sizes of BNL and RT suggest different strength of the acoustic metrics in the native than in the non-native listener group.

3. Do the subjective perception of task workload by listeners support the design thresholds identified from the speech comprehension measure?

   Hypothesis: The trends of BNL and RT on the subjective perception of task performance should be similar to those observed from speech

comprehension performance. The actual level of subjective perception degradation depends on statistical analysis.

1.3.2 *Research questions in Study 2 (Chapter 6)*

4. Do non-native listeners receive the interlanguage benefit of matched accent on speech comprehension?

   Hypothesis: Yes. Non-native listeners who share the same foreign accent with the talkers (i.e., native Mandarin Chinese talker to native Mandarin Chinese listeners) should see a greater improvement on comprehension performance than their non-native counterparts who do not share the accent (i.e., native Mandarin Chinese talker to other non-native English-speaking listeners).

5. Do the effects of BNL and RT on the comprehension of Chinese-accented speech replicate those from native English speech in Study 1? At what level is significant performance deficit observed in speech comprehension?

   Hypothesis: The main effects (trends) of BNL and RT are similar to findings of research question 1, although the level of significant performance deficit may differ.

1.3.3 *Research questions in the combined study (Chapter 7)*

6. How does talker foreign accent affect different listener groups under the assorted BNL and RT conditions?

Hypothesis: Listeners are expected to perform worse in comprehending speech with Chinese accent under assorted acoustic conditions. The severity may depend on the levels in the acoustic metrics. If the interlanguage benefit of matched accent is found, the performance deficit may be less severe for the listeners who share the same accent as the talkers. In addition, the negative effect of BNL and RT may also be less detrimental for these matched-accent talkers.

7. What are the design thresholds for BNL and RT in the comprehensive sample, including listeners from both studies, considering non-native English speakers among both talkers and listeners?

   Hypothesis: The levels of significant performance deficit are lower or equal to those identified in research question 1.

Chapter 2 - Previous Research

2.1 Introduction

Clear communication is the key to successful learning in traditional lecture-style classroom settings. Although teaching style and instruction techniques may be more influential on overall learning outcomes, the room acoustic environment can still impede or enhance the learning experience. A review of existing literature has been performed on the three major topics that are core to this dissertation work: 1) effects of room acoustics on speech perception, 2) performance measures of speech perception, and 3) the non-native English-speaking population. The following sections summarize and discuss the findings from previous research studies on these three topics.

2.2 Classroom Acoustics

The role of classroom acoustics on student learning outcomes has been the interest of investigation since the 1970s. An early set of studies conducted in Manhattan, New York correlated lower standardized reading scores with higher background noise level in classrooms due to road traffic noise among elementary school students (Cohen *et al.*, 1973; Bronzaft and McCarthy, 1975; Bronzaft, 1981). Two decades later, the RANCH project (Road traffic noise and Aircraft Noise exposure and children's Cognition and Health) conducted in several European countries performed an even more elaborate longitudinal investigation on children's cognition and health, which included reading comprehension performance as a learning outcome, under the long term exposure of transportation noise in classrooms (Clark *et al.*, 2006). It was found that higher background noise due to aircraft traffic was associated with lower standardized reading

comprehension scores, while controlling for confounders such as demographics, socioeconomic status and mother's education level. It was further suggested that standardized reading scores dropped below average if aircraft noise present in classrooms exceeded 55 dBA.

While quietness is recommended in classrooms, good acoustical design is equivalently advocated to ensure optimal speech delivery to the listeners. Bradley and colleagues studied a broad range of objective metrics as predictors of speech intelligibility performance (Bradley, 1986; Bradley *et al.*, 1999; Bradley *et al.*, 2003). They showed that A-weighted signal-to-noise ratio (SNR) at the listener's position was positively related to subjective speech intelligibility as perceived by listeners. Adults with normal hearing scored 80% correct on speech intelligibility tests with SNR at 0 dBA and plateaued at nearly 100% correct with SNR at +15 dBA (Figure 5, (Bradley, 1986)). They also showed that reverberation, though contributing to slightly increased background noise level, provided useful sound energy from early reflections within the first 50 milliseconds to improve speech intelligibility (Bradley *et al.*, 1999; Yang and Bradley, 2009).

Good room acoustics is even more critical in speech perception for younger children and listeners with special needs (i.e., hearing impairment and non-native English speakers). Bradlow *et al.* (2003) compared speech intelligibility under two adverse SNR conditions for children with and without learning disabilities. By reducing SNR from -4 dB to -8 dB, both groups of children experienced a significant drop in speech intelligibility performance, as much as nearly 40% for those with learning disabilities. Iglehart (2009) suggested that children with cochlear implants require an even higher

SNR of +21 dB to achieve acceptable speech intelligibility scores. It has also been found that non-native English speakers perform more poorly than native English speakers in perceiving speech in noise and reverberation, even when these non-native listeners became English dominant as early as during preschool years (Nelson *et al.*, 2005; Rogers *et al.*, 2006).

2.2.1 *Effect of Background Noise*

Background noise in classrooms can be grouped into two general categories of babble and non-babble noises. Babble noise is often found in open-plan classrooms or activities involving collaborations among students in enclosed classrooms. Shield *et al.* (2010) performed a meta-analysis on open-plan classroom studies of the past 40 years and concluded that intrusive noises, particularly unwanted speech from adjacent classrooms, were the major source of distraction and annoyance during classroom learning sessions. The lack of effective sound barriers (i.e., walls, full height partitions, and closed doors and windows) in the architectural designs of open-plan classrooms often impedes noise control treatments. In recent years, enclosed classrooms with careful noise control considerations are the preferred architectural designs recommended in design guidelines.

While babble noise is difficult to predict and quantify, non-babble or environmental noise is much more predominant in enclosed classrooms, particularly when using the conventional lecture-style teaching mode. Excessive transportation noise from road and air traffic has been found to pose challenges to children's cognitive development and academic achievement (Evans *et al.*, 1998; Haines *et al.*, 2001; Hygge

et al., 2002; Hygge *et al.*, 2003; Hygge and Kjellberg, 2010; Matheson *et al.*, 2010). Mechanical equipment of the heating, ventilating and air-conditioning (HVAC) system is another major source of non-babble background noise that negatively affects students' academic achievement (Nelson and Soli, 2000; Knecht *et al.*, 2002; Nelson *et al.*, 2005). The American National Standard Institute (ANSI) standard S12.60 for classroom acoustics recommends that the background noise level not exceed 35 dBA in unoccupied core learning spaces. However, several studies with *in situ* measurement results have indicated that most existing classrooms do exceed the standard recommendation (Knecht *et al.*, 2002; Bradley *et al.*, 2003; Ronsse and Wang, 2013). Although lower background noise level is preferred, classrooms are not likely to be retrofitted merely to meet such a standard unless they undergo major renovations and the local school district specifies the standard as part of the construction requirement. It is therefore anticipated that the majority of existing classrooms still maintain a background noise level much higher than the recommended 35 dBA in the unoccupied mode.

Ronsse and Wang (2010) studied the relation between classroom background noise level and student academic achievement from data collected in 58 grade school classrooms in Nebraska over one academic year. Results suggested that background noise level due to HVAC equipment measured in the unoccupied mode negatively correlated with standardized reading comprehension scores. They showed that, with 1 dBA increase in the unoccupied background noise level, the standardized reading comprehension score was expected to decrease by approximately 1.6% for both 2^{nd} and 4^{th} grade students. In another field study in the UK, Shield and Dockrell (2008) showed that environmental noise had a negative impact on the academic performance and attainment among primary

school children. Significant effects were found for environmental noise generated both internal and external to the classroom. Internal noises were identified as those due to mechanical equipment operation and student activities (i.e., chair scratching, paper tearing, light babbling and coughing); external noises were mostly due to road and air traffic. However, their results were countered by Xie *et al.* (2011) who did not find such significant relationships.

2.2.2 *Effect of Reverberation*

While excessive background noise level is unanimously regarded as an impairment to speech perception, there is less agreement on the role of reverberation time particularly in the lower range of less than 1 second. Reverberation time (RT) is the time for sound energy to decay 60 dB. The ease of its calculation and prediction from room geometry has made it one of the most popular metrics used in architectural acoustical designs. The ANSI S12.60 standard provides guidelines on designing reverberation time in core learning spaces depending on the enclosed room volume. It is recommended that the reverberation time should not exceed 0.6 second for typical classrooms of 283 m^3 or smaller and 0.7 second for larger classrooms up to 586 m^3.

A follow-up survey by Knecht *et al.* (2002) after ANSI S12.60 was first published in 2002 showed that over half of the 32 classrooms measured exceeded the RT design recommendation. The ideal reverberation time, as recommended in ANSI S12.60, did not seem to be always honored by existing classrooms. Hodgson and Nosal (2002) calculated the optimal reverberation times to be less than 0.3 second in order to achieve SNR above +20 dB for classrooms between 300 and 500 m^3. In contrast, Bradley and colleagues

(1999; 2003; 2008; 2009) conducted a series of experiments to argue that early reflections are critical in reinforcing and supporting the direct arrival sound, providing useful sound energy for listeners to resolve auditory information. It was further shown that speech intelligibility performances were at maximum for both adults and children of different ages when reverberation time was at approximately 0.6 second (Figure 12, (Yang and Bradley, 2009)). With performances at 0.3 and 0.9 second only slightly lower, they recommended an optimal range of reverberation time between 0.3 and 0.9 second.

However, there is not enough research to further support the optimal range of reverberation time identified by the Bradley group. In addition to background noise level, Ronsse and Wang (2013) also investigated the relation between student academic achievement and reverberation time. Unfortunately, the *in situ* measured reverberation times fell within a narrow range of values (0.4 to 0.6 second) and well below the ANSI S12.60 recommended 0.7 second. The performance scores hence suffered from range restriction and did not vary sufficiently to draw meaningful conclusions.

Several recent studies have specifically investigated the effect of reverberation on speech perception in laboratory controlled environments. Ljung and Kjellberg (2009) studied word and sentence recalls with 32 native Swedish-speaking adults under two reverberation time conditions (0.5 vs. 1.2 seconds). It was found that participants experienced more errors and reported investing more efforts during the recall tasks under the longer reverberation time. In Germany, Klatte *et al.* (2010b) digitally simulated two virtual rooms with mean reverberation times of 0.5 versus 1.1 seconds. For both adults and children from 1st and 3rd grades, the decrement of speech perception performance using word recall tasks was significantly greater for the longer reverberation time

condition. The main effect of reverberation has a large effect size with an η_p^2 of 0.36. In the U.S., a study by Valente *et al.* (2012) provided further supporting evidence on keeping reverberation time below 1 second. They also digitally simulated two reverberation time conditions of 0.6 versus 1.5 seconds and tested both adults and children of 8 and 11 years old. The main effect for reverberation time on sentence recognition tasks was again found to be significant and with a comparable effect size denoted in Pearson's r of 0.53 (equivalent to $\eta_p^2 = 0.31$). Furthermore, Wróblewski *et al.* (2012) demonstrated that adults performed even worse under a reverberation time of 0.4 when the SNR reduced from -5 dB to -10 dB, when a long-term averaged speech spectrum was utilized as the noise source.

Although reverberation adds to the negative effect of background noise when it is embedded in the target auditory stream (i.e., speech) as demonstrated by the previous studies cited above, it may help alleviate such negative effect when it is mixed with the irrelevant auditory stream (i.e., non-babble noise). Beaman and Holt (2007) studied the cognitive process by comparing performances of memory tasks in digitally simulated reverberations for three conditions (quiet, low and high). Although without precise descriptions of the reverberant conditions (i.e., reverberation time), Beaman and Holt suggested that the low and high reverberation conditions emulated those of "large lecture hall or opera theatre." It was found that higher reverberation embedded in the steady-state noise improved serial recall task performance, for which the stimuli were presented visually. Perhaps it was most valuable in this paper that Beaman and Holt pointed out research by Perham *et al.* (2007), which denoted the small effect size of reverberation. Beaman and Holt claimed that, in order to provide significant statistical results (power

over 0.8), the sample size necessary to study a small difference (<0.2 seconds) in reverberation time was as large as 100 participants. Such claim echoed the choices from the Klatte *et al.* (2010b) and Valente *et al.* (2012) studies, both of which compared two extreme reverberation times.

2.3 Speech Perception Measures

2.3.1 *Speech Intelligibility*

Speech intelligibility is often used to describe how clearly speech can be perceived in acoustic environments. There are two ways of quantifying speech intelligibility, either through measuring the physical acoustic environment or through human subject experiment.

In architectural acoustics, speech intelligibility is commonly expressed in terms of the speech transmission index (STI) or speech intelligibility index (SII). STI was first introduced by Steeneken and Houtgast (1980) to measure the quality of acoustic transmission channels (e.g., telephone line, room). The rating spans continuously between 0 and 1, synonymous with bad to excellent quality. It was later standardized through IEC 60268-16 (International Electrotechnical Commission, 2003). Most acoustic data acquisition programs nowadays have the ability to calculate STI from the measured impulse responses, which are also used to derive other acoustic metrics such as reverberation time. SII is also a physical measure similar but not identical to STI, following the similar rating scale between 0 and 1. SII highly correlates with intelligibility rating as evaluated by human subjects. The ANSI S3.5 (2012) specifies

procedures to derive SII from measured speech levels and background noise levels across octave and 1/3 octave band frequencies.

In psychoacoustics, speech intelligibility is acquired through human subjects performing mental tasks, which often involves recalling words or sentences. Contrary to the physical measures of STI and SII, participation of human subjects is mandatory in obtaining the subjective ratings of speech intelligibility. Several word lists (i.e., CID W-22 and NU-6) and sentence lists (i.e., SPIN, HINT) are among the popular test materials for subjective speech intelligibility ratings, with percent correct as the outcome score (Hornsby, 2004). Research studies cited in this dissertation have relied heavily on this particular method in collecting the subjective speech intelligibility while exposing participants to target acoustic conditions. Furthermore, recommendations of background noise level and reverberation time in ANSI S12.60 are based on assorted research studies using subjective speech intelligibility to indicate speech perception performance.

To relate the physical and subjective measures, Hornsby (2004) pointed out that intelligibility rating in percent correct can be predicted by SII using an empirically derived psychometric function. With subjective speech intelligibility rating on the vertical axis and SII on the horizontal axis, the transfer function follows the shape of an ogive curve. It was highlighted specifically that an SII rating of 0.5 corresponded to at least 80% correct using both word and sentence lists. Analogous to a cumulative distribution function, the psychometric function rises drastically in the mid-range. As a continuous and linear scale, SII lacks granularity in describing subjective speech perception even though it has shown consistent correlation with the subjective rating.

2.3.2 *Speech Comprehension*

Although subjective speech intelligibility can be a reliable measure and has had a long history of successful research application, it has not correlated well with student learning outcomes when the design of background noise level and reverberation time is in compliance with ANSI S12.60. Conceptually, speech comprehension as the ability to understand and infer spoken speech based on context, involving more upper level cognitive processing, is perhaps the more appropriate measure of learning outcome.

Two recent studies employed both speech comprehension and speech recognition tasks under assorted acoustic conditions in controlled laboratory settings. Klatte *et al.* (2010b) investigated language comprehension in a classroom-like setting under four combinations of noise type (activity noise vs. babble noise) crossed with RT (0.5 vs. 1.1 seconds). Reverberation was simulated using a virtual room technique through an electroacoustic system *in situ* in the test lab. Participants were randomly assigned as a group to one of the four acoustical conditions. In addition to the significant negative impacts of noise and reverberation, the results indicated that listening comprehension (paper-pencil instructional task) was more impaired than speech recognition (word-to-picture matching task) under the presence of both types of noises. This is further supported by Valente *et al.* (2012), who also tested four combinations of SNR (+7 vs. +10 dB) crossed with RT (0.6 vs. 1.5 seconds). All four acoustic conditions were simulated by augmenting the simulated virtual sound field *in situ* on the test lab. In this study, each participant was randomly assigned to and tested individually for one of the four conditions for both speech comprehension (clear speech or group discussion task) and speech recognition (sentence recognition task). Although no direct comparison was

made for speech comprehension versus recognition, their results implied that the detrimental effect of reverberation and noise was more prominent in speech comprehension tasks than in speech recognition tasks.

Both studies provide some empirical evidence that the negative effects of background noise and reverberation are more detrimental to speech comprehension, which involves higher level cognitive processing. The neighborhood activation model (NAM) by Luce and Pisoni (1998), although later updated, may grant some merits on such interpretation. According to NAM, a set of acoustic-phonetic patterns become activated with a stimulus presented. A recursive process is carried out in the "word decision unit" based on the probability of the activated pattern matching the target stimulus. The process terminates when the activated pattern matches that of the stimulus, thus arriving at word recognition. The time lapse during the recursive process is affected by the characteristics of the target stimulus (i.e., phonological neighborhood density and neighborhood frequency). To extrapolate using the NAM recursive framework, other factors may also contribute to the delay and even error in word recognition. If speech perception in noise and reverberation requires a portion of attention to eliminate the distracting acoustic artifacts, delay can be expected in the recursive process before arriving at word recognition. On the other hand, the recursive process may be further complicated if the individuals' inherent lexical characteristics differ from the norm. For an extreme example, the same target stimulus may activate a very different acoustic-phonetic pattern for a non-native listener with low English proficiency than that for a native English-speaking listener, increasing the chance of delay and even error during the recursive process. As delays and errors on the word recognition level compound over

time, the resources available to resolve meaning becomes scarce, eventually leading to poor speech comprehension performance.

2.4 Non-Native English Speakers

Most of the research studies cited in the previous section focused on the perception of native English speech by native English-speaking listeners. But the population in American classrooms is not exclusive to only native English speakers. A recent Institute of Education Sciences survey showed that 21% of students in the U.S. ages 5-17 (or 10.9 million students) speak a language other than English at home (Aud *et al.*, 2010). In addition to this population entering college in the future, the presence of non-native English speakers may be even more prominent with increasing enrollment of international students in American colleges. The Institute for International Education (2012) reported that international students consist of a record high of 3.7% (or 764.5 thousands) of all enrollments in U.S. higher education during the academic year of 2011-2012. Many of these international students have been hired to academic positions and remained in the U.S. In fact, the National Science Foundation (NSF) reported the 2008 survey that foreign-born postsecondary teachers in the fields of science, technology, engineering and mathematics (STEM)consist of 19% in psychology to 54% in engineering of the full-time academic positions requiring terminal doctoral degrees (National Science Board, 2012). Unfortunately, speech perception and production of this growing population have not been considered in the current ANSI S12.60 classroom acoustics standard.

2.4.1 *English speech perception*

Without manipulating the acoustic environments, Mackay and Flege (2004) and Højen and Flege (2006)found that non-native English-speaking listeners were more impaired than native listeners in speech recognition, even with early English language immersion (<5 years old). Several studies have suggested that non-native listeners with normal hearing experience more difficulties in speech perception than do native English-speaking listeners, particularly in noisy or overly reverberant environments (Takayanagi *et al.*, 2002; Rogers and Lopez, 2008; Shi, 2009).

A set of speech intelligibility studies specifically compared native and non-native listeners' performances on recall tasks by varying SNRs, mostly below 0 dB with the speech level lower than the background noise level. The stimuli used in the recall tasks varied between different levels of the phonological units including vowels and consonants (Cutler *et al.*, 2004), words (Rogers *et al.*, 2006; Bent *et al.*, 2010), and sentences (Bradlow and Bent, 2002; Bradlow and Alexander, 2007). They all suggest that non-native English-speaking listeners perform worse than natives under these extremely adverse listening conditions.

However, these intelligibility studies share similar limitations in the experimental methods in that they lack practical implication for acoustical design recommendations. First, many of the SNRs used in the aforementioned studies were lower than realistic SNRs in daily listening environments. The background noises used to create the SNR conditions varied between white noise and babble noise, which are rarely found in typical classrooms. Second, the stimuli were played back via headphones with participants seated in sound attenuated test chambers. This approach helped control the ambient noise

level experienced by the participants. But participants often have difficulties externalizing the sound source if the signal is presented through headphones. Furthermore, it has been found that apparent source distance is often underestimated when stimuli are played back via headphones (Zahorik, 2002). The listening experience may be biased with a sensation that the sound source is much closer than intended in a realistic classroom.

2.4.2 *Foreign-Accented Speech*

Besides experiencing more difficulties in perceiving speech, non-native English-speaking talkers are also likely to find themselves speaking with accents. Flege *et al.* (1999) studied the relation between age and degree of foreign accent in English (specifically native Korean speakers) and found that non-native talkers who arrive in the U.S. at a later age are more likely to produce more heavily accented speech throughout their lifetime. The ability to perceive foreign-accented speech has been found to deteriorate under the presence of noise, even for native English-speaking listeners. Munro (1998) found that the addition of cafeteria babble noise worsened the native listeners' ability to identify true or false single-sentence statements spoken by non-native speakers. Rogers *et al.* (2004) further demonstrated that native English listeners' performance on sentence recognition decreased faster for English sentences produced by native Mandarin speakers (even mildly accented) than by native English speakers, when reducing SNR from +10 dB to -5 dB.

The perception of speech from non-native talkers by non-native listeners has been even less researched. Bent and Bradlow (2003) identified an interlanguage speech

intelligibility benefit whereby it was easier for non-native listeners to perceive English sentences spoken by highly proficient non-native speakers, rather than by native English speakers. This phenomenon was found even if the non-native speaker and non-native listener did not share the same native language. However, little work has been done to investigate the role of background noise or reverberation on speech comprehension, when both the talker and listener are non-native English speakers.

Chapter 3 – Methodology

3.1 Introduction

This chapter discusses the general methodology used for both studies, including the creation of the assorted acoustics conditions and considerations in choosing various performance measures. The following table summarizes the similarities and differences in the methodologies between the two studies.

Table 3.1 - Summary of methodological similarities and differences between Study 1 and Study 2

Methodology	**Study 1**	**Study 2**
Acoustic Conditions	Background Noise Level (BNL): RC-30, 40 and 50 (or +21, +11 and +1 dB SNR) Reverberation Time (RT): 0.4, 0.6, 0.8, 1.0, and 1.2 seconds	
Testing Facility	Listening chamber with low ambient BNL and RT	
Test Materials (Initial Screen & Main Experiment)	Same materials	
Testing Procedures	Same procedures	
Talkers	Native American English (NAE)	Native Mandarin Chinese (NNC)
Listeners	Group 1: Native American English Group 2: Non-Native English	Group 1: Native American English (NAE) Group 2: Native Mandarin Chinese (NNC) Group 3: Non-Native English and Non-Native Mandarin Chinese (NNO)

Note: RC stands for Room Criteria. Different listeners were recruited for Study 1 and Study 2

3.2 Testing Facilities and Equipment Setup

3.2.1 *Listening Chamber*

All listening tests were conducted in the listening chamber at the University of Nebraska. The listening chamber was constructed using a room-in-room design, situated on 3-inch Kinetics Roll-out Floor Isolation system with secondary interior walls around all four sides that isolate external noise from migrating through building structural members. It has a floor area of 10 m^2 (107 ft^2) with a ceiling height of 2.56 m (8 ft-5 in) to the secondary drop-down ceiling grid. The back wall and one side wall are slightly slanted at 8° and 6° respectively to reduce flutter echo. Two 1.2 m by 2.4 m Tectum acoustical wall panels of 25-mm thickness (NRC 0.40, type "A" mounting) and four ATS corner bass traps were introduced to the interior to further reduce the ambient reverberation. The ambient mid-frequency (averaged across 500 to 2000 Hz) reverberation time is 0.22 second as measured at the listener position, located approximately at the center of the listening chamber. The ambient background noise level of the listening chamber is measured at RC-28 hissy (or 38 dBA), with the air ventilation system in operation during the active testing mode. Detailed ambient reverberation time and background noise level per 1/3 octave band frequency data can be found in Appendix A.

3.2.2 *Equipment Setup for Speech Comprehension Testing*

A pair of monitor loudspeakers (Yamaha HS80M, 8-inch cone) was utilized for playing back speech materials in the listening chamber during speech comprehension testing. The loudspeakers and the listener seat were positioned to form an equilateral

triangle with spacing at 1.52 m, with the loudspeakers cone axles oriented at the listener. A customized computer program interface was developed for displaying test materials and recording listener participants' responses during the speech comprehension testing. The program was operated on a Dell (Precision M2400) laptop computer, which was connected to an external PreSonus AudioBox 44VSL USB audio interface to bypass the computer internal sound card then to the two-channel monitor loudspeakers. Since all speech materials were digitally convolved with reverberation conditions prior to playback (discussed later in Section 3.2.1), additional equipment was not necessary for adding reverberation into the speech materials during real-time playback. A 23-inch monitor screen was placed in the listening chamber between the monitor loudspeakers to display the test program interface for listeners during speech comprehension testing. A second monitor screen on an 11-inch laptop was placed directly underneath the main screen for a different task. Appendix A includes photographs of the listening chamber interior and equipment set-up as seen by the listener participants during the main experiment.

A separate equipment setup was arranged for introducing background noise in the listening chamber. A desktop computer was connected to an Armstrong i-Ceiling amplifier that delivered signals to an overhead i-Ceiling loudspeaker and a corner sub-woofer in the listening chamber. All auxiliary equipment in the listening chamber during speech comprehension testing was placed in the monitor chamber and away from the common partition to prevent noise from leaking into the listening chamber. Schematics showing equipment connections are included in Appendix A.

3.2.3 *Facilities and Equipment Setup for Recording Speech Materials*

Recording of the speech materials was conducted in an anechoic chamber with native American English talkers for Study 1 and a sound attenuated booth with native Mandarin Chinese talkers for Study 2. The sound booth has heavy metal enclosure with a floor area of 3.4 m^2 (36 ft^2) and a height of 1.98 m (6 ft-6 in). It has very low background noise level measured at RC-23 hissy (or 33 dBA), and low mid-frequency reverberation time of 65 milliseconds averaged across 500 to 2000 Hz. The detailed ambient background noise levels and reverberation times per 1/3 octave band frequency are included in Appendix B.

The hardware used for recording speech materials in the sound booth included a Bruel and Kjaer microphone (½-inch transducer with wind screen) with flat frequency response, an Alesis MultiMix8 multichannel USB audio interface, and the Dell Precision laptop computer. The open source software Audacity (version 2.0.5) was used for recording and editing the speech materials. The talkers were instructed to speak in front of the microphone at no further than 20 cm (approximately 8 inches) away. The close-microphone recording technique was expected to minimize artifacts in the recorded speech in the low reverberant sound booth.

The sampling frequency was set at 44.1 kHz with 16 bits resolution for all recordings in both studies. No re-sampling was performed on the recorded speech materials during audio editing in Audacity. All audio segments were saved into the WAV format before embedding reverberations using the acoustic stimuli.

3.3 Acoustic Metrics and Control Measures

In order to study the acoustic effect, BNL and RT were systematically manipulated and presented to listener participants during the main experiment of speech comprehension testing. Since the ability to comprehend speech, regardless of acoustic environment, is highly dependent on the listeners' baseline English proficiency levels, a measure of English proficiency was developed to control for the comprehension performance when investigating the effect of assorted acoustic conditions.

3.3.1 *Acoustic Stimuli*

To expand beyond research conducted by Klatte et al (2010; 2 noise-type X 2 SNR) and Valente et al (2012; 2 SNR X 2 RT), a wider range of realistic acoustic conditions were utilized in this dissertation. A total of 15 acoustic conditions were created from combinations of three conditions of BNL (RC-30, 40 and 50) and five scenarios of RT (0.4 to 1.2 seconds).

3.3.1.1 *Background Noise Levels*

As mentioned in the previous section, background noise was introduced via a subwoofer at the corner of the chamber and an i-Ceiling loudspeaker integrated behind an acoustical panel above the listener position. To calibrate the test signals, pink noise was first introduced then digitally filtered to create three conditions of BNL that followed the Room Criteria contours of RC-30, 40 and 50. The steady-state BNL values for the three test conditions were measured at the listener position and shown in Figure 3.1. During

main experiment testing, the BNL test signals in WAV format were played back continuously.

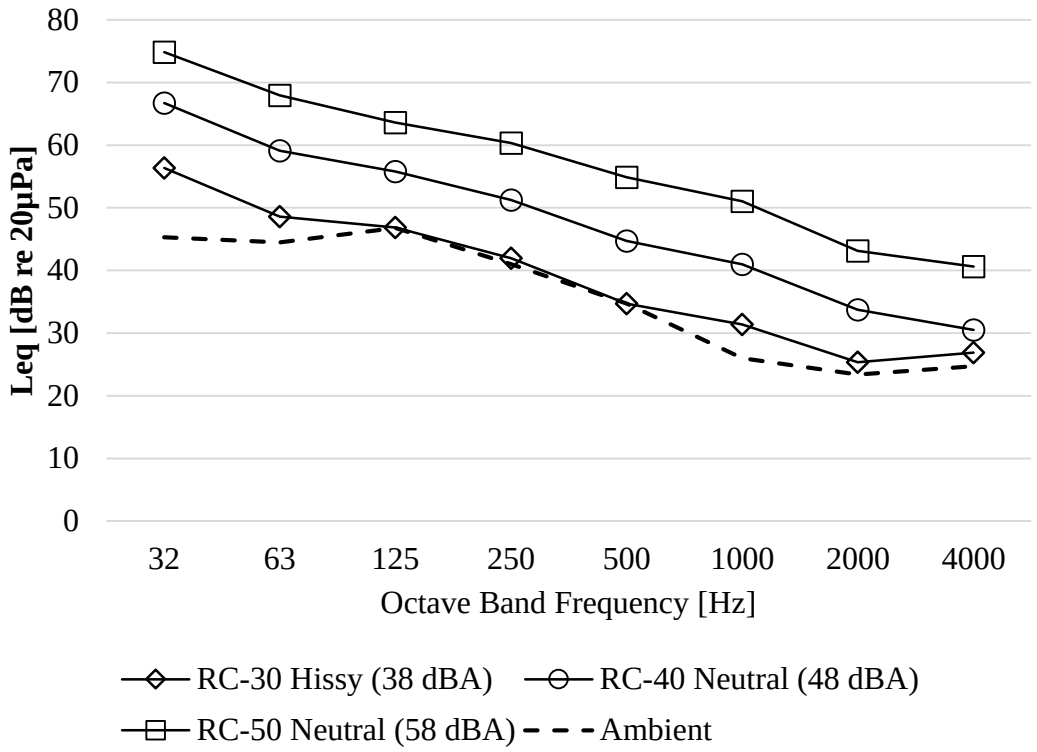


Figure 3.1 - Background noise levels measured at the listener position in the listening chamber during ambient and test conditions

3.3.1.2 *Reverberation Time Scenarios*

To create the RT scenarios, a typical classroom of 260 m^3 (9182 ft^3) was simulated in the auralization program ODEON. Different ceiling materials in combination with 25-mm acoustical panels (NRC 0.70), applied full height on the side and back walls with uniformly scaled absorption coefficients, were utilized to create the five RT scenarios from 0.4 to 1.2 seconds with approximately equal intervals. The

simulated RT under each material configuration is documented in Table 3.2. In the ODEON model, the source and receiver were designated at a relative 4-meter distance to simulate a typical middle seat in the classroom with the talker on center at 1.5-meter away from the front wall. The binaural room impulse responses (BRIR) of the RT scenarios were then exported from ODEON after adjusting for the relative location of the two-channel loudspeaker and the listener position in the listening chamber. The BRIRs were then digitally convolved with speech comprehension materials in Matlab.

Table 3.2 - Documentation of simulated RT scenarios

RT Scenario [sec]	Simulated RT [sec]	Measured RT [sec] in Listening Chamber	Uniform Scale Factor	Ceiling Material
0.4	0.34	0.37	75%	NRC 0.70
0.6	0.6	0.62	30%	NRC 0.70
0.8	0.81	0.84	15%	NRC 0.55
1.0	1.01	1.05	5%	NRC 0.55
1.2	1.18	1.19	9%	GWB

Since the listening chamber was not anechoic, the actual RT measured at the listener position slightly differed from the simulated RT (see Table 3.2). Hak and Wenmaekers (2013) suggested that, for playback in a non-anechoic chamber, the relative error of the resulting RT is less than 10% of the input RT if the ratio between the input and chamber RTs is less than 2. With an ambient reverberation time of approximately 0.22 second across octave band frequency and much shorter than most of the test conditions, the artifacts introduced in the speech materials were expected to be at most 8% for the 0.4 second RT scenario in the high frequency range. The RT measured in T_{20}

on octave band frequencies from 125 Hz to 8000 Hz for each test scenario are shown in Figure 3.2.

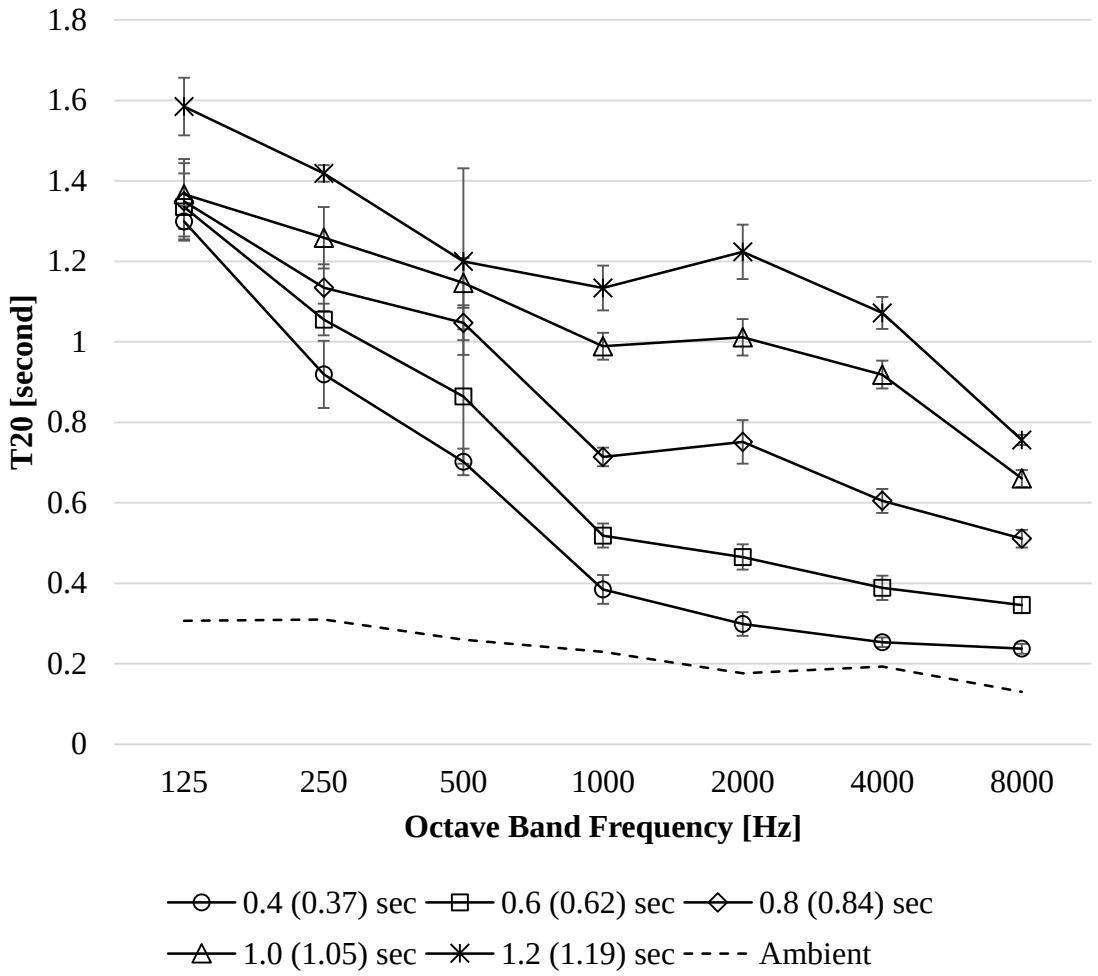


Figure 3.2 - RT in T20 from 125 Hz to 8000 Hz, measured at the listener position in listening chamber, for the ambient and five RT scenarios from 0.4 to 1.2 seconds. Error bar indicates one standard deviation from 10 *in situ* measurements. Single numbered T_{20} in parenthesis are actual measured RT averaged from 500 Hz to 2000 Hz.

The determination of speech level was based on previous studies. Klatte *et al.* (2010b) used a source level of 66 dB at 1-meter for raised voice during lecturing. With ODEON's recommendation of -3.5 dB per doubling distance in a diffuse reverberant field, a 7 dB reduction in sound pressure level is expected from the virtual talker to the listener at 4 meters away. As a result, all convolved speech comprehension materials were calibrated to playback at the listener position at 59 dBA, across all RT scenarios. A similar sound pressure level of 60 dBA was utilized for signal presentation by Valente *et al.* (2012). The resulting signal-to-noise ratio (SNR) was +21, +11, and +1 dB for the RC-30 (38 dBA), RC-40 (48 dBA) and RC-50 (58 dBA) condition, respectively.

The speech intelligibility index (SII) was calculated per ANSI S3.5-1997 for each acoustic combination and are shown in Figure 3.3. The speech transmission index (STI), calculated using monaural room impulse responses in WinMLS 2004, is reported for each acoustic condition in Figure 3.4. In general, both SII and STI reduced drastically for the RC-50 condition in comparison to the two lower BNLs. They also reduced slightly with increasing RT. STI seemed to be more sensitive than SII to the change in BNL and RT. Based on the qualitative designations proposed for STI by Houtgast and Steeneken (1984), the intelligibility of speech ranged from "poor" under RC-50 to "fair to good" under RC-30 and RC-40 BNL. Quantitatively, Hornsby (2004) summarized the psychometric function between percent correct in recognition tests (i.e., CID W-22, NU-6, and Connected Speech Test) and SII and showed that 0.6 SII corresponded to at least 80% correct in speech intelligibility as perceived by participants.

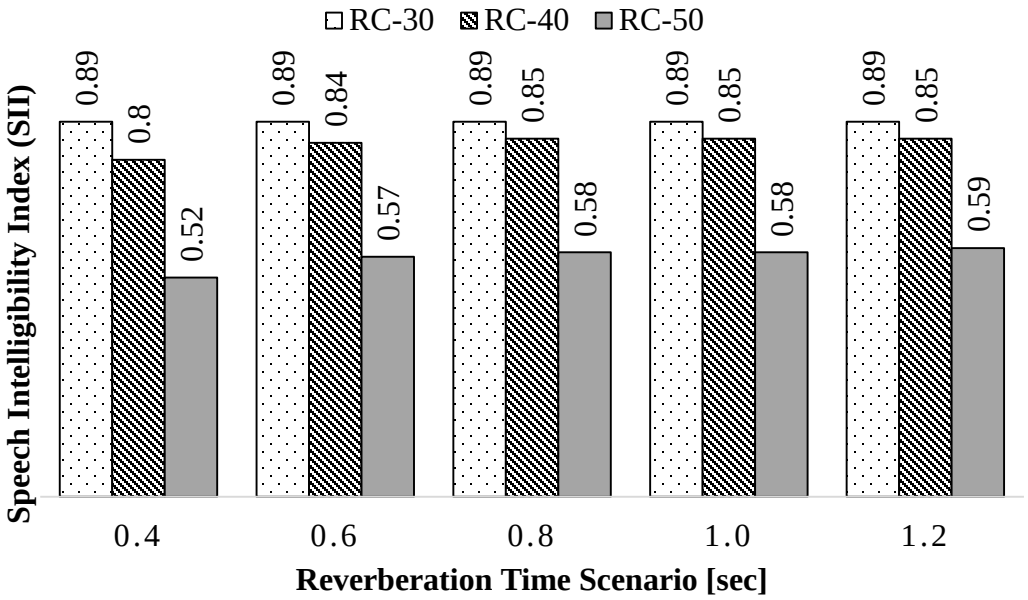


Figure 3.3 - Speech intelligibility index (SII) for each acoustic condition

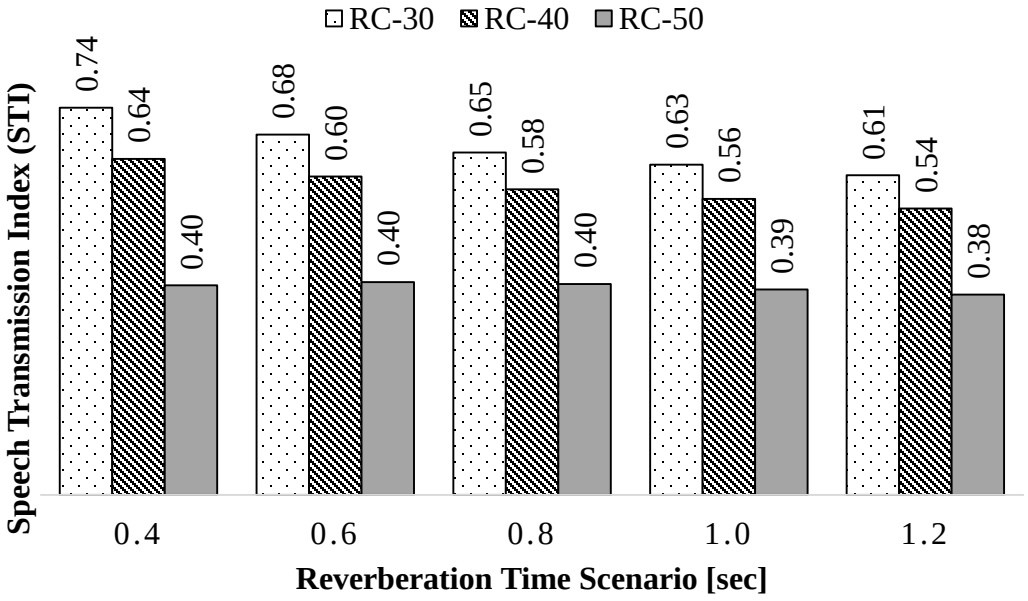


Figure 3.4 - Speech transmission index (STI) for each acoustic condition

3.3.2 *Composite Scale of English Proficiency Levels*

Conceptually, speech comprehension performance relies heavily on listeners' proficiency in using the language. Individual listeners' English proficiency level will confound speech comprehension performance under assorted acoustic environments, and hence must be controlled in the statistical analysis in order to better understand the genuine effects of room acoustics. During initial screening, all listener participants were individually given three tests pertinent to English language proficiency, covering listening span, oral comprehension, and verbal abilities.

3.3.2.1 *Listening Span*

A study conducted by Daneman and Carpenter (1980) showed that comprehension (both reading and listening) correlated significantly with working memory, as measured in listening span. Individuals' working memory capacity determined the amount of information available during the cognitive processing of speech comprehension. Furthermore, the differences in listening span may not only lie in individuals' cognitive abilities, but also the linguistic characteristics of their native languages (Ellis and Hennelly, 1980). To measure listeners' individual working memory, the listening span subtest was adopted from the Woodcock-Johnson III NU Tests of Cognitive Abilities (Woodcock *et al.*, 2001b). In this test, participants were asked to repeat each spoken sentence after it was played via headphones. The recorded sentences became increasingly longer and the test ended when participants could no longer recite these sentences perfectly.

3.3.2.2 *Oral Discourse*

The subtest of English oral discourse was chosen from the Woodcock-Johnson III NU Tests of Achievement (Woodcock *et al.*, 2001a) to measure listener participants' baseline ability of oral comprehension in English. Both Woodcock-Johnson III test packages had been previously normed for measuring cognitive abilities and oral language abilities of individuals from 2 to over 90 years of age. For the oral comprehension test, recorded sentences, each with a missing last word, were presented to participants. They were asked to verbally respond what the missing word should have been based on the context of the sentence.

Both listening span and oral comprehension tests involved spoken materials. These materials were recorded by a female native American English speaker in a former listening chamber (BNL < 30 dBA) using a closely aligned microphone. During the individual English proficiency testing, these recorded materials were played back for participants via headphones. Participants were encouraged to choose a comfortable listening level of L_{eq} between 65 and 68 dBA re 20 μPa (L_{max} between 70 to 75 dBA).

3.3.2.3 *Verbal Abilities*

The English portion of the Bilingual Verbal Ability Tests (BVAT) by Muñoz-Sandoval *et al.* (1998) was selected to be the third measure of English proficiency during initial screening. The BVAT has been normed for measuring overall verbal ability (English only in this project) of individuals from 5 to over 90 years of age. The BVAT test is typically first given in English, then supplemented with materials in the test

participant's native language to obtain the overall verbal ability. In this dissertation, only the English portion of BVAT was deployed in the proficiency testing.

During the BVAT test, a test book was utilized along with verbal instructions provided by the author (non-native English speaker) to assess participants' verbal abilities in three areas: 1) picture vocabulary, 2) oral vocabulary (i.e., synonyms and acronyms), and 3) verbal analogies. The majority of the BVAT test utilized visual materials displayed on the test book. The author administered the BVAT test to all participants in this project and adhered to the test guidelines on giving succinct verbal instructions, mostly to encourage participants and during transition between test items. The effect of the author's foreign accent was considered minimal in obtaining this measure.

3.3.2.4 *Composite Scale*

The three tests were used to form a composite scale to measure individual participants' overall English proficiency level. The raw scores from each test were first verified to conform to normality before being converted into standardized z-scores. The composite scale was then calculated by taking the mean of the z-scores of the three proficiency tests. The composite scale achieved excellent internal consistency with Cronbach's α of 0.938 using data from both studies, suggesting a near perfect measure of English proficiency.

3.3.3 *Native versus Non-Native English-Speaking Listener Groups*

The Language Experience and Proficiency Questionnaire (LEAP-Q) developed by Marian *et al.* (2007) with revision was used to survey the English language experiences for all listener participants (see Appendix C). Based on self-report, the revised LEAP-Q provided a comprehensive understanding of participants' English language experience, including survey items on order of language acquisition, order of language dominance, age of English onset, length of English immersion, and perceived English proficiency levels in reading and listening. Although the definition of non-nativeness remained debatable, the order of language acquisition provided the best prediction of listeners' English proficiency levels in this study (see Chapter 7 for discussion on confounding factors). Therefore in both studies, listener participants were placed into listener groups based on the first language they acquired during early childhood. Chapters 5 and 6 provide more descriptions of both native and non-native English-speaking listeners tested in both studies.

3.4 Performance and Perception Measures

This dissertation aims at studying the acoustic effects on both objective performance on speech comprehension tasks and subjective perception of task workload by the listener participants in order to determine the acoustic design guidelines. The following section provides descriptions on the measures used to obtain objective performance and subjective perception ratings, which were entered into statistical analyses as dependent variables.

3.4.1 *Dual-Task Scheme for Measuring Performance*

A bimodal dual-task paradigm was utilized in testing for performance under assorted acoustic conditions. During the main experiment, participants were asked to simultaneously perform an adaptive pursuit rotor (APR) task and speech comprehension tasks while immersed in the acoustic test conditions. The equipment set-up for the dual-task scheme is outlined in the equipment schematics in Appendix A.

The dual-task paradigm was adopted based on two considerations. First, during a pilot study where only the speech comprehension tests were administered, both native and non-native listeners achieved at least 80% correct even under the worst acoustic condition. Little variation of the percent correct score was observed among other acoustic conditions, suggesting signs of performance plateau perhaps due to the simplicity of the speech comprehension test materials. A secondary task of a different modality was incorporated, assuming it would uniformly diminish listeners' comprehension performance by removing a consistent amount of attention away from the speech comprehension tasks. The APR task revised from the conventional pursuit rotor task by Srinivasan (2010) was hence chosen as the secondary completing task. It was re-designed to include an algorithm to change speed adaptively to keep participants at an 80% on-target accuracy while tracing the dot. The performance of the APR task was recorded as rounds per minute (RPM). The second consideration of incorporating a simultaneous competing task was the reality of classroom activities, in which listeners are expected to multi-task during speech comprehension in the learning experiences (e.g., note taking and critical thinking).

3.4.2 *Primary Performance Measure: Speech Comprehension*

A total of 18 sets of speech comprehension tests in English, of which 15 sets shared equivalent difficulty level, were created from preparation materials for the listening tests of the Test of English for International Communication (TOEIC). These test items were recorded by native English speakers (one male and four females) in an anechoic chamber for Study 1 and by native Mandarin Chinese speakers (one male and one female) in a sound attenuated booth for Study 2. These materials were created to target daily life events with simple vocabularies and could be understood easily by non-native English-speaking listeners with low English proficiency. Each test was randomly paired with one of the 15 acoustic conditions for each participant and lasted no more than 15 minutes. There were 32 multiple choice items in each test, comprised of four tasks as outlined below. Performance was recorded in percent correct based on the accuracy of the 32 test items.

1) Photograph Recognition (4 items): Participants identified one of four spoken sentences that best matched the photograph displayed on the computer screen.
2) Question and Response (10 items): Participants identified one of three spoken sentences that best responded to the spoken question.
3) Conversation (3 conversations X 3 questions, 9 items): Participants listened to a conversation exchanged between a male and a female talker and answered three spoken questions related to the content with answer options displayed on the computer screen.

4) Paragraph (3 paragraphs X 3 questions, 9 items): Participants listened to a short paragraph and answered three questions pertinent to the content, again with answer options displayed on the computer screen.

The test items were presented with talkers of alternating gender within each task. For example, in task 2) Question and Response, if the question was asked by a male talker, the response options would be spoken by a female talker; the subsequent test item would change talker gender with the female asking the question and the male responding.

To ensure equivalent content difficulty level across the 15 sets of tests to be disseminated under acoustic test conditions, all test items were individually screened by five native English-speaking listeners. During the content screening, the speech materials were played back using the version recorded by the native American English speakers and under the same set-up as the actual speech comprehension testing (see Section 3.2). The five native English-speaking listeners (all male) were individually seated in the listening chamber, with speech materials played back under the ambient chamber condition without introducing the test conditions of BNL or RT. Each test item received a percent correct score as answered by the five listeners. Ambiguous items were identified if individual test items were answered incorrectly by more than two of the five listeners. All ambiguous items were excluded from the equivalent test sets for testing under acoustics, but some were used in the practice trials at the beginning of each new BNL condition. Each test received an overall content score for the 32 items between 89% and 91% as understood by the five native English listeners under ideal acoustics of the ambient condition in the listening chamber. The five native English listeners who

participated in the content screening were asked not to participate further for either main study.

3.4.3 *Secondary Performance Measure: Adaptive Pursuit Rotor*

The APR dot-tracing task was developed by Srinivasan (2010) by adding an adaptive speed algorithm to the conventional pursuit rotor task. During the APR task, participants were asked to trace a dot that continuously rotated around a fixed ring. The speed of the dot rotation changed adaptively to engage participants on target at 80% accuracy. The steps in updating the rotation speed was set at 5% of the previous speed, which was updated every second. The APR task was operated on an 11-inch Dell Inspiron laptop computer with the screen directly below the primary monitor screen for speech comprehension tasks. Listeners were asked to switch their visuals up and down to accommodate the visual cues on both tasks during the main experiment. A wired stylus and pad was connected to the laptop computer and provided to the listener for the tracing task using their dominant hand.

It was expected that the simultaneous APR task would require a portion of listener participants' attention while performing the speech comprehension tasks. It was hypothesized that, under divided attention, the performance on speech comprehension tasks would decrease with the implementation of the simultaneous APR task.

3.4.4 *Subjective Perception Measure*

The self-report NASA Task Load Index (TLX) questionnaire was developed by Hart and Staveland (1988) and has a long history of application to survey subjective

assessment of task workload. In a 20-year review of NASA TLX and its application, Hart (2006) pointed out that 31% of over 500 studies using the questionnaire involved visual or auditory evaluation. The NASA TLX surveys task workload using six subscales, with the computerized version included in Appendix C. The original NASA TLX applied weighting on the raw rating of each subscale based on pair-wise comparison. A simplified application of NASA TLX eliminated the weighting scheme by examining individual subscales closely instead of a weighted overall rating. The simplified approach was supported by Hart (2006).

Based on its relevance to auditory evaluation and simplicity in application, the NASA TLX was chosen to survey participants' subjective perception of the dual-tasks to complement their objective performances under assorted acoustic conditions. The questionnaire was given immediately after each speech comprehension test, and repeated for all 15 acoustic conditions tested.

3.5 Other Measures

3.5.1 *Self-Report English Language Experience*

As previously mentioned, in order to obtain a comprehensive understanding of participants' English language experiences, the LEAP-Q developed by Marian *et al.* (2007) was adopted with minor revisions. The revised LEAP-Q used during the initial screening session for all listener participants is included in Appendix C.

The LEAP-Q was normed for obtaining self-reported history and proficiency across all known languages on adults, who have obtained at least high school education in their native language. A subset of the original LEAP-Q items was utilized in this

dissertation. The following items were included as the language history measures to obtain self-report English language experiences among all listener participants.

1) Order of acquisition and dominance of all known languages
2) Self-report proficiency in understanding, speaking, reading, and writing in English
3) Onset age of learning and fluency of speaking and reading English
4) Duration of English immersion in the country, family, and school settings

3.5.2 *Noise Sensitivity*

A reduced version of the original Noise Sensitivity Questionnaire (NoiSeQ-R) was deployed to examine the role of noise sensitivity in speech perception under acoustic environments. The NoiSeQ-R is extracted from the full length NoiSeQ (Sandrock *et al.*, 2007; Schutte *et al.*, 2007a; Schutte *et al.*, 2007b; Griefahn, 2008). It was originally disseminated online as part of a cross-country study to investigate the social attitudes toward traffic noise in Europe.

The online NoiSeQ-R was incorporated into a paper-pencil format as part of the demographic survey for all listener participants (see Appendix C). It contained 13 items using a four-point scale that surveyed three domains of noise sensitivity: sleep, work, and residential surroundings. The outcome of the NoiSeQ-R included individual ratings of noise sensitivity in the three domains and an overall rating.

3.5.3 *Potential Confounding Factors*

To better explain the variance observed in speech comprehension performance under acoustic environments, several potential confounding factors were identified and discussed below.

3.5.3.1 *Talker Speech Rate*

Talkers with faster speech rate are generally more difficult to understand, particularly for non-native listeners with lower language proficiency levels (Bradlow and Pisoni, 1999). During the speech material recordings, talkers were instructed to speak comfortably without specific requirements on maintaining a particular speech rate. To calculate speech rate in syllables per second, the original recordings without embedding the simulated BRIRs were imported into Audacity to examine the sentence duration by highlighting the waveform. The number of syllables were manually counted from the audio scripts. This task was performed by two undergraduate research assistants who were both native English-speakers. Because in the design of experiment to counterbalance the appearance of each talker voice in the speech comprehension test sets, the effect was in fact unable to quantify and be treated as random effect. The speech rate of each talker is reported in Chapters 5 and 6.

3.5.3.2 *Temperature*

Thermal comfort was previously found as a stronger predictor than acoustics in affecting participants' perception and task performance (Tiller *et al.*, 2010). Therefore, the temperature in the listening chamber was monitored and recorded either at the

beginning or end of each one-hour main experiment session. It was observed that temperature did not often fluctuate more than 1ºF during the hour-long session; hence a finer resolution of temperature recording was not necessary.

3.5.3.3 *Handedness*

Handedness was inquired prior to the main experiment testing, mainly for equipment set-up purpose. Since the dual-task scheme involved fine motor skills for the APR task and cooperation of both hands during testing, it was later analyzed for its potential confounding effect.

3.6 Listener Testing Procedure

Both Study 1 and Study 2 followed the same general procedures during individual initial screen and the main experiment of speech comprehension testing. The following section provides details of the screening and testing procedures.

3.6.1 *Initial Screen*

At the beginning of the initial screen, the listener participants were given an orientation program created in PowerPoint for previewing the testing procedures utilized throughout the study. They were then asked to read and sign the informed consent form, and were provided a signed copy to take with them. Participants were encouraged to ask questions during the screening process.

After the signed informed consent form was collected, an audiometric screen was given either in the sound booth or the listening chamber using a Grason-Stadler GSI17

audiometer. Eligible participants needed to be able to listen to pure tones of 25 dB hearing level or lower from 125 Hz to 8000 Hz for both ears. If participants failed to meet the hearing screen requirements, they were given a $5 gift card and asked not to participate further in the study.

Once the participants passed the hearing screen, they were given a demographic survey which included select items from the LEAP-Q and NoiSeQ-R. An additional items on furthering the understanding of English dominance were incorporated to ask whether participants have ever dreamed in English. Additional demographic questions included those regarding gender, age, ethnicity group, and past experience with standardized tests (i.e., TOEIC, TOEFL, GRE, SAT, and ACT).

Next, the three sets of English proficiency tests were given to the participants. All three proficiency tests were administered by the author to maintain consistency of oral instructions. Although a range of English proficiency levels were preferred, several potential non-native English-speaking participants were disqualified and asked not to participate further. These participants either recently began residency in an English dominant country, usually for less than a month, or had no experiences studying in an English classroom (e.g., spouses of foreign students). And, they all scored very low on the proficiency tests. Hence, they were asked not to participate further in the main experiment due to their lack of representation of the target population.

3.6.2 *Main Experiment*

After completing the initial screening, participants were invited back over six one-hour long sessions on separate days to conduct the main experiment. Each session consisted of three speech comprehension tests, which corresponded to testing for three acoustic conditions. From the investigators' previous experience, participants tend to become more conscious of the environmental change from changing background noise level. To reduce participants' sensitivity toward the experimental design, the three tests in each hour-long session contained the identical BNL but with varying RT embedded in the speech materials. The test sequence of each one-hour session is illustrated in Figure 3.5. A practice trial was also given every time a new BNL test condition began and was excluded from data analysis. A nested Latin square design was utilized to counterbalance the order of presentation for both BNL and RT. A two-factor within-subject design, 3 BNL (RC-30, 40 and 50) X 5 RT (five scenarios from 0.4 to 1.2 seconds), was achieved by exposing each participant to all 15 acoustic conditions.

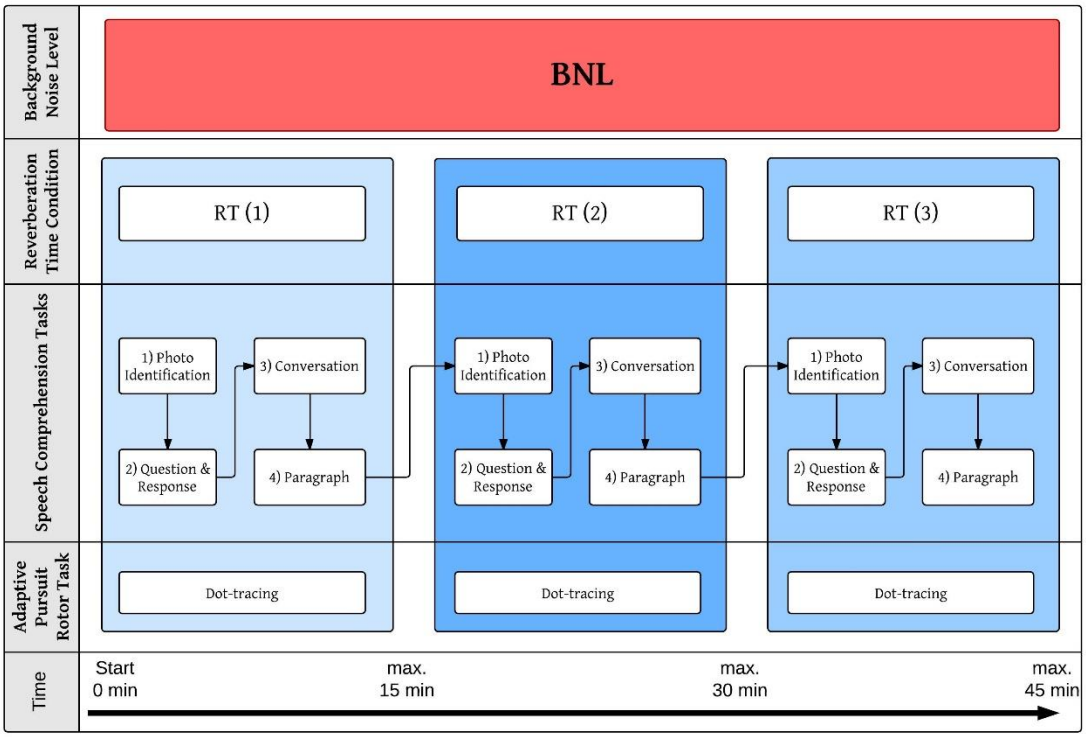


Figure 3.5 - Flow diagram showing test sequence within each one-hour session in the main experiment

Prior to the first speech comprehension test, participants were given a 3-minute practice trial on the APR task only. During each test, all questions in the speech comprehension tasks were in the multiple choice format. Participants responded using a labeled number keypad with their non-dominant hand. Simultaneously, participants performed the APR dot-tracing task using their dominant hand using a wired stylus and pad. Participants were asked to shift their visuals up and down between the two monitor screens (see Appendix A) to accommodate the dual-tasks and not to take priority of either. They were also instructed to refrain from leaning forward or moving sideways if the speech materials became difficult to listen to. After each test, participants were given

a computerized NASA TLX survey to express their subjective opinions regarding the test completed. The APR task was inactive when participants were filling out the survey.

Once participants submitted the NASA TLX survey, the customized computer program would prompt them to start the next test. Each test lasted no more than 15 minutes total including the subjective survey. Participants were allowed to take breaks between tests within the same one-hour session if necessary. They were also encouraged to share their testing experiences with the proctor after each test session.

Listener participants received $5 per hour during the initial screen and main experiment. If all main experiment sessions were completed, participants received an additional lump sum to reach a total of $100 for completing the study.

Chapter 4 - Statistics

4.1 Introduction

This chapter discusses the techniques related to conducting statistical analysis in this dissertation. Only parametric tests, which require normal (also known as Gaussian) distributions, were applied in data analysis due to the variety of statistical models available in answering the research questions in this dissertation. In the case when a variable was not normally distributed, transformation was applied to scale it to approximate normality. All data analysis was performed in IBM SPSS (version 22) and G*Power (version 3.1, (Erdfelder *et al.*, 1996; Faul *et al.*, 2007)).

4.2 Data Examination and Treatment

Before conducting any statistical testing, Hair *et al.* (2006) recommend a thorough examination of all applicable variables to understand their properties and to discover anomalies in the data.

4.2.1 *Variable Type*

There are three types of variables: continuous, ordinal and categorical. Continuous and ordinal variables are also often known as metric variables and categorical variables as non-metric variables. Table 4.1 lists the variable type and possible value range for each major measure in this dissertation.

Table 4.1 - Variable type and value range for select measures

Measure	Variable Type	Possible Range
Performance and Perception		
Speech Comprehension	Continuous	0 to 100 in percent correct (-23 to 123 in RAU)
Adaptive Pursuit Rotor	Continuous	> 0 RPM
NASA Task Load Index	Ordinal	0 to 100
Acoustic		
Background Noise Level	Ordinal	RC-30, RC-40 and RC-50 (or +21, +11 and +1 dB SNR)
Reverberation Time	Ordinal	00.4, 0.6, 0.8, 1.0 and 1.2 sec
Talker and Listener		
English Proficiency Level	Continuous	-3 to 3 in standardized Z-score
Listener Group	Categorical	Study 1: Native vs. Non-native Study 2: English (NAE) vs. Chinese (NNC) vs. Other Non-native English (NNO)
Talker Accent	Categorical	English (NAE) vs. Chinese (NNC)

4.2.2 *Missing Data and Outliers*

4.2.2.1 *Missing Data*

Missing data are generally more common in data collection via questionnaires, where the participants provide no response to one or more items. According to Hair *et al.* (2006), the first step to treating missing data is to determine whether the amount of missing values is substantial in the whole dataset (i.e., > 10%). Subsequently, the pattern of the missing data should be evaluated to check for randomness. These two steps are to prevent losing useful data by the simple treatment of listwise deletion, in which all

responses from a participant will be excluded in analysis if he or she fails to respond to a single item. Once the decision is made to retain participants with missing data, the treatment includes pairwise deletion (participant retained for non-missing variables) or missing value replacement. Since the repeated-measure design cannot facilitate pairwise deletion, the latter approach was adopted for treating missing data in this dissertation.

The majority of the testing was conducted either under supervision during initial screens or with computer prompts in the main experiment. Only under rare circumstances of hardware system failure did the computer not archive results from the APR dot-tracing task. This only occurred in one trial for two listener participants (one from each study) among the 11,725 trials administered. The missing value was then replaced by the mean calculated from the remaining participants in the same study under the same acoustic condition. A different approach was utilized to replace missing data for the temperature measure (discussed in Chapter 7 as potential confounder). The missing temperature record was replaced by the reading from another participant tested during the similar time frame during the same day, since temperature did not change rapidly in the lab controlled environment.

4.2.2.2 *Outliers*

Outliers are observations identified as distinctively different from the remainders. They may substantively skew the distribution and, in some extreme scenarios, lead to biased results in the subsequent statistical testing. Hair *et al.* (2006) discussed several ways to detect outliers in the data (p68-70). The treatments of outliers (Hair *et al.*, 2006; Field, 2009) include case removal, data transformation, and value replacement. With the

massive amount of data in this dissertation, the detection of outliers was completed by exploring the boxplots of the dependent variables (e.g., comprehension performance and NASA TLX subscales) under each acoustic condition. In SPSS, a boxplot signals outliers of two kinds, mild outliers as data points between the 1.5 and 3 times of the interquartile range (IQR) away from the median and extreme outliers beyond the 3 IQR.

As a precaution of potential non-native English speaker with exceptional English proficiency levels, a slightly different outlier detection approach was utilized before all data could be obtained. A non-native English speaker was determined as an outlier if he or she scored within one standard deviation below the mean as calculated from all native English speakers in the study on all three English proficiency tests. One non-native listener participant from Study 1 was found to achieve outlying English proficiency level using this criterion, who was also identified as outliers on most of the boxplots of speech comprehension performance of non-native listeners. After careful consideration, the outlier participant was removed from analyses involving listener groups but included when English proficiency level was controlled.

4.2.3 *Assumptions of Parametric Data*

In order to conduct parametric tests, the data needs to satisfy four statistical assumptions (Hair *et al.*, 2006; Field, 2009), including normal distribution, homoscedasticity, linearity, and independence in error.

4.2.3.1 *Normal Distribution*

According to the central limit theorem, the sampling distribution will conform to normality if the sample drawn from the population is large enough. It is fundamental to

the sampling method, in which a good sample should represent the intrinsic characteristics of the population. Deviation from normality implies (but not necessarily determines) the possibility of poor sampling in the research method.

There are many ways to assess the normality of a distribution as suggested by Hair *et al.* (2006). Graphically, one can visually examine the histogram and the Q-Q plot. To quantify normality, metrics such as skewness and kurtosis are also available. In this dissertation, the Shapiro-Wilk test of normality was utilized to practically examine the large datasets. A significant Shapiro-Wilk test suggests that the actual distribution differ significantly from a normal distribution, and the assumption of normality is thus violated. In that case, data transformation should be considered to scale the distribution to approximate normality.

Among many empirical transformations, the rationalized arcsine unit (RAU) is the most commonly used transformation in auditory perception studies. It was first proposed by Studebaker (1985), who successfully scaled the non-normally distributed percent correct scores to achieve normality. The following equations to calculate RAU were adopted from the updated version by Sherbecoe and Studebaker (2004).

$$\theta = \sin^{-1}\sqrt{\frac{X}{N+1}} + \sin^{-1}\sqrt{\frac{X+1}{N+1}} \quad (1)$$

$$\text{RAU} = \frac{146}{\pi} \times \theta - 23 \quad (2)$$

where N = total number of test items

X = number of correctly answered items

θ in radian

4.2.3.2 *Homoscedasticity*

This assumption is also known as homogeneity of variance, which states that the variance across all levels of the variable should be consistent. It can be evaluated using Levene's test. A significant Levene's test suggests that unequal variance exists across different levels of the variable, and thus homoscedasticity cannot be assumed. The remedy to heterogeneous variance is to apply data transformation similar to the approach to correct non-normal distribution (Hair *et al.*, 2006). In processing the dissertation data, Levene's tests were verified for the error variance in the dependent variables to ensure that the homoscedasticity assumption had been satisfied.

4.2.3.3 *Linearity*

This particular assumption requires that the relation among variables can be modeled mathematically. It does not mean that the relation has to be linear in the sense of a straight regression line. In this dissertation, the research questions (see Chapter 1) were proposed based on extensive literature review (see Chapter 2), from which the results indicated and projected relations among the measures in the statistical models in Chapters 5, 6 and 7. As a result, the linearity assumption was confirmed via logical reasoning rather than additional statistical analysis, although it is possible according to Hair *et al.* (2006).

4.2.3.4 *Independence in Error*

Unlike the previous assumptions, the independent error assumption cannot be confirmed prior to performing statistical testing. In every parametric model using

dependence technique, there will always exist a portion of variance in the dependent variable that the independent variables fail to explain. The unexplained portion of variance, also known as residual or error, should not be correlated with each other. The definition of this assumption may seem like an abstract concept. In fact, dependent error is often the result of confounding factors not accounted for in the model. The assumption helps reinforce a comprehensive examination of the variables in the statistical model to answer the research question. To verify this assumption, Chapter 7 provides a thorough examination of potential confounding factors in the statistical models for this dissertation.

4.3 Statistical Analysis

4.3.1 *Hypothesis Testing*

As previously mentioned in the linearity assumption for parametric testing, the relations among variables of interest can be modeled mathematically to answer research questions. All parametric testing techniques fall into the hypothesis testing framework, which is based on comparing a null hypothesis and an alternative hypothesis. In the view of this framework, all research questions can essentially be reduced to the search of an effect, whether it was a difference between groups or relations among observed phenomena. Two hypotheses (or statements) are fitted into the research question by the following designations.

Null hypothesis (H_0): A default opposition to the alternative hypothesis that there exists no effect

Alternative hypothesis (H_a): Description of an effect based on the research question

By comparing the null and alternative hypotheses, there are two possible outcomes of hypothesis testing: success or failure in rejecting the null hypothesis. However, mathematical models expressed for the hypothesis testing can never perfectly describe the relationship between the observed phenomena. Mismatched results are likely to occur between hypothesis testing and the underlying principle of the specific effect. Therefore, any result from hypothesis testing will lie in one of the four quadrants illustrated in Table 4.2.

Table 4.2 - Relations between hypothesis testing results and the underlying principle of the target effect

	Effect exists	**Effect does not exist**
Reject null hypothesis	Correct (1- α) "True Positive"	Type I Error (α) "False Positive"
Fail to reject null hypothesis	Type II Error (β) "False Negative"	Correct (1- β) "True Negative"

There are several steps in the hypothesis testing process to answer each research question.

1) Establish a null hypothesis and an alternative hypothesis
2) Select the *a priori* significance level, α
3) Compute inferential statistics, particularly the p-value
4) Determine whether the null hypothesis can be rejected

First, the null and alternative hypotheses should be carefully constructed based on the research question. Next, the *a priori* significance level α serves as a criterion in determining the rejection of the null hypothesis later and should be selected before computing the inferential statistics. A typical but arbitrarily selected value for α is .05, suggesting that if the null hypothesis is subsequently rejected, the conclusion tolerates a probability of 5% that the effect actually does not exist in the population (i.e., Type I error). In other words, the probability of a "true positive" (i.e., finding an effect where it truly exists) is 95%. Depending on the context of the research question, the value of α may vary to adjust for the tolerance of Type I error. Once the *a priori* significance level is chosen, an appropriate parametric test can be applied to compute a set of test statistics, which include the p-value. If the p-value is less than or equal to α, the null hypothesis is rejected and the alternative hypothesis is then accepted. On the contrary, if the p-value is greater than α, it suggests that there is not enough evidence to support the rejection of the null hypothesis. In this case, it is often tempting to accept the null hypothesis. But as seen from Table 4.2, the result is indecisive since the possibility of committing Type II error has not been eliminated.

In fact, the probability of Type II error β (i.e., failure in finding an effect where it actually exists) is less commonly discussed in the results from parametric tests, although it can be calculated retrospectively. The caution to avoid Type II error should be applied in determining the sample size before data collection rather than during hypothesis testing. It is well understood that a representative sample from the population is critical in research method to provide good observations of the intended phenomena (Field and Hole, 2002; Hoyle *et al.*, 2002). A misconception of sample size that is large enough to

capture the population characteristics is to follow the rule of thumb set forth by the central limit theorem (i.e., N = 30 for ANOVA and N = 200 for regression analyses). However, the strength of the population characteristics (or the effect size) can also affect the sample size needed (Field, 2009). Logically, the smaller the effect size the more observations are necessary and hence the larger sample size. The determination of sample size is governed by both effect size and the statistical power, which is the probability of a "true negative" (1- β). A conventional value, also arbitrarily selected, for statistical power is 0.80. The calculation of sample size is given in the following equation for an independent t-test.

$$n_1 = \frac{(r+1)d^2}{r}\left(Z_{power} + Z_{\alpha/2}\right)^2 \quad (3)$$

where n_1 = number of participants in group 1

r = n_2/n_1, n_2 = number of participants in group 2

d = effect size in Cohen's d

Z_{power} = Z-score corresponding to statistical power (0.84 for 80% power, β = .20)

$Z_{\alpha/2}$ = Z score corresponding to two-tailed significance level (1.96 for α = .05)

During the development phase of this dissertation, the sample size was determined primarily based on the effect sizes of the acoustic variables derived from Klatte *et al.* (2010b) and Valente *et al.* (2012). It was calculated, using G*Power (version 3.1), that the largest sample size needed was 18 participants in each listener group to achieve an 80% statistical power. The final sample size in both Study 1 and Study 2 does

satisfy the sampling requirement. And, the *a priori* significance level was set at the conventional .05 level.

4.3.2 *Inferential Statistics*

To conduct hypothesis testing, or determine whether $H_0 = H_a$, the general philosophy of test statistic is given by (Field, 2009)

$$\text{test statistic} = \frac{\text{variance explained by model}}{\text{variance not explained by model}}$$

The calculated test statistic (e.g., t, F, χ^2) can then be used to compare with the critical value to determine the rejection of the null hypothesis. The p-value is also often calculated from the test statistic as the actual significance level and used to compare with the *a priori* significance level of α.

4.3.2.1 *t-test*

A t-test is conducted for comparing two group means. The default null hypothesis states no significant difference between the two means, while the alternative hypothesis suggests that significant difference does exist. There are two categories of t-test: independent sample and dependent (paired) sample.

For the independent sample t-test, different participants are used to provide responses in each condition and the group variable is known as a between-subject variable. For example, in this dissertation, both listener group and talker accent were

between-subject variables, where a listener or a talker could not be identified as both native and non-native English-speaking. The independent t-test is given as

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} \tag{4}$$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \tag{5}$$

where $\bar{X}_1, \bar{X}_2$ = group mean of group 1 and 2, respectively

s_1^2, s_2^2 = variance of group 1 and 2, respectively

n_1, n_2 = number of participants in group 1 and 2, respectively

s_p^2 = pooled variance

For the dependent or paired sample t-test, on the contrary, participants provided responses on all conditions in the variable, which is also known as within-subject variable. For example, the acoustic variables of background noise level and reverberation time in this dissertation were both within-subject variables. A paired t-test should be applied to compare the means calculated for any two levels in the acoustic variables. The paired t-test is given as

$$t = \frac{\bar{D} - \mu_D}{s_D/\sqrt{N}} \tag{6}$$

where $\bar{D}$ = mean difference between two groups

μ_D = expected mean, 0 if testing null hypothesis suggests group difference

$s_D/\sqrt{N}$ = standard error of the difference

Once the value of the t-test is computed, it can be used to compare against the critical value of the t-distribution determined by α and degree of freedom. Numerically, the p-value can also be calculated for direct comparison with α. If the calculated t value is greater than or equal to the critical value (or $p \leq \alpha$), the null hypothesis is rejected suggesting a significant difference between the two group means.

4.3.2.2 *F-test*

Besides comparing two group means, there are sets of parametric tests (e.g., regression and ANOVA) that examine the strength of the predictors (or also known as independent variables) in explaining variation observed in the dependent variable. These parametric models take the general form of

$$\text{Data} = \text{Model} + \text{Error} \tag{7}$$

The hypothesis testing therefore utilizes the F-ratio as a measure of the systematic variation to unsystematic variation (error or residual). The F-ratio is given as

$$F = \frac{SS_{model}/df_M}{SS_{error}/df_{error}} \tag{8}$$

$$SS_{model} = \sum n_k\left(\bar{x}_k - \bar{x}_{grand}\right)^2 \tag{9}$$

$$SS_{total} = \sum\left(x_i - \bar{x}_{grand}\right)^2 \tag{10}$$

$$SS_{error} = SS_{total} - SS_{model} \tag{11}$$

where df_M, df_{error} = degree of freedom in the model and residual, respectively

$\bar{x}_k$ = group mean for group k

n_k = number of participants in group k

$\bar{x}_{grand}$ = grand mean

x_i = observed data

Analogous to the t-test, once the F-ratio is computed it will be used to compare against the critical value identified by the degrees of freedom and the *a priori* significance level. Alternatively, the p-value can also be calculated for direct comparison with α. The null hypothesis is rejected when the F-ratio is greater than or equal to the critical value (or $p \leq \alpha$).

4.3.2.3 *Effect Size*

As seen in the previous sections, the use of significance testing (comparing p-value with α) in both t- and F-tests has restricted the outcome to be dichotomous. Although the magnitude of the calculated p-value provides some insights of strength, it has limitation in providing a direct measure as a probability metric. Two measures of effect size are utilized for this dissertation in the context of mean difference and variance explained, corresponding to t-test and F-test respectively.

Mean difference. Cohen's d (Cohen, 1977) is a measure of effect size to indicate the degree of separation between two independent distributions. It is expressed mathematically by

$$d = \frac{\bar{X}_1 - \bar{X}_2}{s_p} \tag{12}$$

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \quad (13)$$

where $\bar{X}_1, \bar{X}_2$ = group mean of group 1 and 2, respectively

s_1^2, s_2^2 = variance of group 1 and 2, respectively

n_1, n_2 = number of participants in group 1 and 2, respectively

s_p = pooled standard deviation

A variation of Cohen's d for repeated measures is given by

$$d = \frac{\bar{X}_1 - \bar{X}_2}{s_{diff}/\sqrt{N}} \quad (14)$$

where s_{diff} = standard deviation of $\bar{X}_1 - \bar{X}_2$

N = number of participants

Variance explained. Both η^2 and η_p^2 are used as measures of effect size for variance explained. They are calculated from the sample size using the following formulae.

$$\hat{\eta}^2 = \frac{SS_{effect}}{SS_{total}} \quad (15)$$

$$\hat{\eta}_p^2 = \frac{SS_{effect}}{SS_{effect} + SS_{error}} \quad (16)$$

Most recently, there are debates regarding the use of η^2 or η_p^2 as the better measure of effect size (Levine and Hullett, 2002; Richardson, 2011). Both measures are biased upward with the sum of squares calculated from the sample rather than from the population. From the mathematical expression above, η_p^2 is more biased than η^2 with a

smaller denominator if the model contains more than one factor. In addition, the η_p^2's are not additive in the factorial design, making it difficult to directly compare factors within the same omnibus model.

In this dissertation, the metric of η_p^2 was adopted as the measure of effect size for several reasons. First, the ratio for η_p^2 is analogous to the definition of F-ratio, hence conceptually more favorable in the philosophy of testing the strength of model prediction. Second, η_p^2 is calculated for the variance explained by the unique effect when controlling all other factors in the model, making it unaffected by the number of factors in the omnibus model. The comparison of η_p^2 of the same effect is hence possible across models with different number of factors. Third, the unbiased estimate of effect size is provided for η_p^2 by Judd *et al.* (2011). (The equation below is slightly revised to contain notations consistent with others in this chapter.)

$$\eta_p^2 = 1 - (1 - \hat{\eta}_p^2)\frac{df_{effect} + df_{error}}{df_{error}} \tag{17}$$

In the statistical analyses for this dissertation, the unbiased estimate of η_p^2 was used to indicate the effect size of all main effects and interactions, and the Cohen's d for pairwise or planned comparisons between two means. Specifically, Equation (12) was applied for comparisons of the between-subject variables (i.e., listener group and talker accent) and Equation (14) for within-subject variables (i.e., background noise level and reverberation time). Based on Cohen's (1992) suggestion, effect size can be categorized into small, medium and large by the following magnitude of Cohen's d and η^2 in Table 4.3.

Table 4.3 - Effect size values for small, medium, and large effects

	Effect Size		
	Small	**Medium**	**Large**
Cohen's d	0.2	0.5	0.8
η^2 and η_p^2	0.02	0.1	0.25
r	0.1	0.3	0.5

4.3.3 *Multivariate Analyses*

The various multivariate statistical analysis techniques used in this dissertation are discussed in this section.

4.3.3.1 *Correlation and Regression*

Correlation. Both bivariate correlation and partial correlation were adopted in this dissertation in examining the linear relation between two variables. For example, the NASA TLX subscales were correlated among each other (Chapter 5) and the two performance measures (Chapter 5 and 7) were also related, as assessed by a Pearson's product-moment correlation coefficient.

$$r = \frac{cov_{xy}}{s_x s_y} \tag{18}$$

$$cov_{xy} = \sum \frac{(x_i - \bar{x})(y_i - \bar{y})}{(N-1)} \tag{19}$$

The Pearson's correlation coefficient r is also a measure of effect size (Cohen, 1992). The values of r associated with different effects are indicated in Table 4.3.

A partial correlation was applied to relate the subjective dual-task performances and perceived performance (Chapter 7), while controlling for English proficiency level, to understand the unique variance in perceived performance explained by either performance measure.

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}} \tag{20}$$

where r_{12} = correlation between X_1 and X_2

r_{13} = correlation between X_1 and the controlling variable X_3

r_{23} = correlation between X_2 and the controlling variable X_3

Regression. This technique allows the examination of the linear relationship among multiple variables. A simple regression model was applied to examine the ability of standardized English proficiency score in predicting speech comprehension performance, which was averaged across all acoustic conditions. The mathematical expression of a simple regression model takes the form of

$$Y_i = b_0 + b_1 X_i + \varepsilon_i \tag{21}$$

where Y_i = *i*th observed score in the dependent variable

X_i = *i*th observed score in the independent variable

b_0 = intercept

b_1 = unstandardized regression coefficient of predictor X

ε_i = residual or error between the *i*th predicted and observed scores

Using the regression model, a set of test statistics can be calculated such as the coefficient of determination R^2 (and subsequently F-ratio) for assessing goodness of fit and effect size of the omnibus model, as well as *t* for individual predictors if multiple predictors exist in the model. In a simple regression with only one predictor, the Pearson's correlation coefficient equals to the square root of the coefficient of determination. They are given in the following equations.

$$R^2 = \frac{SS_{model}}{SS_{total}} \tag{22}$$

$$t = \frac{b_{observed}}{SE_b} \tag{23}$$

where SS_{model} = sum of squares of the model including all factors

SS_{total} = sum of squares total

$b_{observed}$ = unstandardized regression coefficient of specific predictor

SE_b = standard error of the mean of $b_{observed}$

An unbiased estimate of R^2, or R^2_{adj}, is defined as

$$R^2_{adj} = 1 - (1 - R^2)\frac{df_{model}}{df_{error}} \tag{24}$$

where df_{model} = degree of freedom in the omnibus model

df_{error} = degree of freedom of residual

4.3.3.2 *Reliability and Intraclass Correlation*

Reliability. All the scales utilized in this dissertation were adopted from existing surveys for their relevance and good internal consistency (also known as reliability), as quantified by Cronbach's α, in measuring the intended construct (Nunnally *et al.*, 1967). However, if a composite scale is formed by combining sets of the scales, it may not sustain the same internal consistency as each individual scale. For measuring English proficiency level in this dissertation, instead of using self-report surveys, a composite scale was created using three individual tests of listening span (Woodcock *et al.*, 2001b), oral comprehension (Woodcock *et al.*, 2001a), and bilingual verbal abilities (Muñoz-Sandoval *et al.*, 1998). The reliability of the composite scale should be therefore confirmed using Cronbach's α, which is given as

$$\alpha = \frac{N^2\overline{Cov}}{\sum s_{item}^2 + \sum Cov_{item}} \tag{25}$$

where N = number of items

$\overline{Cov}$ = averaged covariance between items

s_{item}^2 = individual item variance

Cov_{item} = individual item covariance

The magnitude of Cronbach's α suggests different degree of internal consistency, as shown in Table 4.4. As previously reported in Chapter 3, the Cronbach's α for the composite scale of English proficiency level achieved over 0.9 from both studies, suggesting excellent internal consistency.

Table 4.4 - Levels of Cronbach's α

Cronbach's α	**Internal Consistency**
$\alpha \geq 0.9$	Excellent
$0.7 \leq \alpha < 0.9$	Good
$0.6 \leq \alpha < 0.7$	Acceptable
$0.5 \leq \alpha < 06$	Poor
$\alpha < 0.5$	Unacceptable

Intraclass Correlation. This analysis was only applied to examine the consistency between two raters in measuring the speech rates of talkers using recorded sentences from the speech comprehension materials in both studies. The intraclass correlation coefficient (ICC) examines the correlation between two raters using the following equation.

$$ICC = \frac{s_b^2}{s_b^2 + s_w^2} \tag{26}$$

where s_b^2 = between rater variance

s_w^2 = within rater variance

The ICC measure is analogous to the Pearson's correlation coefficient, whereas ICC quantifies the linear relation between participants (e.g., Do raters always observe the same phenomenon?) and the Pearson's r quantifies such relation between factors (e.g., Does the change in one phenomenon affect another phenomenon?).

4.3.3.3 *Analysis of Variance and Covariance*

The analysis of variance (ANOVA) is a large family of variance analysis techniques, which were most frequently used in answering the research questions in this dissertation. Rather than comparing two group means in a t-test, ANOVA is capable of comparisons of multiple group means. Beginning from an omnibus model with calculated F-statistics, ANOVA is analogous to regression but with categorical or ordinal variables as predictors or independent variables. It also allows comparisons of multiple group means without the inflation of Type I error. Four variations of the analysis of variance technique were utilized and illustrated in Table 4.5 for their distinct characteristics.

Table 4.5 - Variations of analysis of variance depending on characteristics of the independent variable (IV) and dependent variable (DV)

	Only one DV	**More than one DV**
IVs contain categorical variables only	ANOVA	MANOVA
IVs contain both categorical and continuous variables	ANCOVA	MANCOVA

Univariate Analysis of Variance. As briefly discussed in Section 4.3.2.1 for t-test, there are two types of variables, between-subject and within-subject, based on the design of experiment. For between-subject variables, participants are only tested for one level of the categorical or ordinal variable. For within-subject variables, participants are measured repeatedly for all levels of the same factor. Depending on the type of variables in the model, the univariate ANOVA is further divided into three types: factorial (between-

subject variables only), repeated-design (within-subject variables only), and mixed-design (both between- and within-subject variables).

Following the philosophy of examining variance explained as shown in Section 4.3.2.2 for F-test, the test statistics reported for the omnibus ANOVA models in this dissertation included F-ratio, degrees of freedom for both model and error, effect size in η_p^2, and p-value. An additional assumption of sphericity is required for models containing within-subject variables. It states that the variance of the differences between conditions in the within-subject variable should be equal. A Mauchly's test of sphericity is always calculated for ANOVA models containing within-subject variables in SPSS. It tests the null hypothesis that the variance of the differences is the same. If the Mauchly's test is found statistically significant, the variance of the difference cannot be assumed equal. In this case, the calculated F-ratio should be corrected using the Greenhouse-Geisser on the degree of freedom (Field, 2009), which may subsequently change the p-value. Neither sum of squares nor effect size is affected by the violation of sphericity. Throughout all ANOVA models in this dissertation, the Greenhouse-Geisser correction did not substantially change the dichotomous outcome from the calculated p-values. For the purpose of avoiding confusion in the reported degrees of freedom, the results assuming sphericity are always reported even though violation existed.

Multivariate Analysis of Variance. As seen in Table 4.5, the distinction between univariate and multivariate ANOVAs is in the number of DVs in the model. Essentially, MANOVA not only calculates the variance explained, but it also takes into account the relation between the DVs. Field (2009) provides a clear conceptual comparison of the different components between ANOVA and MANOVA (Chapter 16.4.2), as illustrated in

Table 4.6. Instead of using single numbers, MANOVA replaces the components with matrices in the test statistic calculations.

Table 4.6 - Conceptual comparisons of variance partitioning between ANOVA and MANOVA models

	ANOVA	**MANOVA**
Total variance	SS_{total}	Total sum of squares and cross-products matrix (Total SSCP, *T*)
Proportion of variance explained by model	SS_{model}	Hypothesis sum of squares and cross-products matrix (Hypothesis SSCP, *H*)
Residual	SS_{error}	Error sum of squares and cross-product matrix (Error SSCP, *E*)

The test statistic for MANOVA comparing the systematic variation over the unsystematic variation is then given as

$$HE^{-1} \tag{27}$$

from which, a set of eigenvectors can be extracted to construct discriminant functions that links the DVs in the form of a multiple linear regression to predict a variate score, where the eigenvalues are the coefficients of determination for the DVs. By calculating the variate scores for each participant, the HE^{-1} matrix can be reduced into a diagonal matrix of $HE^{-1}_{variate}$.

$$HE^{-1}_{variate} = \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_i \end{bmatrix} \tag{28}$$

The test statistics for MANOVA include Pillai's trace, Hotelling's T^2, Wilks' lambda, and Roy's largest root. In this dissertation, the Pillai's trace is reported for all MANOVA models. It is given as

$$V = \sum_{i=1}^{s} \frac{\lambda_i}{1 + \lambda_i} \tag{29}$$

where λ_i = eigenvalue in the $HE^{-1}_{variate}$ matrix

s = number of eigenvalues in the $HE^{-1}_{variate}$ matrix

The Pillai's trace V approximately follows an F-distribution, from which the conventional test statistics (e.g., F-ratio and significance level) can be calculated. Follow-up tests to a significant MANCOVA are recommended by Field (2009) in two variations, either separate univariate ANOVAs or discriminant analysis.

In this dissertation, the multivariate model was most relevant to the dual-task scheme in measuring two performance DVs of speech comprehension and APR dot-tracing tasks. A multivariate model was first fitted to the data involving the two performance measures. Since speech comprehension was a more relevant performance measure than dot-tracing in most research questions, the effect of individual IVs (e.g., background noise level, reverberation time, and English proficiency level) on speech comprehension performance was preferred over discrimination between the two performance measures. Therefore, separate ANOVAs were conducted as follow-up tests to the significant MANOVAs.

Analysis of Covariance. The ANOVA and MANOVA models are not limited to only containing categorical IVs. The analysis of covariance commence when a

confounding factor is identified and required for control in the statistical models, turning them into ANCOVA and MANCOVA. An example from this dissertation was the standardized English proficiency score, which was a significant and strong confounder to the speech comprehension performance under assorted acoustic conditions. It can also be regarded as a hybrid model of ANOVA and regression, where continuous covariate can be represented by a regression line for each condition in the categorical IVs. If the DV is plotted against the covariate, as seen in the conceptual illustration in Figure 4.1, the main effects of the categorical IV are the relative position of the regression lines across conditions. Their interaction is suggested by the different slope of the regression lines under different conditions.

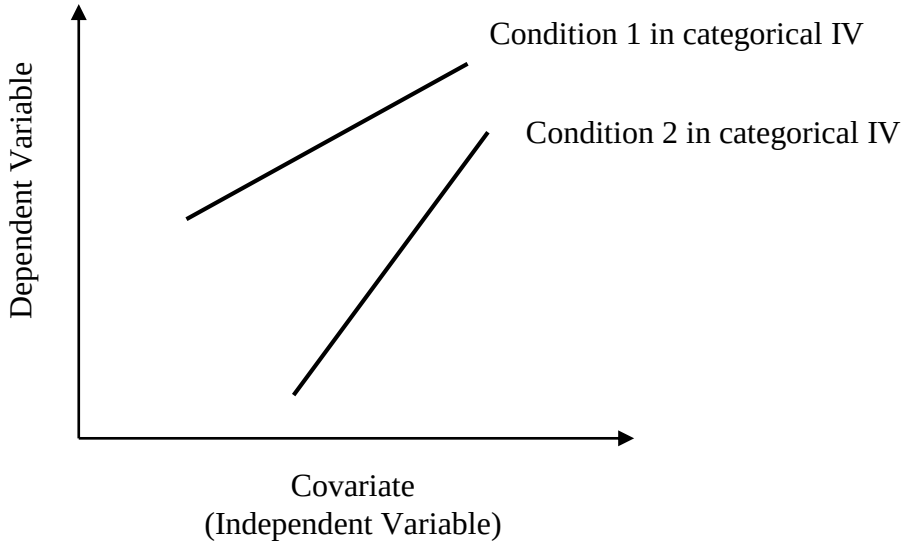


Figure 4.1 - Conceptual illustration of analysis of covariance

4.3.3.4 *Planned Comparison and Post Hoc Analyses*

As mentioned in the previous section, the family of ANOVA models have the ability to conduct comparisons of multiple group means through either planned comparison or post hoc analysis, while maintaining the *a priori* significance level for Type I error. The difference between planned comparison and post hoc analysis, as Field (2009) pointed out, is whether a hypothesis exists on the relation among the multiple group means.

Planned Comparison. It is also known as planned contrasts, for which a hypothesis exists on the relation among multiple group means. A set of contrast codes can be applied to the various levels to test the specific hypothesis. In order to maintain the *a priori* significance level, the number of comparisons should not exceed the degree of freedom of the categorical IV.

In this dissertation, both background noise level and reverberation time were hypothesized to correlate negatively with speech comprehension performance. In order to provide practical acoustic design guidelines, the research question sought to identify the level beginning at which a significant speech comprehension deficit occurred. Hence, the first level in both acoustic variables (i.e., RC-30 for background noise level and 0.4 second reverberation time) was used as the reference level for multiple comparisons against the higher levels individually.

It should be noted that the contrast coding applied for the planned comparison for the acoustic variables were non-orthogonal. In the SPSS (version 22) output, if non-orthogonal contrast codes are used in a mixed-design ANOVA, the sum of squares effect and error (both in Type III) for the between-subject effects were coincidentally reduced

by a factor of 15. Although this does not affect the results of F-ratio, effect size, or partial η_p^2, they should be corrected for deriving the total sum of squares of the corrected omnibus model in the manual calculation of η^2.

Post hoc. Without a proper hypothesis of the relation among various group means, post hoc analysis using pairwise comparisons is possible by applying corrections. Both Bonferroni's and Tukey's honestly significant difference (Tukey's HSD) are conservative corrections to sufficiently control the Type I error rate. Field (2009) pointed out that Bonferroni's correction has more statistical power for a small number of comparisons, whereas Tukey's HSD is more powerful for a large number of comparisons. However, from the experience working with the data in this dissertation, the Bonferroni's seemed to over correct the significance level more often than the Tukey's HSD. Hence, the Tukey's HSD was applied for post hoc analysis on all between-subject IVs. For within-subject IVs, only Bonferroni's correction was available for pairwise comparisons.

4.4 Summary and Discussion

This chapter examined the fundamentals of statistics and the related analysis techniques utilized in this dissertation. The procedures and decisions were documented in greatest details when conducting the specific statistical analysis relevant to answering the research questions. An issue was identified in the SPSS output for ANOVA using non-orthogonal contrast codes in the planned comparisons analysis.

Chapter 5 – Study 1: Effects of Room Acoustics on Native American English Speech Comprehension

5.1 Introduction

Study 1 focused on investigating effects of room acoustics on speech comprehension by native and non-native English-speaking listeners while the speech materials were produced by only native American English-speaking talkers. This chapter discusses the experimental procedures and findings from data collected from these listeners.

5.2 Speech Material Recording

Recording of the speech materials for Study 1 was conducted in an anechoic chamber. Five native English-speaking talkers, one male and four females, were recruited as volunteers to record the speech comprehension materials described in Chapter 3. They were instructed to read the audio scripts at their normal conversational speed. The anechoic audio recordings were first edited in Audacity before being convolved with each BRIR in Matlab for presentation in the speech comprehension test program. No special effects (spectral or temporal) were added in the anechoic recordings during post-processing in Audacity.

Due to the large amount of audio recording, four female talkers were recruited for Study 1 and assigned to record for different parts in the four speech comprehension tasks. The recording assignment, as seen in Table 5.1, for the female talkers was done so that their voice appearance remained consistent across the final 15 sets of test materials. Furthermore, the speech comprehension test program presented test items with

alternating gender voice within each set of test. Although listeners experienced different voices during testing within each test set (i.e., male vs. females and different female talkers for different parts), the effect of varying speech rate from the talkers was counterbalanced across the 15 test sets.

To calculate the speech rate in syllables per second, two research assistants who were native American English speakers counted the number of syllables from the audio scripts and manually measured the speech duration of the corresponding audio recording in Audacity. At least 5 minutes of audio recordings were sampled for each talker. The speech rate of each talker is reported in Table 5.1. The two raters highly agreed with each other on the calculated speech rate, with an intraclass correlation coefficient (ICC) of 0.992.

Table 5.1 - Talker role assignment and speech rate

Native English-speaking Talker	Recording Assignment of the Speech Comprehension Materials	Speech Rate [Syllables per Second]	
		Mean	95% CI
Male	All four tasks	5.3	[5.1, 5.4]
Female 1	Task 1 and 2	3.4	[3.3, 3.5]
Female 2	Task 2	3.8	[3.7, 4.2]
Female 3	Task 3	4.8	[4.3, 5.3]
Female 4	Task 4	5.0	[4.7, 5.3]

5.3 Listener Participants

Two groups of total 58 listener participants, both native and non-native English speakers, were recruited on the University of Nebraska at Omaha campus. As previously

mentioned in Chapter 3, they were grouped by the first language learned in the self-report LEAP-Q. The native language profile of all participants in Study 1 is included in Appendix D. It was later found that two listeners (one from each listener group) were unable to complete the dual tasks simultaneously during the speech comprehension experiment. They were hence removed from data analysis. One participant self-identified as a native Arabic speaker (non-native English-speaking) but scored highly on the English proficiency tests, within the one standard deviation below the mean calculated from the native listeners. Furthermore, this non-native listener was later identified as an outlier, with much better speech comprehension performance among other non-native listeners. Although including this outlier in the native listener group did not substantially change the conclusions, the listener was only included in the reported analyses where the statistical models did not distinguish difference between listener groups.

The final set of participants comprised of a total of 56 participants, with 27 native English-speaking listeners (13 female) and 29 non-native listeners (13 female). The average age for the native English-speaking listener group was 23.7 years (SD = 5.8 years) and for the non-native group 26.5 years (SD = 5.2 years). Speech comprehension performance was not found to differ significantly between male and female; and it was not significantly predicted by age either.

Each listener participant was screened and tested according to the procedure outlined in Chapter 3. All listeners participated in the study have normal hearing. The non-native English-speaking listeners reported a variety of native languages from the language experience section in the LEAP-Q, as shown in Appendix D. (It should be noted that in the analysis for Chapter 7, the subgroup of native Chinese-speaking listeners was

separated from the non-native listeners from Study 1.) Besides two non-native listeners with extensive residency of 20 and 25 years, the average length of immersion in the English-spoken community is 23.6 months (range = 1-90 months). In addition to the self-report language experiences, all participants were individually given three English proficiency tests, involving listening span, oral comprehension, and English verbal skills. The composite scale of English proficiency level was highly reliable in Study 1, resulting in a Cronbach's alpha of 0.94.

5.4 Results

5.4.1 *English Proficiency Level*

The non-native English-speaking listeners who participated in this study were mainly foreign students attending degree programs at the University of Nebraska. A majority of these participants had taken English proficiency tests, such as the Test of English as Foreign Language (TOEFL), to gain entry to academic programs and had been living in an English dominant country for an extended period of time. The results of the composite English proficiency tests showed that the non-native participants as a group scored significantly lower than the native English-speaking participants, $t(54) = 14.36$, $p < .001$. However, as shown in Figure 5.1, there was no clear gap of the English proficiency levels between the native and non-native listeners, suggesting that the sampled non-native listeners were mostly at least moderately proficient in English. When averaged across acoustic conditions, English proficiency level significantly and strongly predicted speech comprehension performance, $b = 6.60$, $t(54) = 8.21$, $p < .001$. English

proficiency level also explained a significant proportion (55%) of the variance in speech comprehension performance, R^2_{adj}= 0.55, F(1,54) = 67.37, p < .001.

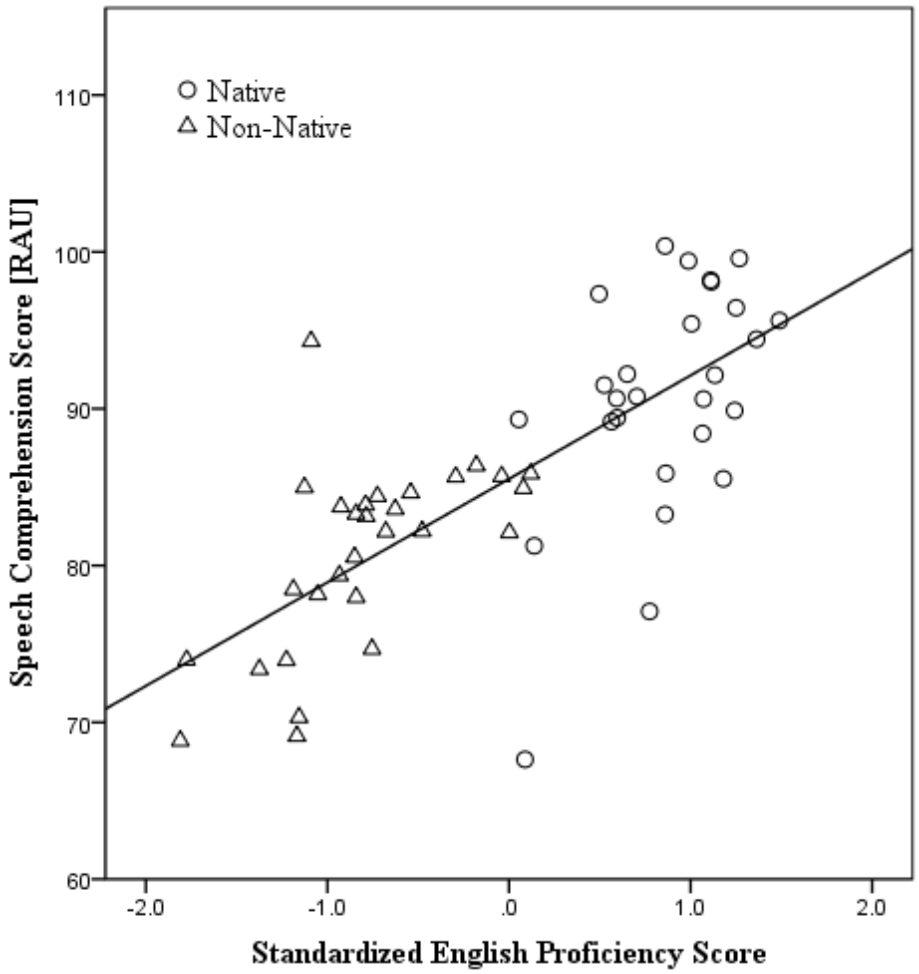


Figure 5.1 - Speech comprehension score, averaged across 15 acoustic conditions, as a function of English proficiency level for both native and non-native English-speaking listeners

.

The significant and strong linear relation provided support that English proficiency was indeed a strong confounding factor contributing to the bias in room acoustic effects on speech comprehension performance. Therefore, English proficiency

level should be entered in the statistical model as a covariate to control for its confounding effects. The investigation of effects of room acoustics on speech comprehension should look beyond listeners' English proficiency level.

5.4.2 *Objective Performance of Speech Comprehension*

5.4.2.1 *Controlling for English Proficiency*

A mixed-design multivariate analysis of covariance (MANCOVA) was applied to examine the room acoustic effects on the performances of speech comprehension and APR dot-tracing tasks together while controlling for English proficiency level. Using Pillai's trace, there was only one significant main effect for BNL, $F(4,51) = 23.85$, $\eta_p^2 = 0.63$, $p < .001$ and one significant interaction between BNL X English proficiency level, $F(4,51) = 4.38$, $\eta_p^2 = 0.20$, $p = .004$ on speech comprehension and dot-tracing performances. English proficiency was still a significant strong predictor of performances under the dual-task scheme, $F(2,53) = 33.21$, $\eta_p^2 = 0.54$, $p < .001$.

As follow-ups to the MANCOVA, separate mixed-design analysis of covariance (ANCOVA) were performed for the output performance measures as the single dependent variables. Prior to analysis, the assumptions of sphericity were confirmed for speech comprehension scores in RAUs by the non-significant Mauchly's W for BNL and RT. However, such assumptions were violated for the APR dot-tracing performance measured as RPM for both BNL ($W = 0.89$, $p = .047$) and RT ($W = 0.50$, $p < .001$) with significant Mauchly's W. The Greenhouse-Geisser corrections ($\varepsilon = 0.90$ for BNL, $\varepsilon = 0.78$ for RT) were checked and suggested no substantial change in the outcome from the calculated p-value than when sphericity was assumed. Therefore, all results were reported

under the assumption of equal sphericity to retain consistent degrees of freedom (see Chapter 4 for more discussion).

For speech comprehension tasks, English proficiency level remained as a significant and strong predictor, $F(1,54) = 67.37$, $\eta_p^2 = 0.55$, $p < .001$. There was a significant main effect for BNL, $F(2,108) = 36.26$, $\eta_p^2 = 0.39$, $p < .001$ and for RT, $F(4,216) = 3.73$, $\eta_p^2 = 0.05$, $p = .006$. It was previously hypothesized that speech comprehension performance decreases as BNL or RT increases. Therefore, planned comparisons were deemed appropriate using the lowest condition (RC-30 for BNL and 0.4 seconds for RT) as the reference level to identify a higher level, at which significant performance deficit was observed. As shown in Figure 5.2, The results showed that, while controlling for English proficiency level, participants scored significantly higher in the RC-30 BNL condition than in RC-50 ($d = 1.18$, $p < .001$) but not in RC-40 ($d = 0.23$, $p = .093$). For RT, as seen in Figure 5.3, participants scored significantly higher in the 0.4 second scenario than in the 0.8 second ($d = 0.38$, $p = .007$) and in the 1.2 second ($d = 0.42$, $p = .003$) scenarios; but not in the 0.6 second ($d = 0.12$, $p = .36$) or 1.0 second ($d = 0.13$, $p = .32$) scenario. There was a significant interaction between BNL X English proficiency level, $F(2, 108) = 5.72$, $\eta_p^2 = 0.08$, $p = .004$. The performance deficit in speech comprehension with increasing BNL, specifically from RC-30 to RC-50 ($p < .004$), was significantly greater for participants with lower English proficiency level (see Figure 5.4 and Table 5.2).

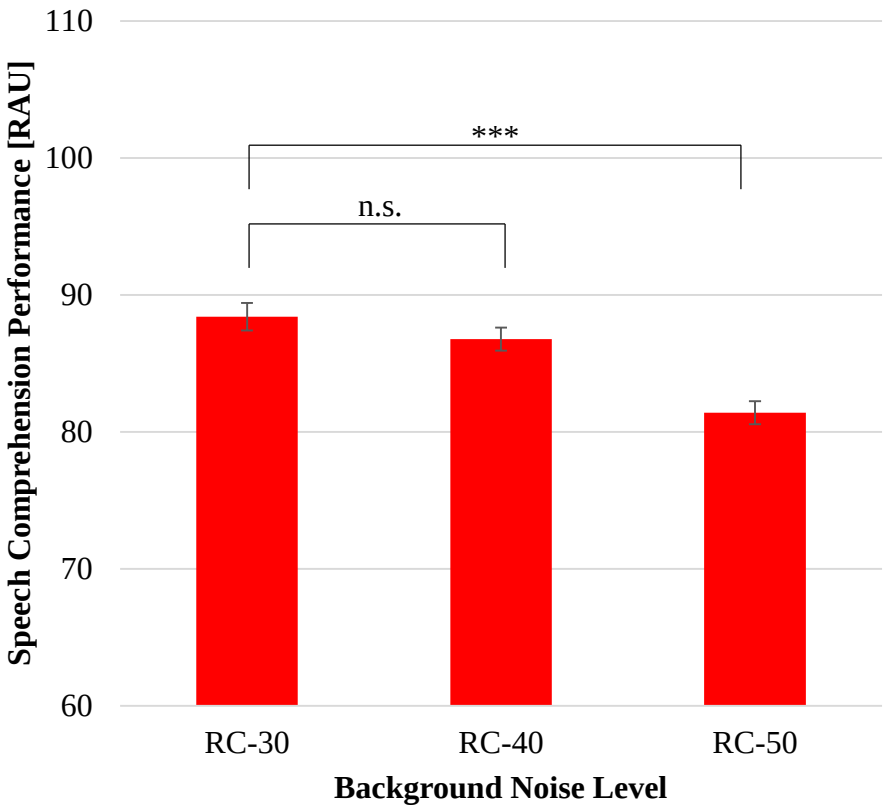


Figure 5.2 - Marginal means of speech comprehension performance, averaged across all RT scenarios for each BNL condition, evaluated at standardized English proficiency score at 0. Error bar indicates 1 standard error. Statistical significance level is shown for each pair tested in planned comparison[1].

[1] * p < .05; ** p < .01; *** p < .001; n.s. for non-significant, p > .05. Same in all following graphs.

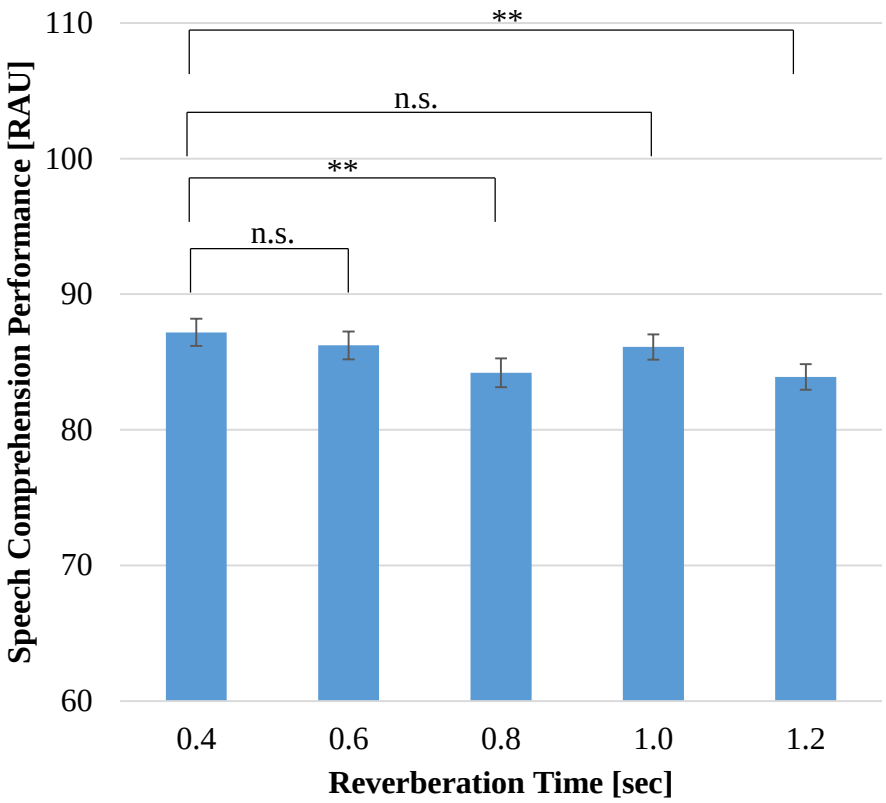


Figure 5.3 - Marginal means of speech comprehension performance, averaged across all BNL for each RT scenario, evaluated at standardized English proficiency score at 0. Error bar indicates 1 standard error.

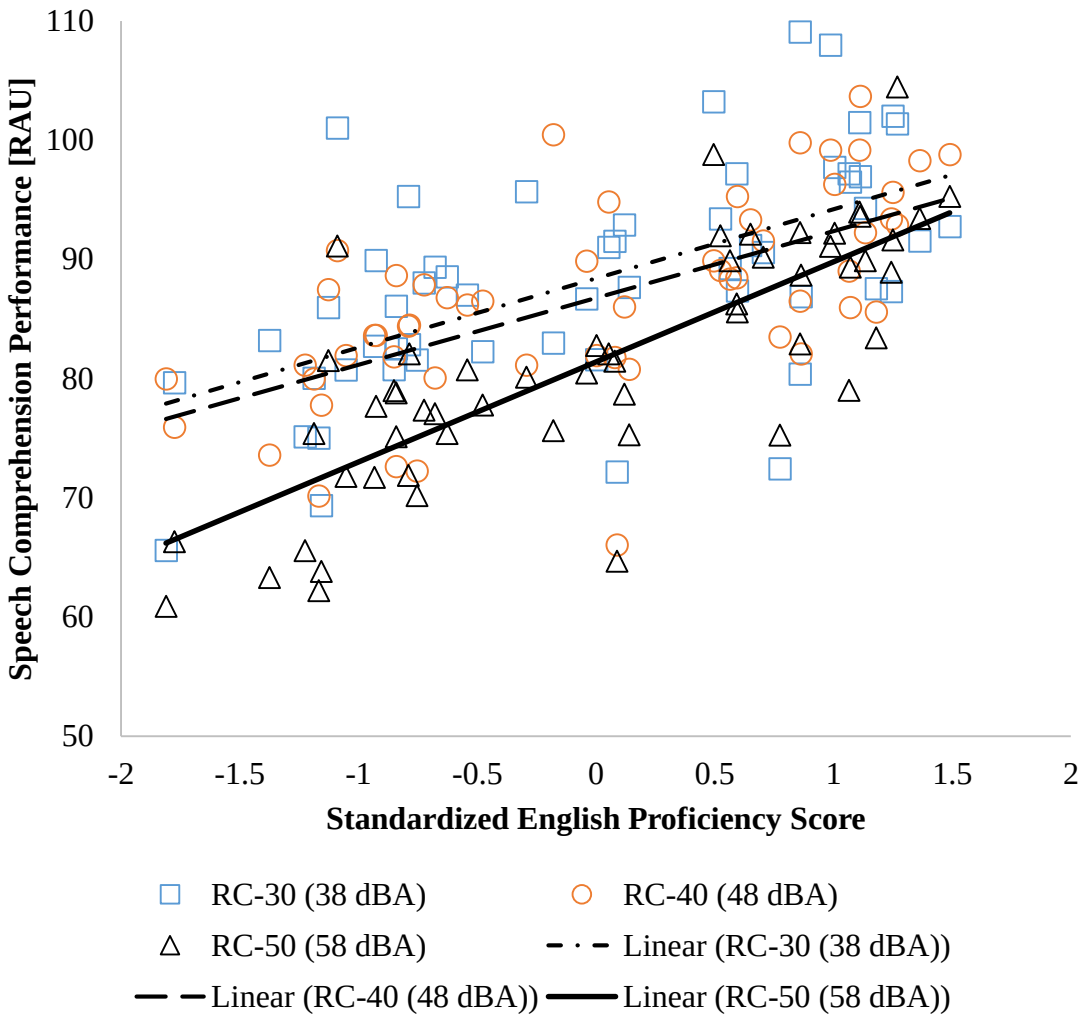


Figure 5.4 - Relation of speech comprehension performance and English proficiency level under three BNL conditions

Table 5.2 - Summary of linear regression lines fitted to the relation between speech comprehension performance and English proficiency level for each BNL

BNL	R^2_{adj}	b	SE b	β
RC-30	0.34	5.80	1.08	0.59***
RC-40	0.41	5.60	0.90	0.65***
RC-50	0.61	8.40	0.90	0.79***

Note: *** p <.001

The follow-up ANCOVA for the secondary competing APR task did not reveal any significant main effect for BNL, RT or English proficiency level ($p > .05$ for all main effects). Planned comparison using polynomial contrasts showed significant quadratic trend for BNL ($p = .040$) and the linear interaction of BNL X RT ($p = .037$). Participants achieved slightly better performance of the APR task under the RC-40 condition by an extra 1.0 RPM than the two other BNL conditions. The two performance measures were correlated using Pearson's correlation for each acoustic condition, as shown in Table 5.3. Both measures were positively correlated across all acoustic conditions, suggesting that the APR dot-tracing performance increased with increasing performance in the speech comprehension tasks.

Table 5.3 – Pearson's correlation coefficient (two-tailed) between performance measures of speech comprehension and adaptive pursuit rotor (dot-tracing) for each acoustic condition

Pearson's Correlation Coefficient (N = 56 for each acoustic condition)					
	Reverberation Time Scenario				
Background Noise Level	**0.4 sec**	**0.6 sec**	**0.8 sec**	**1.0 sec**	**1.2 sec**
RC-30	0.15	0.28*	0.28*	0.33*	0.18
RC-40	0.16	0.18	0.28*	0.17	0.25
RC-50	0.10	0.35**	0.17	0.16	0.35**

Note: *p < .05, **p < .01

5.4.2.2 *Speech Comprehension Performance between Native and Non-native English-speaking Listeners*

In the previous statistical model, listener group was not included as a between-subject variable because such property was more accurately described by English proficiency level. To examine the acoustic effects between listener groups, the same ANCOVA models on speech comprehension performance were conducted separately for the native and non-native English-speaking listener groups. The effect sizes of BNL and RT were compared between the two listener groups. As mentioned in Chapter 4, effect size is utilized to quantify the strength of the independent variable (IV) in affecting the dependent variable (DV). Both η^2 and η_p^2 are reported in Table 5.4 for significant main effects and interaction in the factorial ANCOVA to describe the proportion of variance explained in speech comprehension performance, either in the omnibus model (i.e., η^2) or while controlling for all other IVs (i.e., η_p^2).

As shown in Table 5.4, English proficiency level significantly and strongly predicted the speech comprehension performance of both native and non-native listeners. Although statistically non-significant, the effect size of BNL in the native listener group is similar to that in the non-native listener group, sharing a moderate effect on speech comprehension performance (see Chapter 4 on magnitude of effect size). Interestingly, the significant main effect for RT was only found among non-native listeners, and its moderate effect size was similar to that of BNL in this listener group. A two-way interaction between RT X English proficiency was found to be significant for the non-native listeners. For native English-speaking listeners, the negative effect of RT is much smaller than that of BNL. Taken altogether, listeners' baseline English proficiency level

greatly influenced their performance on the speech comprehension tasks. When English proficiency level was controlled, the negative impact of BNL was similar for both listener groups though slightly weaker for the native listeners. However, the effect of RT on speech comprehension differed substantially between native and non-native listeners. While native listeners did not seem to be affected by RT, its impact on non-native listeners was almost as equivalently negative as BNL.

Table 5.4 - Effect size comparisons of the significant main effects and interaction in the factorial ANCOVA of speech comprehension performance between native and non-native English-speaking listener groups

	Native Listeners (N = 26)			**Non-Native Listeners (N = 29)**		
	p-value	η^2	η_p^2	**p-value**	η^2	η_p^2
English Proficiency Level	.006	0.12	0.28	.001	0.1	0.39
BNL	.053	0.01	0.12	.005	0.03	0.20
RT	.62	0.004	0.03	.007	0.04	0.18
RT X English Proficiency	.68	0.004	0.02	.01	0.02	0.12

5.4.3 *Subjective Perception of Task Workload*

5.4.3.1 *NASA TLX Subscales*

The NASA Task Load Index (TLX) of workload assessment questionnaire was given to participants as a measure of subjective perception. As previously mentioned, only the individual scale rating was administered, without the supplementary subscale rank order through pairwise comparisons (Hart, 2006). Among the 90 distributions in the NASA TLX ratings (6 subscales X 15 acoustic conditions), 53 of them resulted in non-

significant Shapiro-Wilk tests ($p > .05$). It suggested that a majority of the NASA TLX distributions under various acoustic conditions conformed to normality. As a result, a mixed-design analysis of variance (ANOVA) was used to test the effects of BNL, RT, and listener group on the individual subscales of workload assessment from the NASA TLX. The assumption of sphericity has either been confirmed or checked for the Greenhouse-Geisser correction when interpreting results. A post hoc analysis of pairwise comparison using the Bonferroni's correction had also been applied. All six subscales in NASA TLX were shown in Figure 5.5 and Figure 5.6, with discussions in the following paragraphs.

Mental Demand. There were significant main effects for BNL [$F(2, 106) = 11.97$, $\eta_p^2 = 0.17$, $p < .001$] and listener group [$F(1,53) = 5.39$, $\eta_p^2 = 0.08$, $p = .024$], as well as a two-way interaction for BNL X listener group [$F(2,106) = 5.03$, $\eta_p^2 = 0.07$, $p = .008$]. Non-native listeners reported higher mental demand than native listeners. Pairwise comparison revealed that the demand for mental activity was significantly higher under the BNL condition of RC-50 than those under RC-30 ($d = 0.57$, $p < .001$) and RC-40 ($d = 0.47$, $p = .003$). The increase in mental demand under the RC-50 BNL condition was greater for non-native English-speaking listeners.

Physical Demand. There were significant main effects for BNL [$F(2, 106) = 7.45$, $\eta_p^2 = 0.11$, $p = .001$] and listener group [$F(1, 53) = 26.26$, $\eta_p^2 = 0.32$, $p < .001$]. Similar to mental demand, the demand for physical activity was significantly higher for BNL of RC-50 than the two other lower levels ($d = 0.46$, $p = .004$ for RC-30; and $d = 0.37$,$p = .026$ for RC-40). Non-native listeners reported higher physical demand than native listeners.

Temporal Demand. There were significant main effects for BNL [F(2,106) = 3.87, η_p^2 = 0.05, p = .024] and listener group [F(1, 53) = 15.91, η_p^2 = 0.22, p < .001]. The interaction of BNL X listener group was found to be significant [F(2,106) = 5.37, η_p^2 = 0.08, p = .006]. Non-native listeners again reported more severe time pressure than native listeners. All listener participants experienced significantly stronger time pressure under the highest BNL of RC-50 than under RC-30 (d = 0.40, p = .013). Again, the increase in temporal demand of the tasks with increasing BNL was rated greater by non-native listeners.

Effort. Significant main effects were found for BNL [F(2, 106) = 17.11, η_p^2 = 0.22, p < .001] and listener group [F(1, 53) = 6.23, η_p^2 = 0.09, p = .016], as well as one significant interaction for BNL X listener group [F(2, 106) = 8.31, η_p^2 = 0.12, p < .001]. Specifically, participants recognized having to work harder to accomplish the simultaneous tasks with increasing BNL (d = 0.69, p < .001 for RC-30 vs. RC-50; p = .004 for RC-40 vs. RC-50; and d = 0.32, p = .06 for RC-30 vs. RC-40). Such increase in effort was again more pronounced among non-native listeners. Non-native listeners reported spending more effort than native listeners in completing the tasks.

Frustration. The significant main effects and interaction and their respective effect size were similar to those of the subscale of effort, for BNL [F(2, 106) = 17.11, η_p^2 = 0.23, p < .001] and listener group [F(1, 53) = 10.47, η_p^2 = 0.15, p = .002], and BNL X listener group [F(2, 106) = 5.09, η_p^2 = 0.07, p = .008]. Non-native listeners reported feeling more frustrated than native listeners in completing the tasks. For the BNL conditions, significant increase in frustration was observed for RC-30 versus RC-50 (d = 0.72, p < .001), RC-40 versus RC-50 (d = 0.58, p < .001). The increase in frustration was

even greater for non-native listeners than for native listeners when BNL increased from RC-30 to RC-40 and to RC-50.

Perceived Performance. Participants were also asked to provide subjective rating of how successful they felt in accomplishing the simultaneous tasks under each acoustic condition. There were significant main effects for BNL [F(2, 106) = 9.65, η_p^2 = 0.14, p < .001] and RT [F(4, 212) = 2.95, η_p^2 = 0.04, p = .021], as well as one interaction between BNL X listener group [F(2, 106) = 3.34, η_p^2 = 0.04, p = .039]. Surprisingly, native and non-native listeners' perception of performance did not differ significantly (p = .50), although its interaction with BNL was significant. This was likely due to the fact that non-native listeners perceived to have performed better than native listeners under the RC-30 condition. Pairwise comparisons revealed that listeners perceived significantly worse performance on the simultaneous tasks under RC-50 than the two lower BNLs for RC-30 (d = 0.51, p = .001) and for RC-40 (d = 0.37, p = .025). In addition, they also felt performing significantly worse under RT of 1.2 seconds than under 0.4 second (d = 0.45, p = .017). The degradation in perceived performance was particularly greater for non-native listeners with increasing BNL from RC-30 to RC-50.

In summary, non-native listeners provided higher ratings than native listeners on all NASA TLX subscales except perceived performance under the RC-30 BNL condition. Most of these attributes of subjective perception on task workload assessment were not sensitive to the change in RT, as seen in Figure 5.6. Listeners only perceived their task performance to decrease when increasing RT from 0.4 to 1.2 seconds (p = .002). However, the effect of BNL on subjective perception was much more pronounced. The degradation in subjective perception was significant when increasing BNL from RC-30 to

RC-50, for some subscales even between RC-30 and RC-40 (i.e., temporal demand and effort). The interaction between BNL and listener group was also found significant in all subscales except physical demand, as plotted in Figure 5.5.

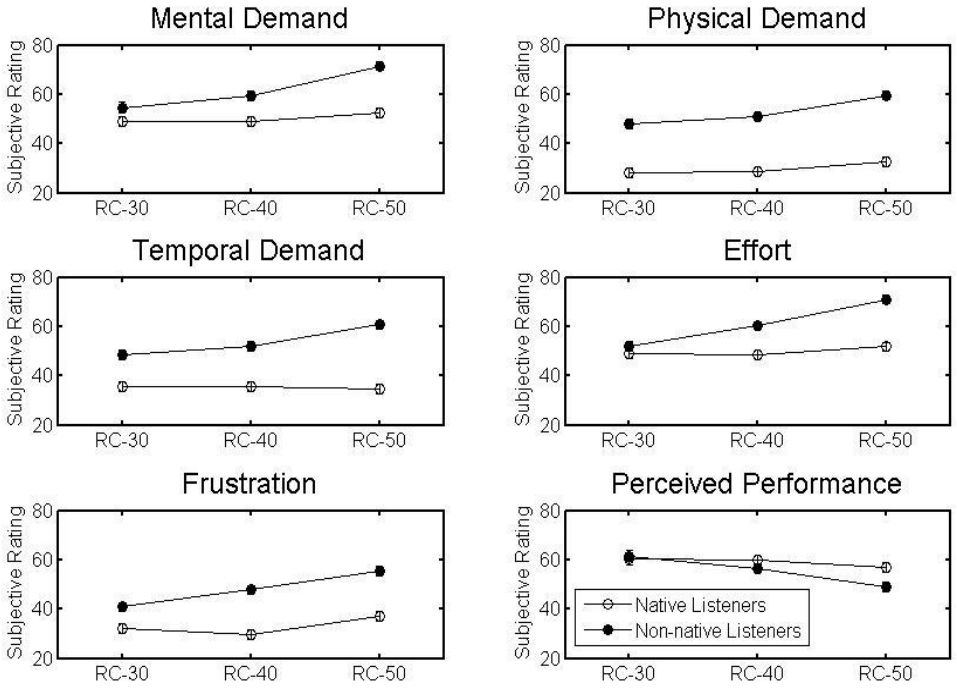


Figure 5.5 - Marginal means of NASA Task Load Index ratings of the dual-tasks in six subscales versus BNL for native (empty circle) and non-native (solid circle) listeners. Error bar indicates one standard error.

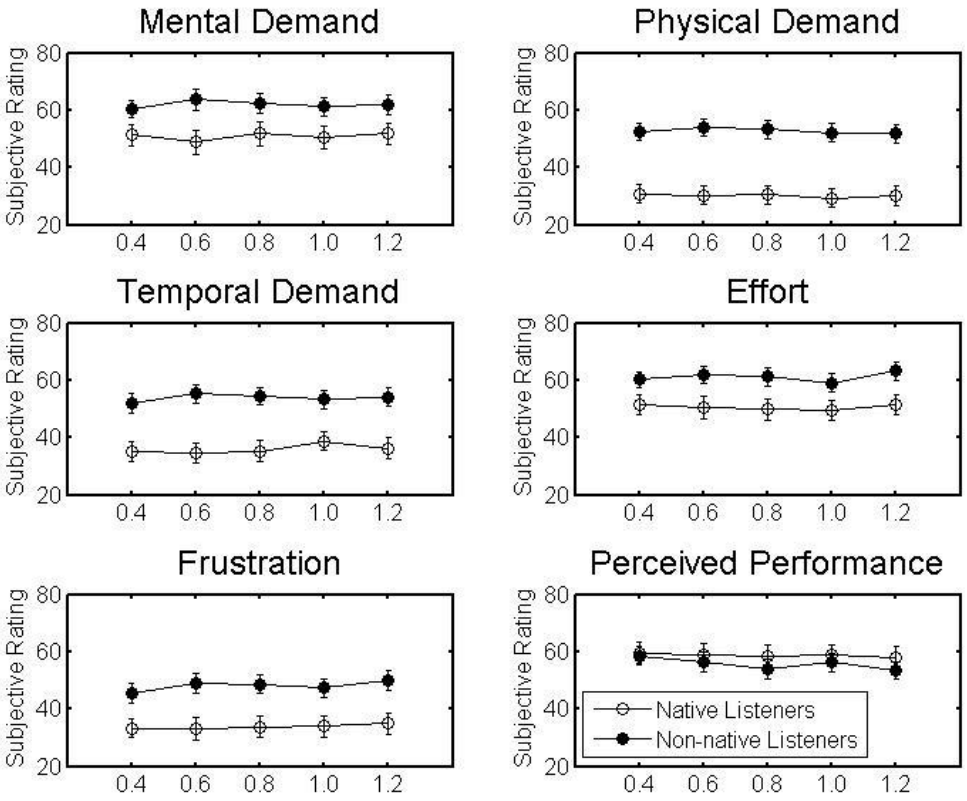


Figure 5.6- Marginal means of NASA Task Load Index ratings of the dual-tasks in six subscales versus RT for native (empty circle) and non-native (solid circle) listeners. Error bar indicates one standard error.

5.4.3.2 *Relating Subjective Perception with Objective Performance under Acoustics*

In order to relate subjective perception and objective performance, a partial correlation was computed between speech comprehension and perceived performance from NASA TLX, holding constant the standardized English proficiency score. Prior to the correlation analysis, the subscale of perceived performance was specifically checked for normality since it would be related to objective performance in RAU. Among the 15 distributions for the perceived performance rating, only three (BNL-RT combinations of RC-30 and 0.4 second, RC-30 and 0.6 second, and RC-50 and 1.2 seconds) were found statistically significant violating the normal distribution assumption. Since the majority of the perceived performance still conformed to normality, no transformation was needed and the raw score on the perceived performance rating was entered into the partial correlation analysis.

The partial correlation coefficient suggested that, while controlling for English proficiency level, objective performance and subjective perception of speech comprehension under assorted acoustic conditions were positively related, $r(837) = 0.27$, $p < .001$. Correlation between RPM from the APR task and perceived performance was not found, though, $r(840) = 0.024$, $p = 0.49$. When rating the perceived performance scale in the NASA TLX, listeners based heavily on their perception of performance from the speech comprehension task rather than the APR task.

Another mixed-design ANCOVA was performed on perceived performance to examine the acoustic effects, replacing the listener group with standardized English proficiency score as the control variable. Results show that BNL and RT have similar effects on perceived performance as they do on the objective performance of speech

comprehension. Significant main effects were found for both BNL [F(2, 108) = 10.44, η_p^2 = 0.15, p < .001] and RT [F(4, 216) = 2.70, η_p^2 = 0.03, p = .032]. English proficiency level was marginally significant [F(1, 54) = 3.63, η_p^2 = 0.05, p = .062]. The interaction between BNL X English proficiency was found significant [F(2, 108) = 3.94, η_p^2 = 0.05, p = .022]. The main effects of BNL and RT on perceived performance are plotted in Figure 5.7 and Figure 5.8, respectively.

In order to identify the level of significant degradation in perceived performance on the speech comprehension tasks, similar planned comparison was conducted using the lowest level as the reference level which yielded the highest perceived performance rating. For BNL, the perceived performance was rated significantly higher under RC-30 than RC-40 (d = 0.27, p = .048) and RC-50 (d = 0.52, p < .001). Specifically, the degradation in perceived performance worsened for those with lower English proficiency level when BNL increased from RC-30 to RC-50 (p = .019). For RT, listeners felt that their performance was significantly better under the RT scenarios of 0.4 second than under 0.8 second (d = 0.31, p = .048) and 1.2 seconds (d = 0.45, p < .001), respectively.

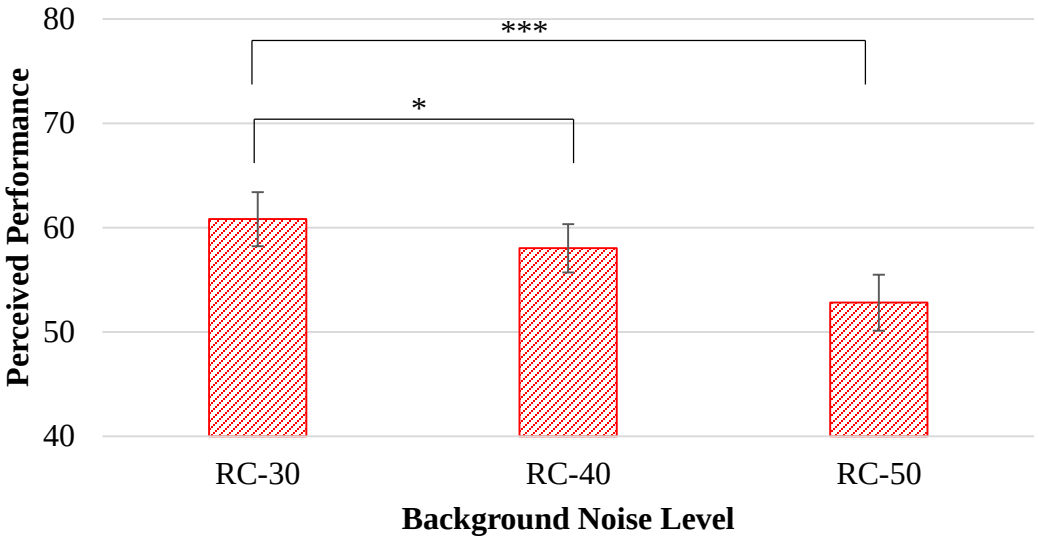


Figure 5.7 - Relation between perceived performance and background noise level, adjusted at standardized English proficiency score at 0. Error bar indicates one standard error.

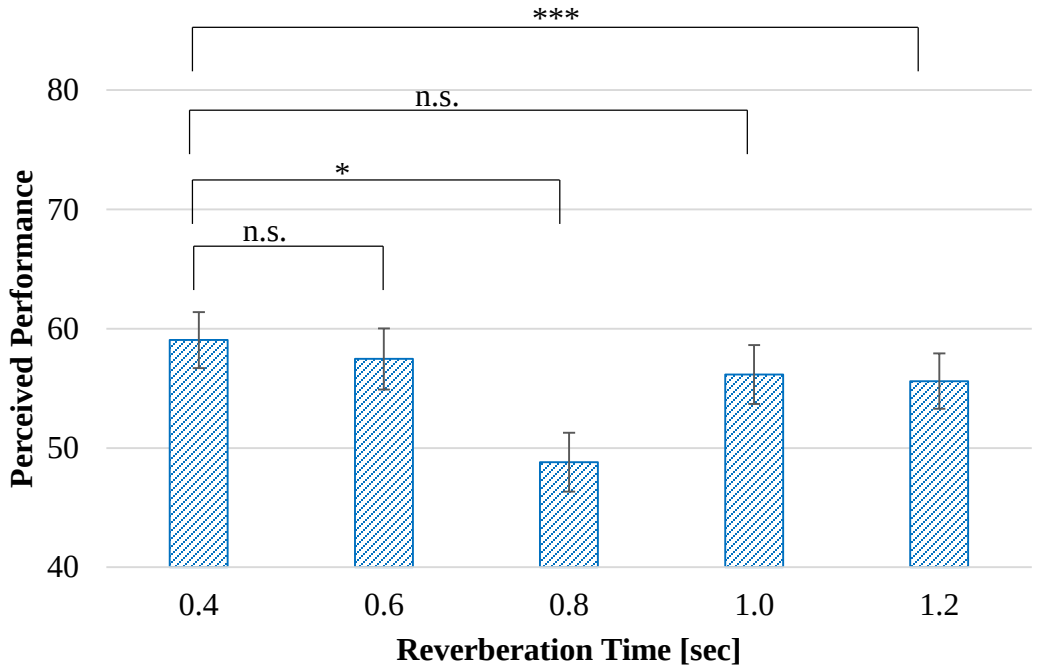


Figure 5.8 - Relation between perceived performance and reverberation time, adjusted at standardized English proficiency score at 0. Error bar indicates one standard error.

5.4 Summary and Conclusions

Study 1 systematically examined the effects of a wide range of BNL (i.e., RC-30, 40 and 50) and RT (between 0.4 and 1.2 seconds) on the objective performance and subjective perception of the comprehension of speech from native American English-speaking talkers by native and non-native English-speaking listeners. In general, the effect of BNL was more detrimental than that of RT on speech comprehension performance, particularly for listeners who were less proficient in English. But the acoustics affected native and non-native listeners differently. BNL and RT were equivalently detrimental to non-native listeners, as indicated in similar effect sizes for the main effects. Non-native listeners with lower English proficiency level are more adversely affected by RT, experiencing greater performance deficit on speech comprehension tasks with increasing RT. On the contrary, native listeners were able to overcome the negative effect of RT, but not for BNL. The strength of BNL on the speech comprehension performance was comparable for both native and non-native listeners. The interaction between BNL and RT was not found to be significant, suggesting that the effects of the acoustic metrics were independent from each other.

Furthermore, the levels of BNL and RT for significant objective performance deficit or subjective perception degradation could be identified to provide guidelines for classroom acoustic designs, if the speech was delivered by native American English talkers and perceived by both native and non-native English-speaking listeners. Interestingly, results showed converging evidence for the acoustic effects on both objective performance and subjective perception of speech comprehension. For BNL, when compared to the most ideal condition of RC-30 among all others, significant

performance deficit of speech comprehension was identified at RC-50 and degradation of perceived performance at RC-40. For RT, significant performance deficit and perception degradation coincided at 0.8 second, when compared against those at the 0.4 second RT scenario. Based on the factor of safety consideration, the RT and BNL design criteria were selected at one level below which significant performance deficit on the speech comprehension tasks was first observed. Therefore, if the design scenario involves both native and non-native listeners in comprehending speech produced by native American English talkers, the classroom acoustics should not exceed 0.6 second RT and RC-40 (or 48 dBA) BNL throughout the room.

Results from Study 1 provided support on relaxing the existing maximum BNL requirement of 35 dBA in the ANSI S12.60-2010 classroom acoustics standard up to 48 dBA (or RC-40), but only for comprehension tasks when the speech is produced by native English-speaking talkers.. The design of RT, however, was shown to be dependent on the nativeness of English of the listeners. Since native English-speaking listeners was not affected by RT, design scenarios involving only native listeners may consider a higher RT up to 1.2 seconds. If the design scenario involves both native and non-native listeners, the existing maximum RT of 0.6 second is still valid.

Chapter 6 – Study 2: Effects of Room Acoustics on Foreign-Accented Speech Comprehension

6.1 Introduction

Instead of native American English speech, Study 2 focused on studying the room acoustic effects on native and non-native listeners' comprehension of foreign-accented English speech. In this study, the speech comprehension test materials from Study 1 were recorded by two native Mandarin Chinese talkers with similar degree of accentedness. Three groups of listeners were recruited to conduct the dual tasks under 15 acoustic conditions (3 BNL X 5 RT, same as in Study 1). The three groups of listeners included: 1) native American English speakers, 2) native Mandarin Chinese speakers, and 3) other non-native English speakers. This chapter discusses the experimental procedures and findings from data collected from these listeners.

6.2 Speech Material Recording

6.2.1 *Recruitment of Native Mandarin Chinese Talkers*

To recruit native Mandarin Chinese talkers with similar degree of accentedness, the commercially available Versant Spoken English Test (Downey *et al.*, 2008) was adopted to screen talker candidates until two (a male and a female) were identified to achieve similar test scores. The Versant Test was administered using a computer test program on a Dell Precision M2400 laptop with internal sound card and an external Sennheisser PC151 headset with microphone included in the listening chamber. The volume setting for the microphone was fixed, but the playback level from the headphone was adjustable for talker candidates in the beginning of the test during calibration.

For the Versant Test, talker candidates were graded in four skill areas, including sentence mastery, vocabulary, fluency, and pronunciation. The two talkers identified for speech recording shared similar scores on the fluency and pronunciation skill areas, as shown in Table 6.1. Although sentence mastery and vocabulary skills also revealed non-native speakers' spoken English proficiency level, these skills were less relevant to the speech recording task in the current study as audio scripts were provided to the talkers. The Versant Test reported t-scores for these skill areas based on the normal distribution from a large database of test takers who were non-native English speakers. Percent rankings were calculated from the t-scores and are reported in Table 6.1.

In addition to spoken proficiency level, speech intelligibility of the chosen talkers was measured as perceived by 10 native English-speaking listeners even though it was not part of the criteria for talker selection. During the individual recording sessions of the speech test materials (detailed description in Section 6.2.2), the two talkers were also asked to record 60 sentences from the revised Bamford-Kowal-Bench (BKB-R) list, included in Appendix E. The BKB-R list was originally developed for testing cochlear devices with British children (Bamford and Wilson, 1979; Bench *et al.*, 1979) but revised for use with American children. Each BKB-R sentence, adopted from Bent and Bradlow (2003), contained three or four keywords and was syntactically simple to non-native English speakers. The recorded sentences were played back via headphones (Sennheisser HE600 with Alexis MultiMix 8 USB 2.0 multichannel mixer) in the sound booth to 10 native English speakers, who were asked to transcribe the sentences into standard English using paper and pencil. The transcriptionists utilized a customized Matlab GUI program

to control audio playback. They were allowed to listen each sentence only once, but could take as long as they wanted to write it down.

Each transcriptionist was first presented a block of 30 randomly selected sentences containing 90 to 95 keywords (depending on the actual sentences) from the BKB-R list spoken by either the male or the female talker, then a second block of the remaining sentences by the other talker. Half of the transcriptionists listened to the male talker first and the other half listened to the female talker first. None of the transcriptionists participated in Study 2; only a few of them previously participated in Study 1 that did not involve foreign-accented speech. Accent intelligibility of the talkers was calculated as percent of the keywords accurately transcribed, as indicated in Table 6.1. The female talker scored significantly higher on accent intelligibility than the male talker, $t(9) = 4.39$, $p = .002$. Despite mediocre percentile rankings among non-native English speakers in the Versant database, the two Mandarin Chinese talkers were highly intelligible to native English-speaking listeners under an ideal listening environment.

To further understand the talkers' foreign accent as perceived under assorted acoustic conditions, the subjective rating on accentedness was also solicited from listener participants at the end of the main experiment sessions. Listener participants were asked to rate the degree of accentedness for each talker using an 11-point scale from 0 to 10, where a "0" represented "no accent at all" and a "10" represented "very heavy accent and impossible to understand." The Shapiro-Wilk test of the accentedness rating suggests normal distribution for the female talker ($p > .05$) but non-normal distribution for the male talker ($p < .001$). As seen in Table 6.1, the female talker with higher intelligibility

score indeed was regarded as less accented than the male talker, as indicated by the non-parametric test of Wilcoxon signed rank test with related-samples, $p < .001$.

Table 6.1 - Tabulated results of Versant Test, accent intelligibility, and subjective acceentedness scale of native Mandarin Chinese talkers

	Native Mandarin Chinese Talker	
	Male	**Female**
Fluency (Versant Test)		
T-score	58	55
Percentile Ranking	81th	70th
Pronunciation (Versant Test)		
T-score	53	52
Percentile Ranking	63th	55th
Accent Intelligibility (Percent Correct)		
Mean	92.2	96.7
SD	3.2	2.6
Accentedness Scale Rating (from 0 to 100)		
Mean	6.9	4.0
SD	1.1	1.7

6.2.2 *Speech Material Recording*

The recording of the speech materials with the two native Mandarin Chinese talkers was conducted in the sound attenuated booth. The sound booth ambient conditions of BNL and RT were reported in Chapter 3

Similar to the native English-speaking talkers in Study 1, the native Mandarin Chinese talkers were also instructed to read the audio scripts at their normal speaking rate for conversations. To preserve the feature of foreign-accent, mispronounced words were

not identified to talkers during the recording sessions. Furthermore, if talkers solicited examples of pronunciation for unfamiliar words, they were encouraged to try without being provided hints or corrections.

The method of calculating speech rate in syllables per second for the Mandarin Chinese talkers was the same as in Study 1. At least five minutes of audio recordings from each Chinese talker were analyzed by two raters who were native English speakers, and the average speech rate is shown in Table 6.2. Again, the two raters showed high agreement on the speech rate calculation with an ICC of 0.95.

Table 6.2 – Talker role assignment and speech rate of native Mandarin Chinese talkers

Native Mandarin Chinese Talker	Recording Assignment of the Speech Comprehension Materials	Speech Rate [Syllables per Second]	
		Mean	95% CI
Male	All four tasks	5.1	[4.9, 5.3]
Female	All four tasks	4.0	[3.9, 4.2]

6.3 Listener Participants

A total of 59 listener participants were recruited on the University of Nebraska at Omaha campus and were categorized in three listener groups, based on their native languages reported on the LEAP-Q described below. The native language profile of the listener participants in Study 2 is included in Appendix D.

Listener Group 1 – Native English-speaking (NAE): This group comprised of 20 participants (12 females), who reported that English was the first learned and currently dominant language. The average age for this group was 22.7 years (SD = 1.3 years).

Listener Group 2 – Native Mandarin Chinese-speaking (NNC): This group comprised of 19 participants (11 females), who reported that Mandarin Chinese was the first learned and currently dominant language. The average age for this group was 26.8 years (SD = 0.9 years).

Listener Group 3 – Other Non-native English-speaking (NNO): This group comprised of 20 participants, whose native and dominant language reported was neither English nor Mandarin Chinese. The average age for this group was 24.8 years (SD = 1.3 years). The native languages spoken by this group of listeners included Ewe (n = 1), Hainanese (n = 1), Hindi (n = 4), Kannada (n = 1), Portuguese (n = 6), and Telugu (n = 7). Although a local dialect in China, the Hainanese-native reported a multi-lingual (non-English) upbringing and inability to communicate fluently in Mandarin Chinese.

None of these listener participants had previously participated in the accent intelligibility tests or any part of Study 1. All listener participants were screened for normal hearing and English proficiency levels, and were tested in the main experiment according to the procedure outlined in Chapter 3. In addition, a talker familiarity screen was given to the listener participants in Study 2 during the initial screen, since the Chinese talkers were recruited from the same community. Among the 59 listener participants, the male talker was correctly identified by one listener and the female talker by two listeners. (The same familiarity screen was not performed in Study 1, because all talkers were recruited from outside of the University of Nebraska community in Lancaster, PA.)

The average length of immersion in the English-spoken community is 78.1 months (range = 2 to 564 months) for all non-native listeners. A histogram showing the

English proficiency levels of all listener participants is included in Figure 6.1. Outliers were not identified in Study 2 among the non-native listeners who achieved exceptional English proficiency.

In the main experiment, all listener participants were able to attend to the dual-tasks simultaneously without losing focus on either task. There was no extreme outlier identified from the performances in either of the dual-tasks.

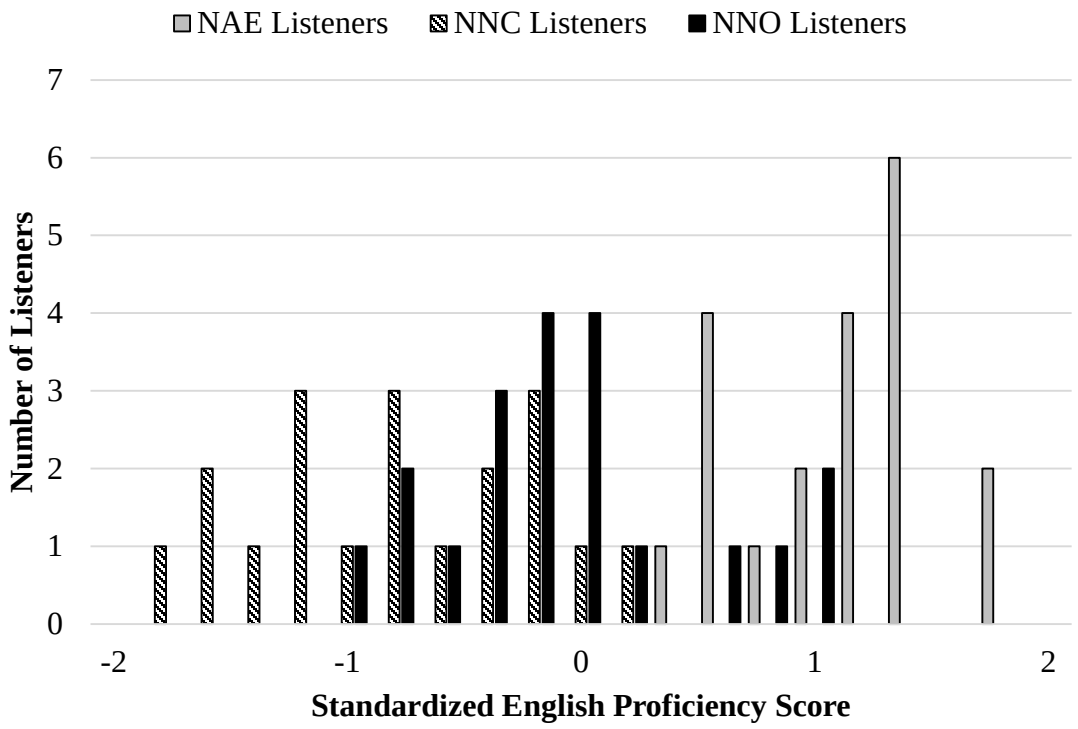


Figure 6.1 - Histogram of standardized English proficiency scores for the three listener groups

6.4 Results

6.4.1 *English Proficiency Level*

A one-way between-subject ANOVA was conducted to compare the differences on English proficiency levels among the three listener groups. There was a significant effect of listener group on English proficiency level, F(2, 56) = 66.16, η_p^2 = 0.69, p < .001. Post hoc comparisons using the Tukey's HSD test indicated that all three listener groups' mean English proficiency levels in standardized scores (for NNC, M = -0.89, SD = 0.60; for NNO, M = -0.17, SD = 0.56; and for NAE, M = 1.02, SD = 0.39) differed significantly from each other at the p < .001 level.

When averaged across all acoustic conditions, the performance on speech comprehension tasks was again significantly predicted by listeners' English proficiency level, b = 5.93, t(58) = 4.52, p < .001. Although a weaker predictor than in Study 1, English proficiency level still explained a significant proportion (25%) of the variance observed in the performance of foreign-accented speech comprehension, R_{adj}^2= 0.25, F(1, 57) = 20.44, p < .001, as seen in Figure 6.2.

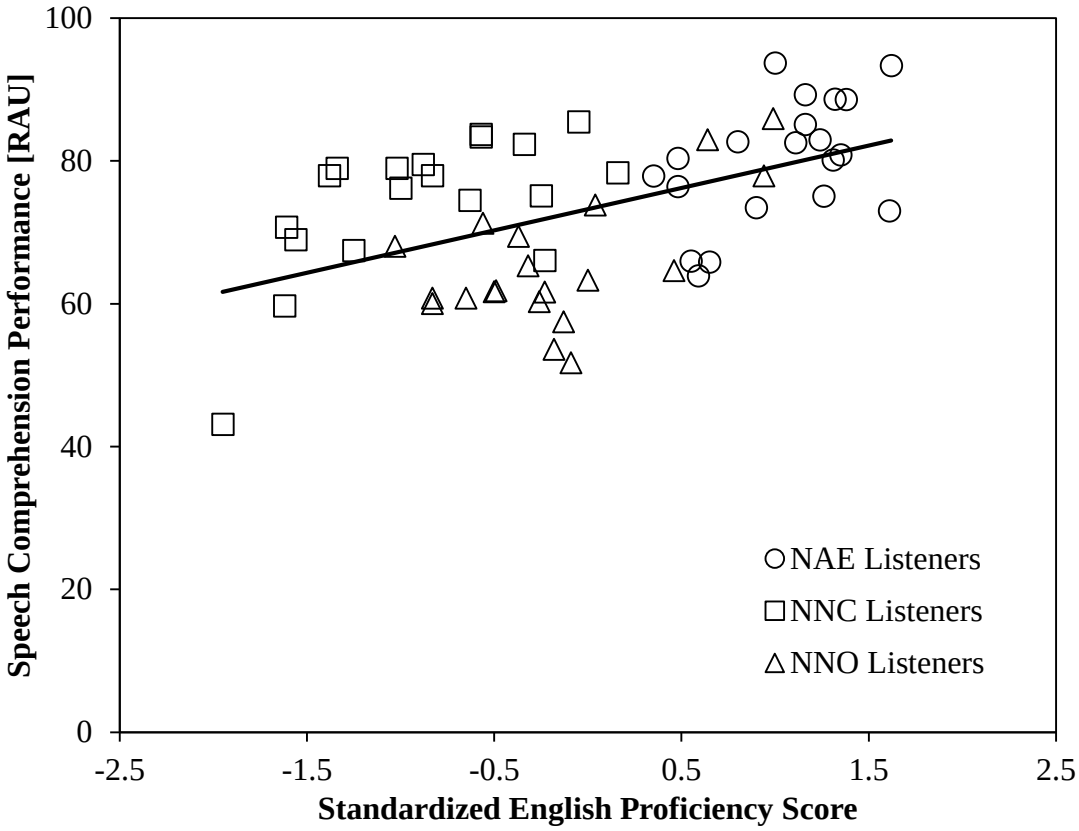


Figure 6.2 - Speech comprehension score, averaged across 15 acoustic conditions, as a function of English proficiency level for both three groups of listeners

6.4.2 *Benefit in Speech Comprehension from Matched Accent*

With the lowest mean English proficiency level as a group, the NNC listener group was likely to perform worst on the speech comprehension tasks than the two other listener groups as predicted by the linear regression model from the previous section. However, a mixed design ANOVA, which examined the within-subject BNL and RT and the between-subject listener group effects on speech comprehension performance suggested otherwise.

There were significant main effects for both acoustic variables of BNL [F(2, 112) = 123.5, p < .001] and RT [F(4, 224) = 6.182, p < .001], as well as for listener group, F(2, 56) = 12.2, η_p^2= 0.28, p < .001. No significant interactions were found. Planned comparisons were performed to compare NAE versus the two non-native listener groups together and between NNC and NNO listener groups, as shown in Figure 6.3. Results show that the NAE listener group scored significantly higher on speech comprehension tasks than the NNC and NNO listener groups together (d = 1.04, p < .001). The NNC listener group scored significantly higher than the NNO group (d = 0.89, p = .006). The results suggest that non-native listeners still perform worse than native listeners on foreign-accented speech comprehension under assorted acoustic conditions. But those who share the same native language with the non-native talkers benefit from the matched accent and are able to understand the accented English speech better than other non-natives who perceive it in mismatched accent.

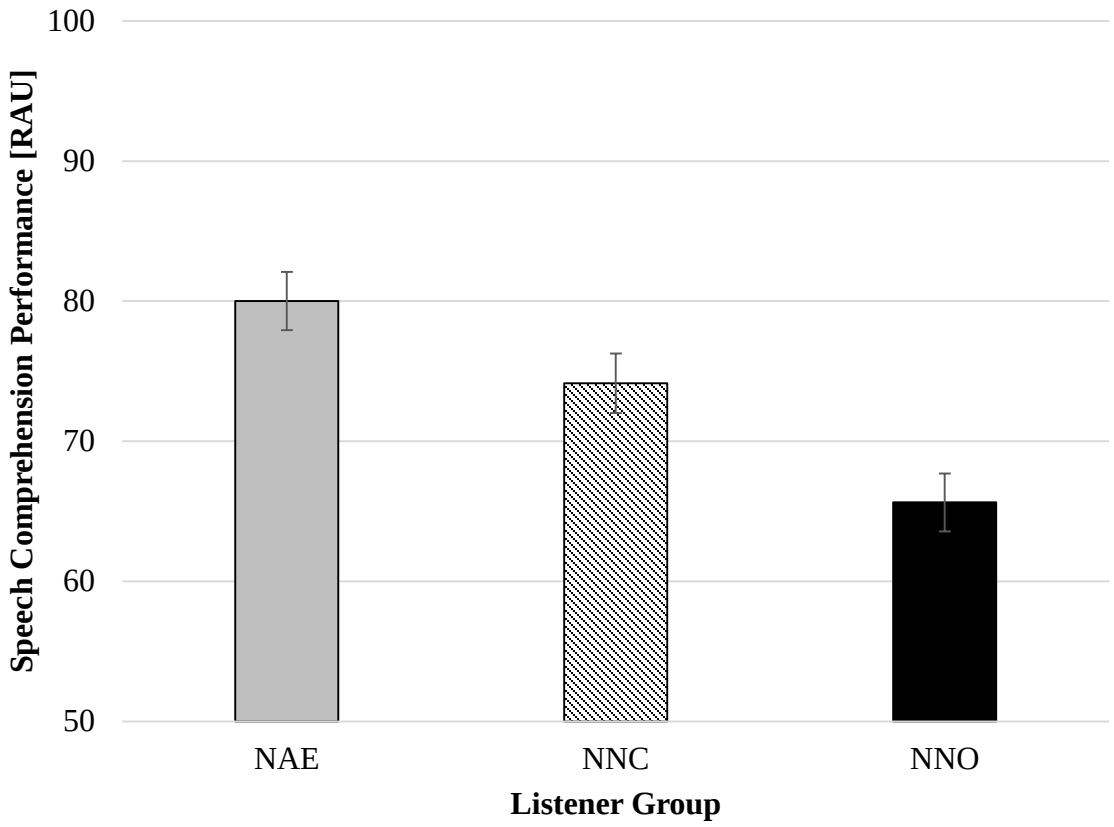


Figure 6.3 - Speech comprehension performance, averaged across all acoustic conditions, for three groups of listeners. Error bar indicates one standard error.

6.4.3 *Objective Performance of Speech Comprehension*

To examine the overall effects of BNL and RT on the simultaneous dual-tasks (speech comprehension and APR tasks), a similar model of mixed-design MANCOVA was applied while controlling for English proficiency level. Using Pillai's trace, there were significant main effects for BNL [$F(4, 54) = 52.04$, $\eta_p^2 = 0.77$, $p < .001$], RT [$F(8, 50) = 4.44$, $\eta_p^2 = 0.33$, $p < .001$], and English proficiency level [$F(2, 56) = 12.35$, $\eta_p^2 = 0.98$, $p < .001$]. No significant interaction was found for the MANCOVA model.

Two follow-up ANCOVAs were conducted using one dependent variable at a time to examine the effects of the acoustic variables. The assumptions of sphericity were satisfied for speech comprehension scores in RAUs, as indicated by non-significant Mauchly's W for BNL (p = .56) and RT (p =.93). Such assumption was violated for the APR dot-tracing measure in RPM for RT only (p < .001; BNL, p = .08). The Greenhouse-Geisser correction for RT (ε = 0.86) in RPM was not applied since it did not suggest different results from calculations with sphericity assumed.

For the speech comprehension tasks, there were significant main effects for BNL [F(2, 114) = 122.85, η_p^2 = 0.67, p < .001], RT [F(4, 228) = 6.12, η_p^2 = 0.09, p < .001], and English proficiency level [F(1, 57) = 20.49, η_p^2 = 0.25, p < .001]. Similar to the findings in Study 1 for the acoustic effects, planned comparisons identified the level of performance degradation with the lowest condition as the reference level. For BNL, as seen in Figure 6.4, listeners performed significantly better in the RC-30 condition than in the RC-40 (d = 31, p = .022) and RC-50 (d = 1.8, p < .001) conditions, respectively. For RT, as shown in Figure 6.5, listeners scored significantly higher under the 0.4 second scenario than in the 0.8 second (d = 0.32, p = .02), 1.0 second (d = 0.42, p = .002), and 1.20 second (d = 0.45, p = .001) scenarios; but not in the 0.6 second scenario (d = 0.04, p = .74). No significant interaction was found between BNL and RT; there existed no interdependence between BNL and RT on the speech comprehension performance. The results suggest that listeners' speech comprehension performance begin to degrade significantly at the RC-40 BNL condition and the 0.8 second RT scenario, respectively. No significant interactions were found in the ANCOVA model for speech comprehension performance.

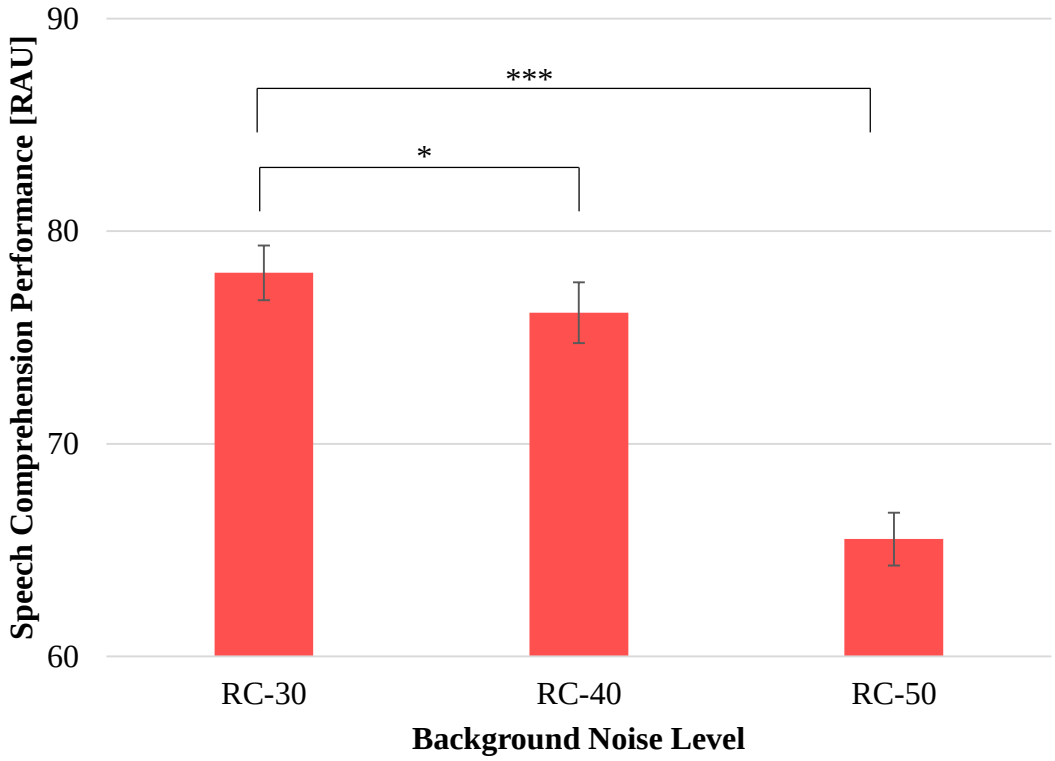


Figure 6.4 - Marginal means of speech comprehension performance on background noise level, adjusted for standardized English proficiency score at 0. Error bar indicates one standard error.

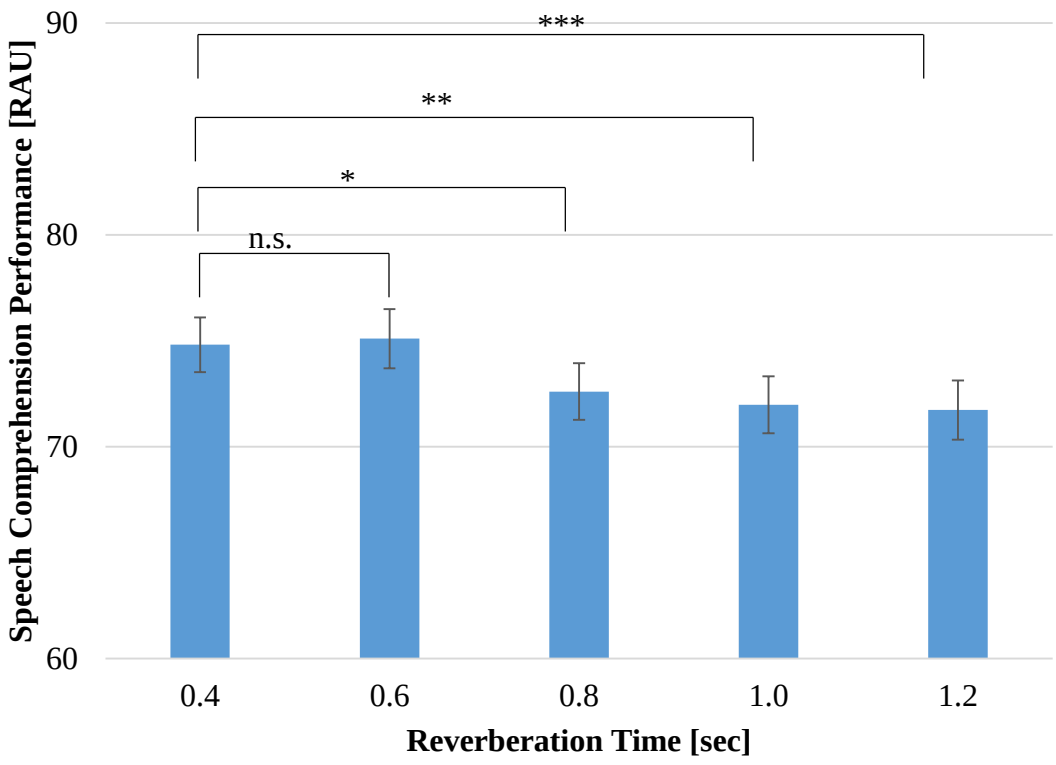


Figure 6.5 - Marginal means of speech comprehension performance on reverberation time, adjusted for standardized English proficiency score at 0. Error bars indicate one standard error.

For the secondary simultaneous APR dot-tracing task, the only significant main effect was found for English proficiency level, $F(1, 57) = 11.05$, $\eta_p^2 = 0.15$, $p = .002$. No significant main effect or interaction was found for BNL or RT ($p > .07$) on the RPM performance. Planned comparison using polynomial contrasts did not reveal any significant trends. Similar to findings in Study 1, results from this follow-up ANCOVA suggest that the dot-tracing task is not affected by the assorted acoustic conditions. Again, the two performance measures of speech comprehension and dot-tracing were found to be positively correlated across all acoustic conditions, as seen in Table 6.3.

Table 6.3 - Pearson correlation coefficient (two-tailed) between performance measures of speech comprehension and adaptive pursuit rotor (dot-tracing) for each acoustic condition.

Pearson's Correlation Coefficient (N = 59 for each acoustic condition)					
	Reverberation Time Scenario				
Background Noise Level	**0.4 sec**	**0.6 sec**	**0.8 sec**	**1.0 sec**	**1.2 sec**
RC-30	0.23	0.26*	0.33*	0.37*	0.26*
RC-40	0.27*	0.17	0.29*	0.40**	0.42*
RC-50	0.48**	0.41**	0.31*	0.35**	0.37**

Note: *p < .05, **p < .01

6.5 Conclusion

Similar testing methodologies from Study 1 were applied to investigate the room acoustic effects on the comprehension of English speech produced by native Mandarin Chinese talkers. Three listener groups were recruited for testing the dual-tasks under the same assortment of acoustic conditions as in the previous study.

It was found that results from Study 2 replicated those from Study 1 on the main effects of BNL and RT on foreign-accented speech comprehension, although a lower BNL condition of RC-30 was preferred when the talkers exhibited moderate foreign accent. Similar to comprehending speech from native American English-speaking talkers, listeners' performance on foreign-accented speech comprehension also degraded significantly beyond 0.6 second of RT. The non-significant interactions between BNL

and RT from both studies suggest that the effects of BNL and RT are relatively independent of each other.

Since both studies agreed on the acoustic effects on speech comprehension, it is reasonable to combine the two datasets to include an additional variable of talker accent for further data analysis in Chapter 7. This chapter on the combined study analyses will discuss the effect of talker accent on speech comprehension by different listener groups under the assorted acoustic conditions.

Chapter 7 – Combined Analysis: Effects of Talker Accent on Speech Comprehension under Acoustic Conditions

7.1 Introduction

In this chapter, data from Study 1 and 2 are combined to investigate the effect of talker accent on speech comprehension performance under assorted acoustic conditions. A comprehensive analysis of the room acoustic effects, specifically background noise level (BNL) and reverberation time (RT), was conducted to discuss the acoustic design criteria for classrooms whose occupants were of diverse linguistic backgrounds.

7.2 Listener Participants from Study 1 and 2

The listener participants from Study 1 and 2 were regrouped into three listener groups: 1) native American English-speaking (NAE), 2) native Mandarin Chinese-speaking (NNC), and 3) other non-native English-speaking (NNO). The descriptive statistics for the listener participants from both studies are shown in Table 7.1.

Table 7.1 - Descriptive statistics of listener participants in both studies

	Study 1 - NAE Talkers			**Study 2 - NNC Talkers**		
Listener Group	**NAE**	**NNC**	**NNO**	**NAE**	**NNC**	**NNO**
N	26	10	19	20	19	20
Age						
Mean	24	26	27	23	27	25
Range	19-40	23-31	19-43	17-36	19-33	19-46
SD	5.9	2.3	6.2	5.8	3.9	5.6
Standardized English Proficiency Score						
Mean	0.96	-0.72	-0.60	0.91	-1.02	-0.30
SD	0.38	0.40	0.52	0.40	0.61	0.57
Speech Comprehension Performance [RAU]						
Mean	90.7	82.9	79.4	80.0	74.1	65.6
SD	7.5	5.9	5.7	8.7	10.1	9.0

Note: Speech comprehension performance averaged across 15 acoustic conditions

The composite scale of English proficiency level achieved a Cronbach's α of 0.94 in the combined dataset. The linear relation between speech comprehension performance, averaged across 15 acoustic conditions, and standardized English proficiency level is plotted in Figure 7.1. In the participant sample combining listeners from both studies, English proficiency level significantly explained 33 % of the variance observed in the speech comprehension performance, R^2_{adj}= 0.33, F(1, 113) = 56.91, p < .001.

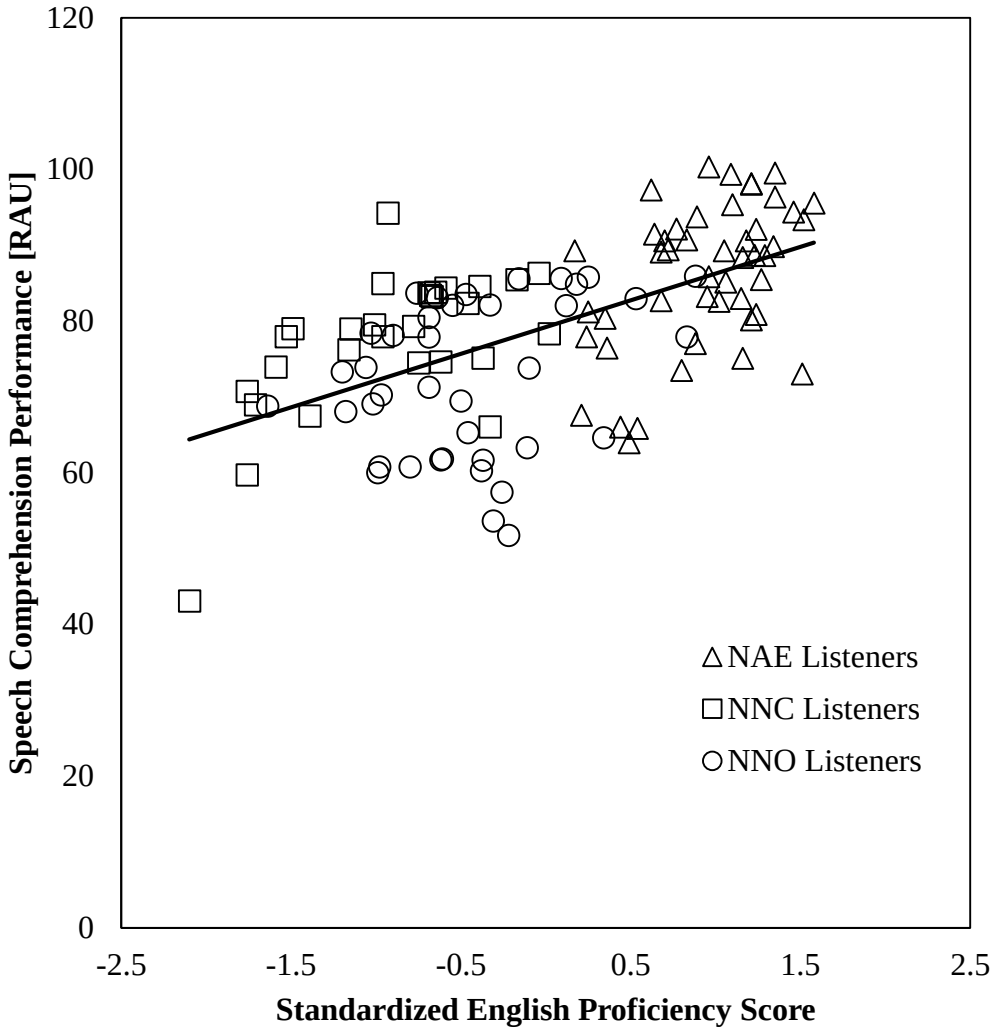


Figure 7.1 - Speech comprehension score, averaged across 15 acoustic conditions, as a function of English proficiency level for all listeners from Study 1 and 2

7.3 Results

7.3.1 *Effect of Foreign Accent*

The effect of talker foreign accent was examined in the context of objective performance of speech comprehension and subjective perception evaluated by the NASA TLX under BNL and RT by different listener groups. The following sections discuss the effect of talker accent on speech comprehension performance using two paradigms of MANOVA and effect size comparisons, as well as its impact on subjective perception from individual subscales in NASA TLX.

7.3.1.1 *Main Effects and Interactions by Listener Group on Speech Comprehension*

After combining datasets from Study 1 and 2, a mixed-design MANOVA was applied to examine the effects of acoustics and foreign accent on the simultaneous dual-tasks of speech comprehension and APR dot-tracing. Two between-subject variables were included in this model for talker accent (American English vs. Mandarin Chinese) and listener group (NAE vs. NNC vs. NNO). In this model, English proficiency level was not controlled for comparisons among listener groups. Both BNL and RT remained as the within-subject variables.

Using Pillai's trace, there were significant main effects for talker accent [$F(2, 107) = 24.08$, $\eta_p^2 = 0.30$, $p < .001$], listener group [$F(4, 216) = 12.67$, $\eta_p^2 = 0.08$, $p < .001$], BNL [$F(4, 105) = 75.05$, $\eta_p^2 = 0.74$, $p < .001$], and RT [$F(8, 101) = 5.75$, $\eta_p^2 = 0.26$, $p < .001$]. The two-way interaction of BNL X talker accent was found to be significant, $F(4, 405) = 4.42$, $\eta_p^2 = 0.11$, $p = .002$. Another two-way interaction of BNL X listener group was not statistically significant, $F(4, 105) = 1.97$, $\eta_p^2 = 0.03$, $p = .052$. A three-way

interaction of BNL X RT X talker accent was also found to be significant, F(16, 93) = 1.81, η_p^2 = 0.11, p = .041.

In the follow-up ANOVA of APR dot-tracing performance in RPM to the above MANOVA, there was only one significant main effect for BNL, F(2, 216) = 3.95, η_p^2 = 0.03, p = .021. The follow-up ANOVA of speech comprehension performance, using talker accent, listener group, BNL and RT as independent variables, revealed several interesting significant main effects and interactions. The statistical significant main effects included talker accent [F(1, 108) = 48.62, η_p^2 = 0.30, p < .001], listener group [F(1, 108) = 26.12, η_p^2 = 0.31, p <.001], BNL [F(2, 216) = 146.38, η_p^2 = 0.57, p <.001], and RT [F(4, 432) = 8.42, η_p^2 = 0.06, p <.001]. The two-way interactions were found significant for BNL X talker accent [F(2, 216) = 7.82, η_p^2 = 0.06, p =.001] and BNL X listener group [F(4, 216) = 2.55, η_p^2 = 0.03, p = .04]. The only significant interaction involving RT was a three-way interaction of RT X talker accent X listener [F(8, 432) = 2.38, η_p^2 = 0.02, p = .016].

For talker accent, post hoc analysis was performed to compare listeners' comprehension performance of speech produced by native English-speaking versus native Mandarin Chinese-speaking talkers. It was found that listeners performed worse in comprehending English speech with Mandarin Chinese accent, (d = 0.65, p <.001], as seen in Figure 7.2. The performance deficit in speech comprehension was as much as 10 RAU, or approximately 10% in accuracy.

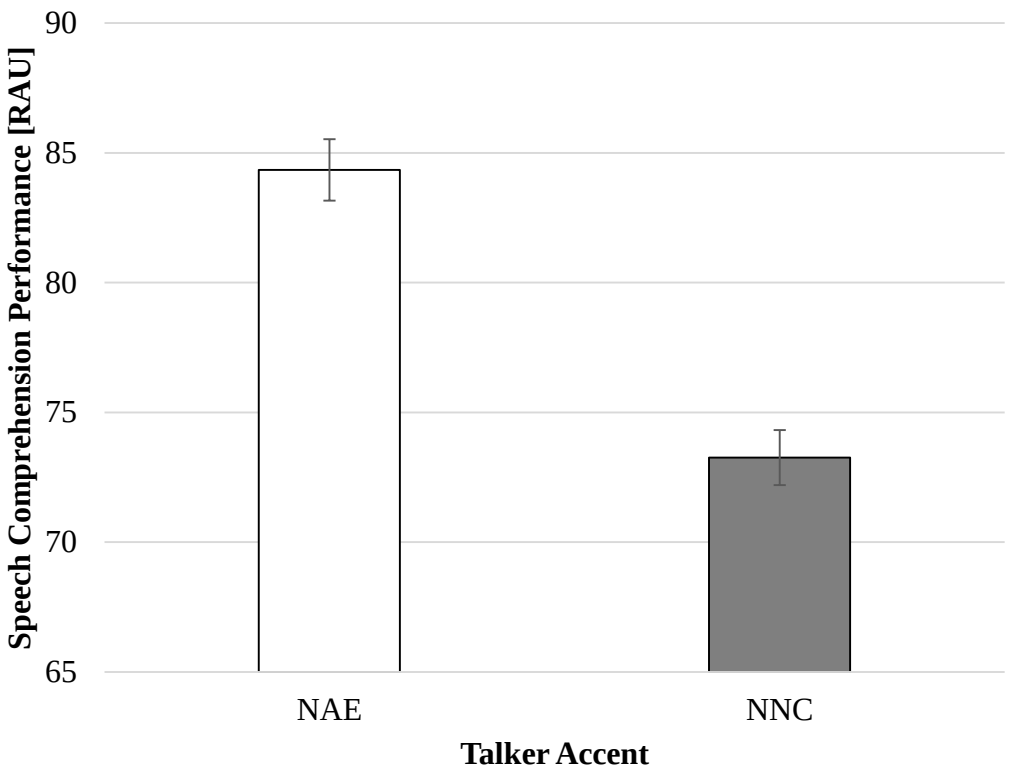


Figure 7.2 - Marginal means of comprehension performance of speech produced by native American English (NAE) talkers versus native Chinese Mandarin (NNC) talkers. Error bar indicates one standard error.

For listener group, pairwise comparison using Tukey's HSD suggested all possible pairs were statistically significant ($d = 0.43$, $p < .001$ for NAE vs. NNC; $d = 0.23$, $p = .045$ for NNC vs. NNO; $d = 0.72$, $p < .001$ for NAE vs. NNO). The marginal means of speech comprehension performance are plotted in Figure 7.3 for all three listener groups. When controlling for the effects of acoustics and talker accent, NAE listeners always achieved higher performance than non-native listeners on speech comprehension. Despite scoring lower on the English proficiency composite scale as a group, NNC listeners actually performed significantly better on speech comprehension

than NNO listeners when averaged across two studies. It implies that NNC listeners may have benefited from the matched accent on speech comprehension in Study 2. The interlanguage benefit of matched accent will be further discussed in the next section.

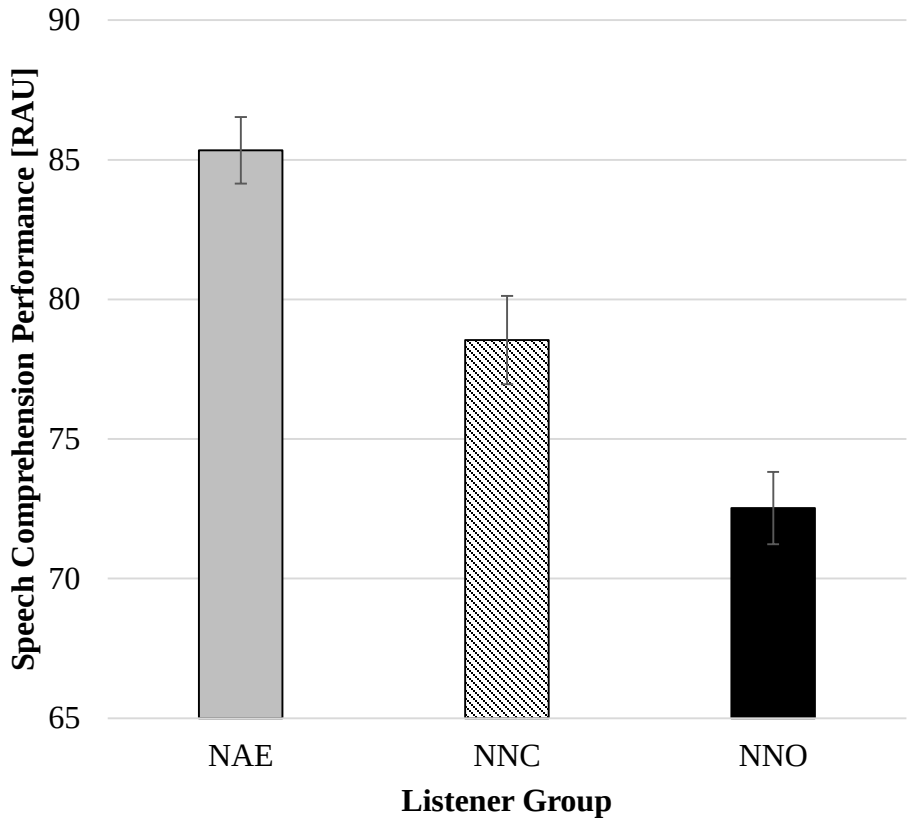


Figure 7.3 - Marginal means of speech comprehension performance of three listener groups. Error bars indicate one standard error.

For BNL and RT, planned comparisons were conducted separately on these two factors following the significant main effects using the lowest levels as the reference comparison. The levels of significant performance reduction were identified at RC-40 (vs. RC-30, $d = 0.26$, $p = .009$) for BNL and 0.8 second (vs. 0.4 second, $d = 0.35$, $p < .001$) for RT, similar to findings in Study 2.

For the significant interaction of BNL X talker accent, planned comparisons showed that the performance deficit of comprehending Chinese-accented speech was significantly greater under the RC-50 than the RC-30 condition, p = .001 (Figure 7.4). BNL, particularly the RC-50 condition, was more detrimental to the comprehension of Chinese-accented speech for all listeners.

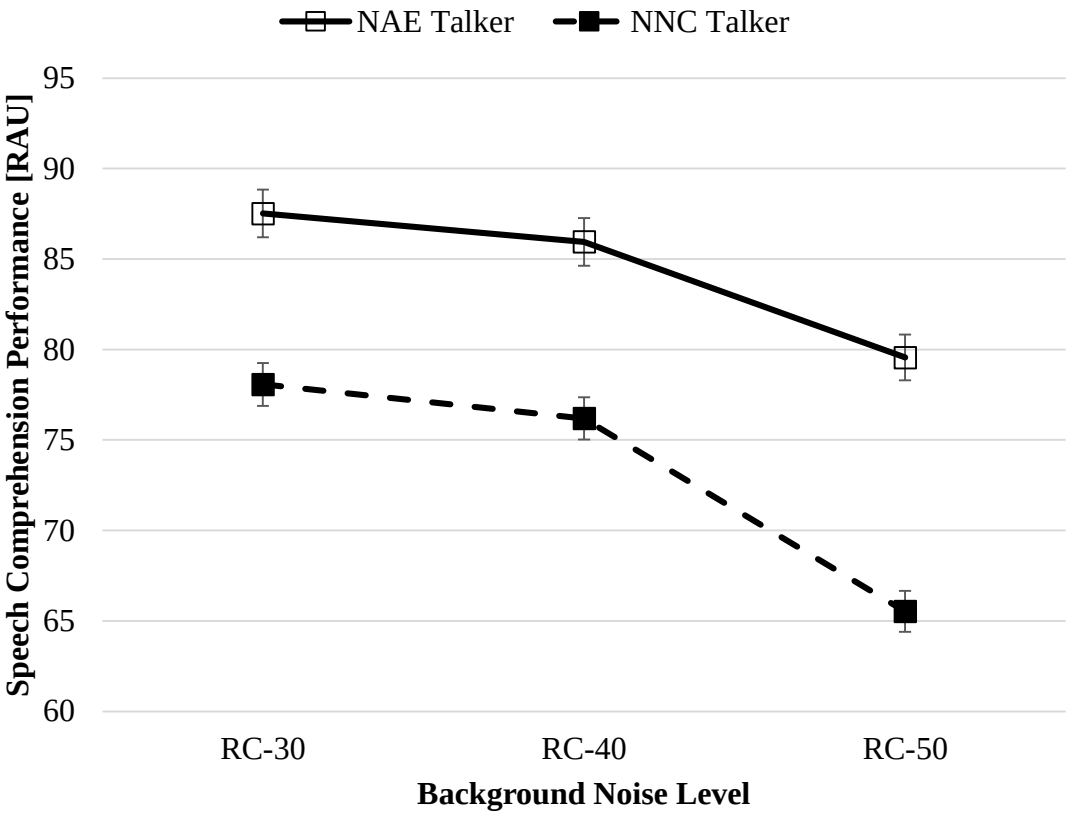


Figure 7.4 - Two-way interaction between BNL and talker accent on speech comprehension performance. Error bar indicates one standard error.

For the significant interaction between BNL X listener group, as shown in Figure 7.5, planned comparisons suggested that performance deficit in speech comprehension

between the RC-30 and RC-50 BNL conditions significantly differed across listener groups (p = .019).

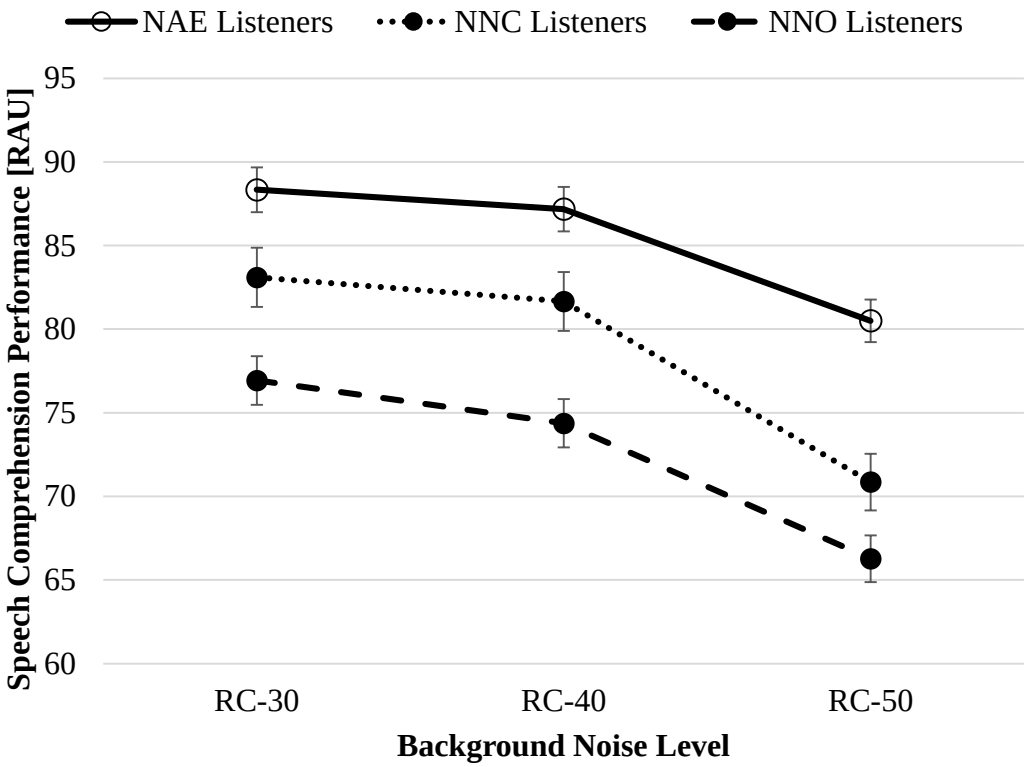


Figure 7.5 - Two-way interaction between BNL and listener group on speech comprehension performance. Error bar indicates one standard error.

To further understand which listener group was more severely affected by increasing BNL, a separate one-way ANOVA was conducted to predict the performance deficit of speech comprehension between the two BNL conditions using the listener group as the between-subject variable. Planned comparisons suggested that, as BNL increased from RC-30 to RC-50, NAE listeners (M = 7.4, SD = 7.17) experienced significantly less performance deficit than NNC and NNO listeners together (M = 11.6, SD = 6.88), p = .001. In general, NAE listeners were less affected by BNL as compared

to non-native listeners. No significant difference in the performance deficit was found between NNC (M = 12.67, SD = 6.97) and NNO (M = 10.72, SD = 6.78), $p \geq .25$.

The significant three-way interaction between RT X talker accent X listener group was slightly more difficult to interpret. Planned contrast comparisons revealed significant pairs of RT between 0.4 versus 0.8 second ($p = .013$) and 0.4 versus 1.2 seconds ($p = .019$). In Figure 7.6, the mean difference of speech comprehension performance between NAE and NNC talker accents are plotted for the three listener groups in the 0.4, 0.8 and 1.2 seconds RT scenarios. The significant three-way interaction suggests that the variations in performance deficit due to foreign accent differed across listener groups. For instance, NAE listeners experienced significantly greater performance deficit under the 0.8 and 1.2 seconds than in the 0.4 second RT. But for NNC and NNO listeners, the Chinese accent did not incur significantly greater performance deficit with increasing RT. NNO listeners experienced the greatest performance deficit among all three listener groups under all scenarios in RT.

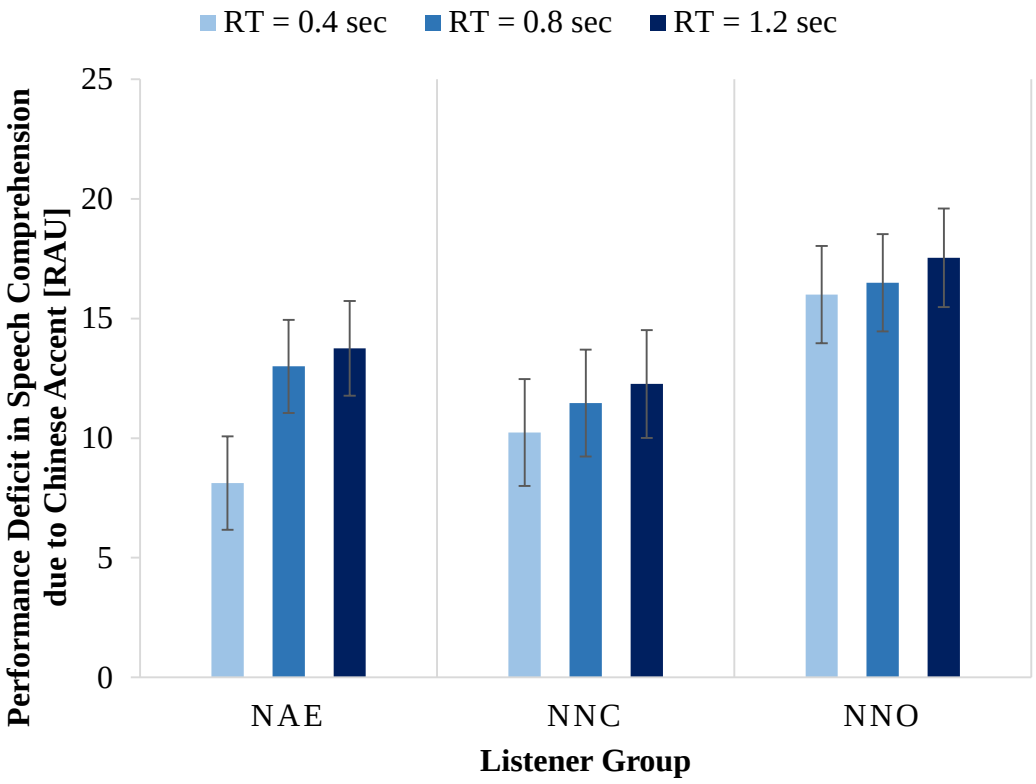


Figure 7.6 - Three-way interaction between talker accent shown as performance deficit due to Chinese accent, listener group (NAE vs. NNC vs. NNO) and reverberation time (0.4 vs. 0.8 vs. 1.2 sec). Error bar indicates one standard error.

7.3.1.2 *Interlanguage Benefit of Matched Foreign Accent on Speech Comprehension*

Bent and Bradlow (2003) suggested that there was an interlanguage benefit for non-native listeners in perceiving foreign-accented English speech using a speech intelligibility task, particularly when the talker and listener shared the same accent. The post hoc analysis in the ANOVA model from Section 7.3.1.1 to compare listener group difference hinted that such benefit of matched accent seemed to also exist in speech comprehension tasks, which are at a higher level of language processing. The next step was to verify such benefit for speech comprehension tasks under assorted acoustic conditions and by controlling for listeners' English proficiency level. The paradigm of effect size comparison was utilized.

For each listener group, an ANCOVA was applied to examine the speech comprehension performance using English proficiency level and talker accent as the between-subject variables and BNL and RT as the within-subject variables. The significant main effects and interactions are listed in Table 7.2 for the three listener groups.

Table 7.2 - Effect size comparison of significant main effects and interactions on speech comprehension performance among three listener groups

	NAE Listeners $\binom{N_1 = 26}{N_2 = 20}$		**NNC Listeners** $\binom{N_1 = 10}{N_2 = 19}$		**NNO Listeners** $\binom{N_1 = 19}{N_2 = 20}$	
	p-value	η_p^2	**p-value**	η_p^2	**p-value**	η_p^2
English Proficiency Level	**<.001**	**0.27**	**.002**	**0.33**	**<.001**	**0.46**
Talker Accent (NAE vs. NNC)	**<.001**	**0.36**	≥.056	0.13	**<.001**	**0.68**
BNL	**.004**	**0.12**	**<.001**	**0.37**	**<.001**	**0.44**
RT	≥.38	0.02	≥.18	0.06	**.001**	**0.12**
BNL X Talker Accent	**<.001**	**0.2**	≥.51	0.03	.068	0.07

Note: N_1 = Number of listeners in Study 1 (NAE talkers); N_2 = Number of listeners in Study 2 (NNC talkers). Bold values indicate statistical significant results.

As seen in the above table, English proficiency level retained the statistical significant main effect in speech comprehension performance with comparable effect size in η_p^2 across all listener groups. Talker accent (NAE vs. NNC) was a significant and strong predictor in both NAE and NNO listener groups. Although marginally significant in the NNC listener group, talker accent had a much weaker effect size, explaining only 33% of the variance in NNC listeners' speech comprehension performance while all other variables were controlled. NNC listeners were less affected by Chinese-accented speech than the other two groups of listeners, suggesting the interlanguage benefit due to matched accent.

Comparisons were also conducted for the main effects and interactions involving the two acoustic variables of BNL and RT. In Chapter 5, similar comparisons were conducted between native and non-native listeners from Study 1, where speech was produced by native talkers of American English. It was previously reported that native listeners were able to overcome the negative effect of RT (non-significant main effect) but not BNL (marginally significant main effect with moderate effect size). But for non-native listeners, both BNL and RT were equivalently detrimental as quantified by the similar η_p^2 for the main effects. In general, non-native listeners were more susceptible than native listeners to both BNL and RT in speech comprehension. It was concluded that larger effect size of BNL and RT on speech comprehension, while controlling for English proficiency level, was a distinct characteristic for non-native listeners. As a result, the similar trend of effect size was expected for the acoustic variables in the updated dataset combining listeners from both Study 1 and Study 2.

For the NAE and NNO listener groups, as shown in Table 7.2 , the significance levels of BNL and RT replicated those in Study 1 (see Table 5.4 in Chapter 5). The effect size of BNL remained similar for both NAE and NNO listener groups. The main effect of BNL has become statistically significant for the NAE group with a larger sample size. And, the effect of BNL became stronger for the NNO listener group increasing from 0.20 to 0.44 in η_p^2. The effect of RT was also in agreement with the previous finding for these two listener groups. It remained weak for the NAE listeners and moderate for the NNO listeners. In summary, NAE listeners who were generally more proficient in English were also better at suppressing negative effects from BNL and RT than NNO listeners. While the hypothesis of similar effect sizes for the acoustic variables was confirmed, effect sizes

calculated for the NNC listener group provided an opportunity to examine the interlanguage benefit of matched accent in speech comprehension in background noise and reverberation.

For the NNC listeners, the main effect of BNL was both statistically significant and strong as indicated by a η_p^2 of 0.37, which was slightly smaller than that for the NNO listeners. However, the RT main effect remarkably weakened and became statistically non-significant. It suggests that NNC listeners were also able to overcome the negative impact of RT, delineating the distinction with their non-native peers in the NNO listener group.

Two potential factors were identified in contributing to the improved ability in suppressing the negative acoustic effects from previous investigations: higher English proficiency level and the interlanguage benefit of matched accent. However, as shown in Figure 7.1 (Section 7.2), NNC listeners as a group actually scored lowest on the composite scale of English proficiency level, eliminating the possibility of improved ability in suppression due to higher language proficiency level. It was thus concluded that NNC listeners received interlanguage benefit in comprehending foreign-accented speech produced by talkers who matched the same accent to improve the ability in suppressing the negative effects of reverberation.

The only significant interaction found in the factorial design across all listener groups was between BNL X talker accent for the NAE listeners. This specific interaction is illustrated in Figure 7.7. As previously reported, all listeners performed worse on comprehension tasks when speech was produced by the NNC talkers as opposed to the NAE talkers; and performance also deteriorated with higher BNL. Furthermore, the

performance deficit between RC-30 and RC-50 was significantly greater for NAE listeners, but not for NNC or NNO listeners.

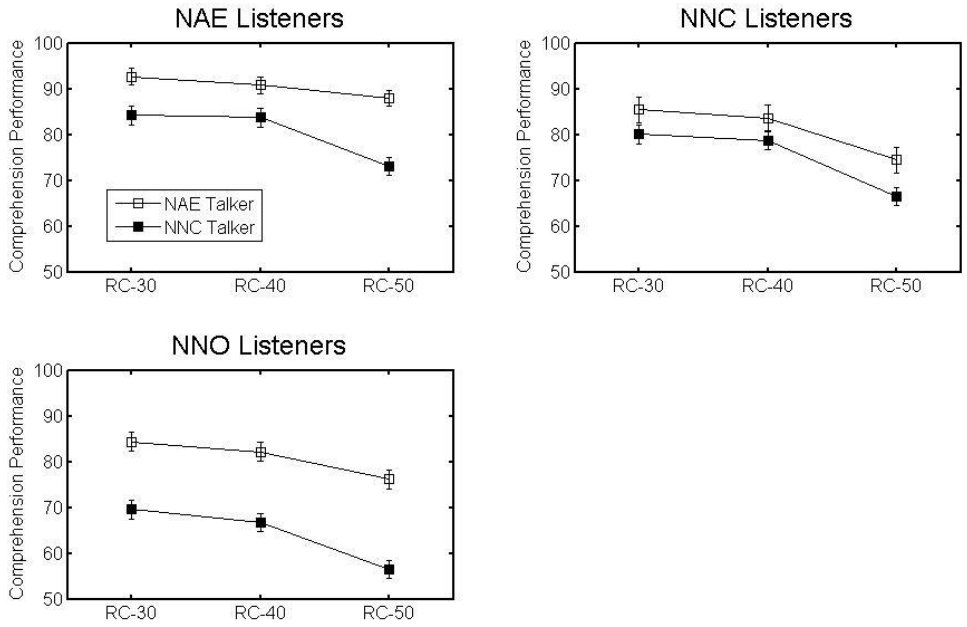


Figure 7.7 - Two-way interaction between BNL and talker accent for the NAE, NNC and NNO listener groups

7.3.1.3 *On Subjective Perceptions of Workload Assessment*

In addition to the objective performance of speech comprehension, it was worthwhile to examine the effect of talker accent on the subjective perception by listeners. A mixed-design ANOVA using BNL, RT, listener group, and talker accent as the independent variables was applied to the six individual subscales (as dependent variables) to answer the following questions.

1) Does foreign accent also degrade the subjective perceptions of listeners?

2) Does the interlanguage benefit of matched accent identified for the NNC listener group improve their subjective perceptions?

The main effect of talker accent was only found significant in the ANOVAs for frustration [F(1, 109) = 7.15, η_p^2 = 0.05, p =.009] and perceived performance [F(1, 109) = 8.20, η_p^2 = 0.06, p =.005]. When Mandarin Chinese accent was introduced, listeners felt more frustrated (M = 44.0, SD = 2.2 for NAE talkers; M = 51.9, SD = 2.0 for NNC talkers) during the speech comprehension tasks. And, they also reported to achieve lower performance on the comprehension tasks (M = 55.5, SD = 2.2 for NAE talkers; M = 47.2, SD = 1.9 for NNC talkers). Interestingly, listeners did not report experiencing increase in mental, physical or temporal demand due to the foreign accent. Furthermore, the accented speech did not incur more effort among listeners to complete the simultaneous dual tasks either.

The interlanguage benefit of matched accent was also realized for listeners' subjective perceptions. The significant two-way interactions of talker accent X listener group were found for effort (p = .030), frustration (p = .032), and perceived performance (p = .027). To examine whether NNC listeners perceived differently, two separate ANOVAs were fitted to the dataset for effort, frustration and perceived performance ratings to test the two-way interaction between talker accent X listener group that contained either NNC and NAE or NNC and NNO listener groups. The observed change in the effort rating between talker accents significantly differed between NNC and NAE listeners (p = .01), but not between NNC and NNO listeners (p = .17). The effort rating is illustrated for the three listener groups in Figure 7.8. Despite the significant interaction in

the effort rating, it does not provide sufficient support of the interlanguage benefit of matched accent by failing to identify the distinction between NNC and NNO listeners. Such benefit was in fact realized from the two other subjective perception ratings.

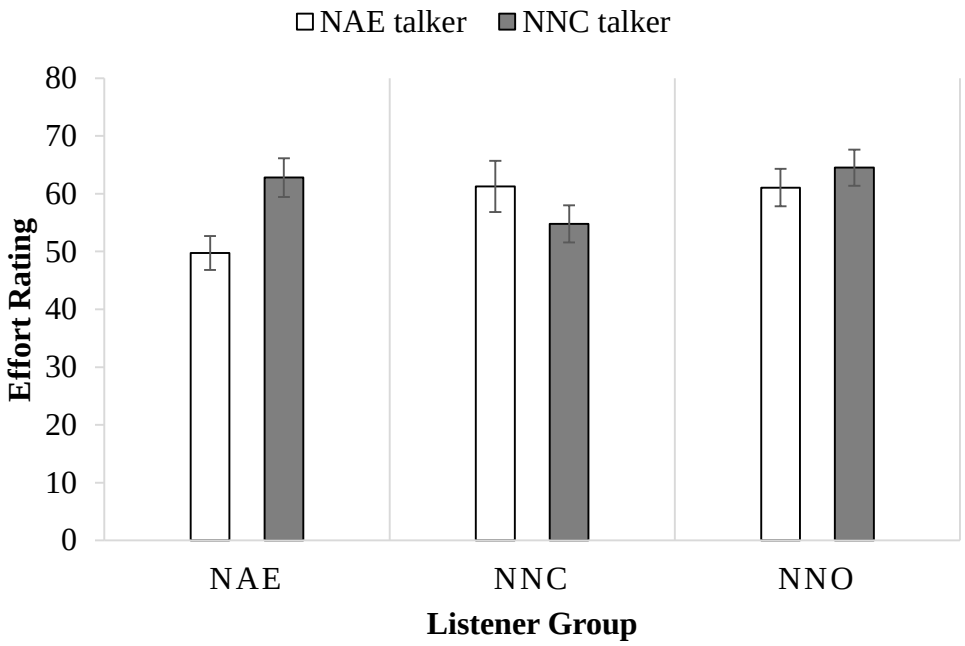


Figure 7.8 - Interaction of talker accent and listener group for the effort rating in NASA TLX. Error bar indicates one standard error.

From the frustration rating, both NNO and NAE listeners reported feeling more frustrated with the Chinese accent, while the NNC listeners reported no significant change in frustration (see Figure 7.9). The change in the frustration rating significantly differed between the NNC versus NAE listeners ($p = .01$) and the NNC versus NNO listeners ($p = .02$). A similar trend was also observed from the perceived performance rating, as shown in Figure 7.10. The change in perceived performance significantly

differed between the NNC and NNO listeners (p = .002), but not between the NNC and NAE listeners (p = .10).

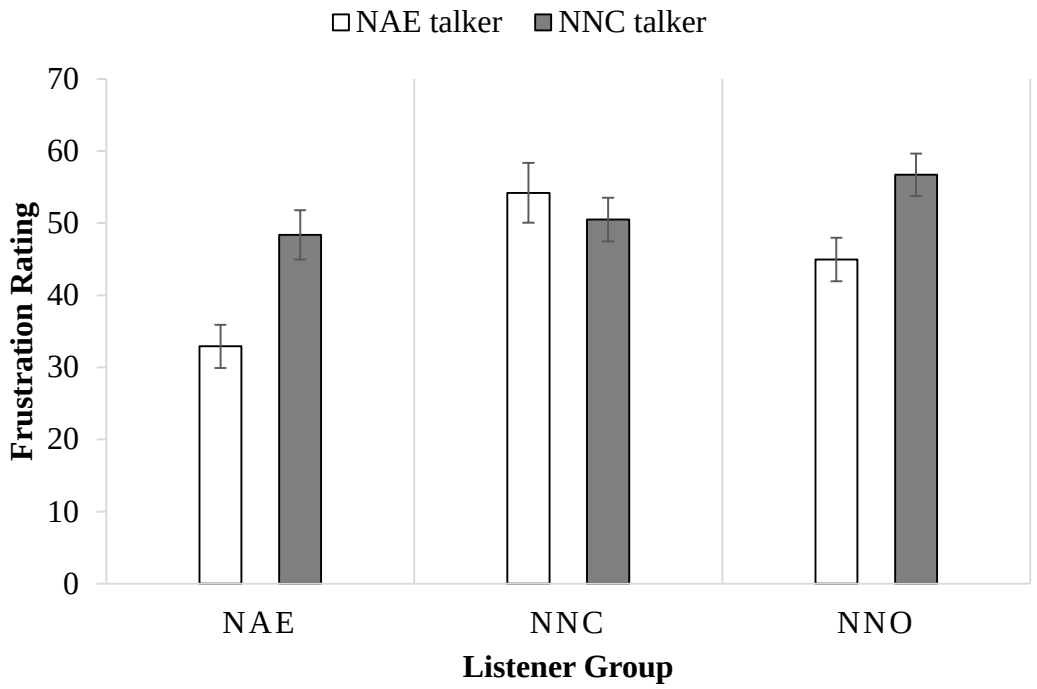


Figure 7.9 - Interaction of talker accent and listener group for the frustration rating in NASA TLX. Error bar indicates one standard error.

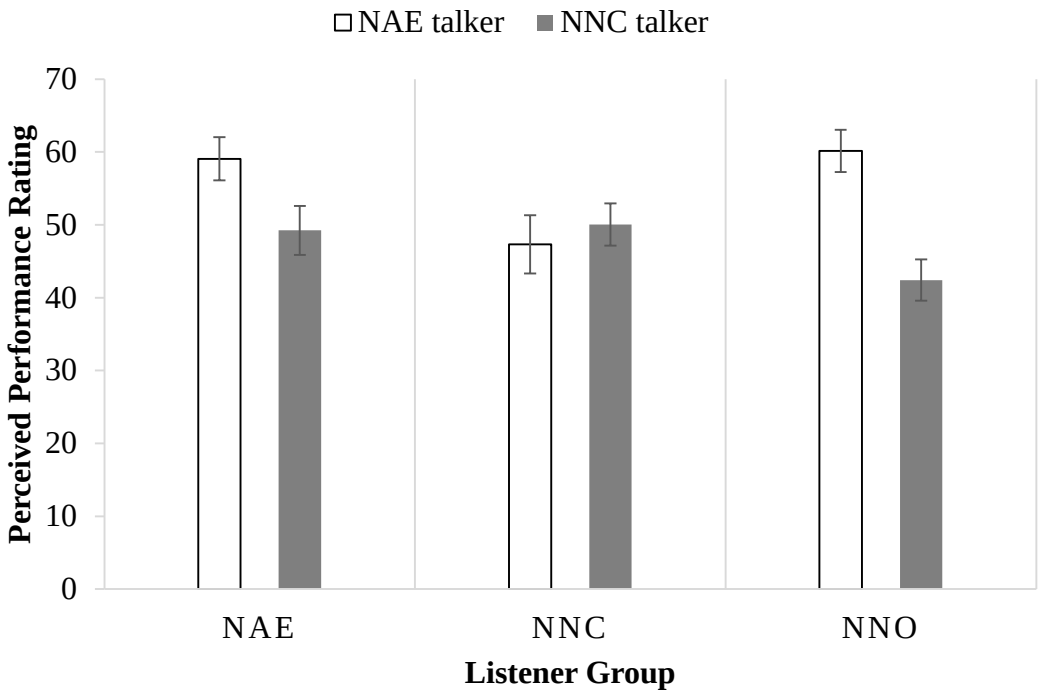


Figure 7.10 - Interaction of talker accent and listener group for the perceived performance rating in NASA TLX. Error bar indicates one standard error.

7.3.2 *Objective Performance of Speech Comprehension*

All analyses discussed so far have mainly focused on examining aspects of the data. The acoustic effects were studied under specific circumstances when the speech materials were produced by native American English-speaking talkers in Study 1 or by native Mandarin Chinese-speaking talkers in Study 2. The effect of talker accent was discussed in the previous section in this chapter and how it has influenced the comprehension performance among different listener groups. The interlanguage benefit of matched foreign accent was identified, suggesting listeners who shared the same accent with the talkers were at an advantage in understanding speech under assorted conditions of BNL and RT.

These detailed discussions of results provided insights to designing classroom acoustics for specific user cases. However, for practical classroom acoustic designs, the precise composition of occupants (e.g., ratio of native vs. non-native listeners) and the specific user cases (e.g., frequency of non-native talkers giving lectures) are often unattainable. The difficulty of categorizing individual occupants into listener groups challenges the applicability of the previous results in practical classroom acoustic designs. Therefore, a comprehensive model controlling for English proficiency level, instead of listener group, is deemed more appropriate to provide guidelines for design purpose.

To examine the effects of acoustics and talker accent comprehensively, a mixed-design MANCOVA was applied with two follow-up ANCOVAs on the performances of speech comprehension and APR dot-tracing. The within-subject independent variables included BNL and RT; and the between-subject independent variables included English

proficiency and talker accent. The full factorial MANOVA revealed significant main effects for BNL [F(4, 109) = 73.12, η_p^2 = 0.71, pp <.001], RT [F(8, 105) = 5.45, η_p^2 = 0.19, p <.001], English proficiency level [F(2, 111) = 32.70, η_p^2 = 0.35, p < .001], and talker accent [F(2, 111) = 26.93, η_p^2 = 0.30, p < .001]. Statistically significant two-way interactions included BNL X English proficiency [F(4, 109) = 2.63, η_p^2 = 0.02, p = .038] and BNL X talker accent [F(4, 109) = 5.92, η_p^2 = 0.12, p < .001]. One three-way interaction between BNL X RT X talker accent was also found statistically significant [F(16, 97) = 2.23, η_p^2 = 0.03, p = .009].

The follow-up ANCOVA of the APR dot-tracing performance, using the same set of independent variables from the MANCOVA, revealed only one significant main effect of BNL, F(2, 224) = 3.59, η_p^2 = 0.02, p = .029. Planned comparisons showed a significant quadratic trend of RPM, p = .004. As seen in Figure 7.11, listeners performed best on the dot-tracing task under the RC-40 condition. There was also a significant three-way interaction between BNL X RT X talker accent, F(8, 896) = 2.24, η_p^2 = 0.02, p = .023.

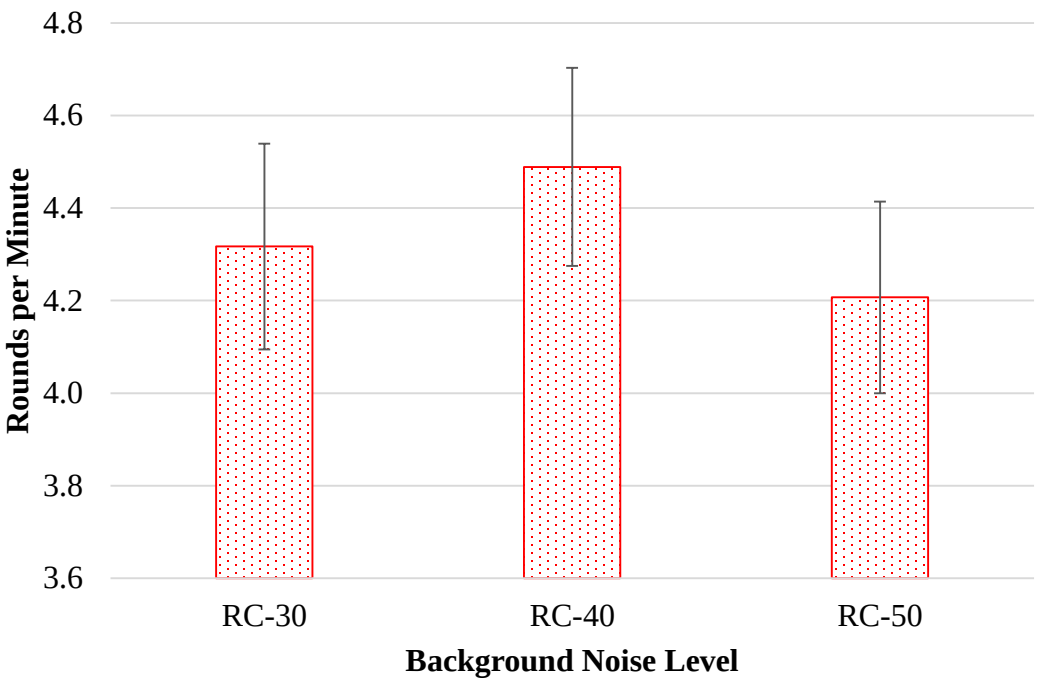


Figure 7.11 - Effect of background noise level on the APR dot-tracing performance (in RPM), adjusted for standardized English proficiency score at 0. Error bars indicate one standard error.

From the follow-up ANCOVA of speech comprehension, significant main effects were found for BNL [F(2, 224) = 144.62, η_p^2 = 0.56, pp < .001], RT [F(4, 448) = 8.20, η_p^2 = 0.06, p < .001], English proficiency level [F(1, 112) = 64.96, η_p^2 = 0.36, p <.001], and talker accent [F(1, 112) = 52.80, η_p^2 = 0.31, p <.001]. There were also significant two-way interactions between BNL X English proficiency level [F(2, 224) = 3.91, η_p^2 = 0.03, p = .021] and BNL X talker accent [F(2, 224) = 10.93, η_p^2 = 0.08, p <.001]. No significant interactions were found between BNL and RT.

Similar to previous findings, English proficiency level still significantly predicted listeners' speech comprehension performance, when averaged across all BNL and RT

conditions. The linear relation between speech comprehension performance and English proficiency was previously shown in Figure 7.1 in Section 7.2. Listeners with higher English proficiency level were more likely to perform better on the speech comprehension tasks under acoustics, η_p^2 = 0.36, p <.001. Their performance also improved with speech produced by talkers who were native American English talkers (M = 84.7, SD = 1.05) than by native Mandarin Chinese talkers (M = 74.02, SD = 1.02), d = 0.95, p < .001. In addition, listeners performed worse when speech was produced by native Mandarin Chinese talkers than by native American English talkers (d = 0.94, p <.001), as seen in Figure 7.12.

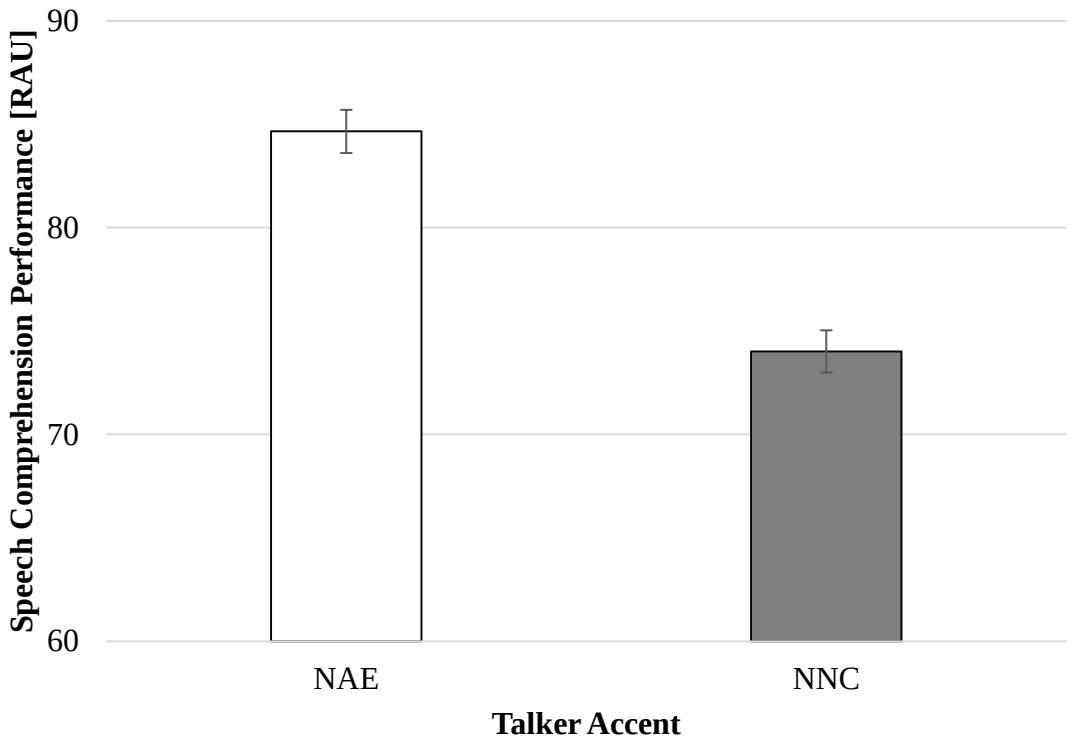


Figure 7.12 - Marginal means of speech comprehension performance, adjusted for standardized English proficiency score at 0. Error bar indicates one standard error.

Planned comparisons using the first condition as the reference level (i.e., RC-30 for BNL and 0.4 second for RT) were conducted for the within-subject acoustic main effects. For BNL, listeners performed significantly better under the RC-30 condition than the RC-40 ($d = 0.26$, $p = .005$) and RC-50 ($d = 1.51$, $p < .001$) conditions, respectively. For RT, comprehension performance was significantly better under the 0.4 second scenario than the 0.8 ($d = 0.35$, $p < .001$), 1.0 ($d = 0.26$, $p = .006$) and 1.2 ($d = 0.43$, $p < .001$) seconds, but not the 0.6 second scenario ($d = 0.05$, $p \geq .62$). The main effects of BNL and RT are illustrated in Figure 7.13 and Figure 7.14, respectively.

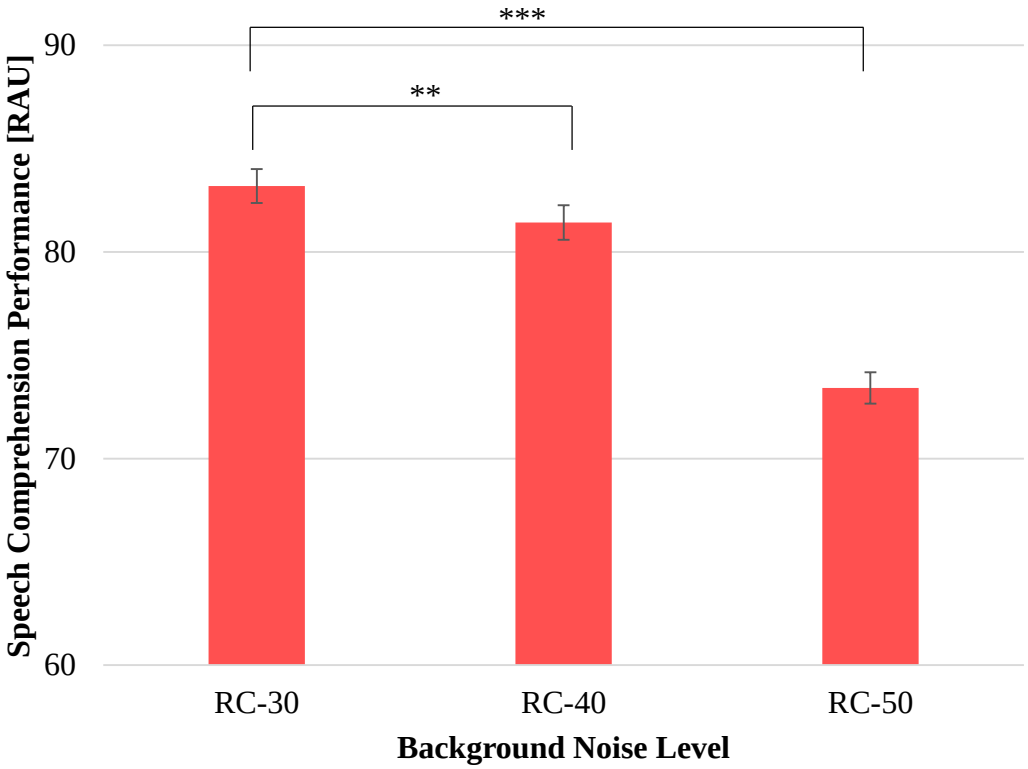


Figure 7.13 - Marginal means of speech comprehension performance on background noise level, adjusted for standardized English proficiency score at 0. Error bars indicate one standard error.

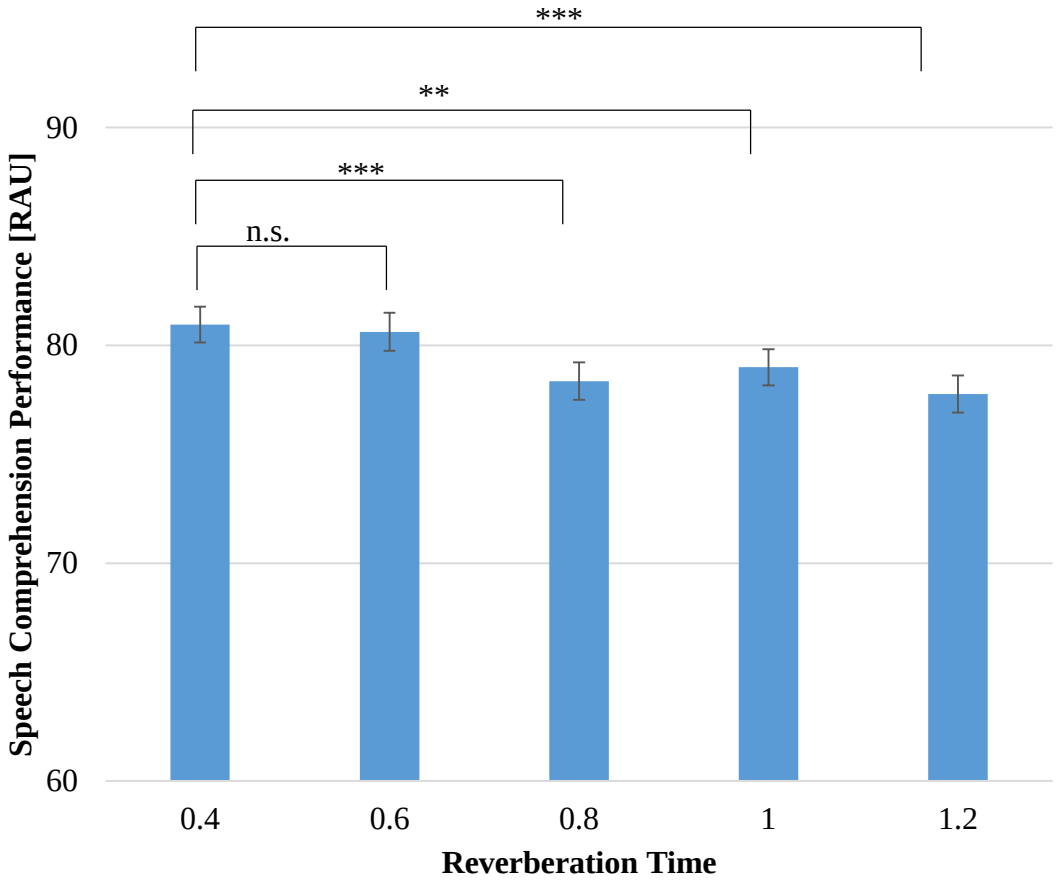


Figure 7.14 - Marginal means of speech comprehension performance on reverberation time, adjusted for standardized English proficiency score at 0. Error bars indicate one standard error.

The significant two-way interaction between BNL X English proficiency level suggests that the performance deficit of speech comprehension between RC-30 and RC-50 was dependent on English proficiency level, p = .01 (see Figure 7.15 and Table 7.3). Listeners with lower English proficiency level experienced greater performance deficit

when exposed to the RC-50 condition. However, such relation was not found for the BNL pair of RC-30 and RC-40, $p \geq .71$.

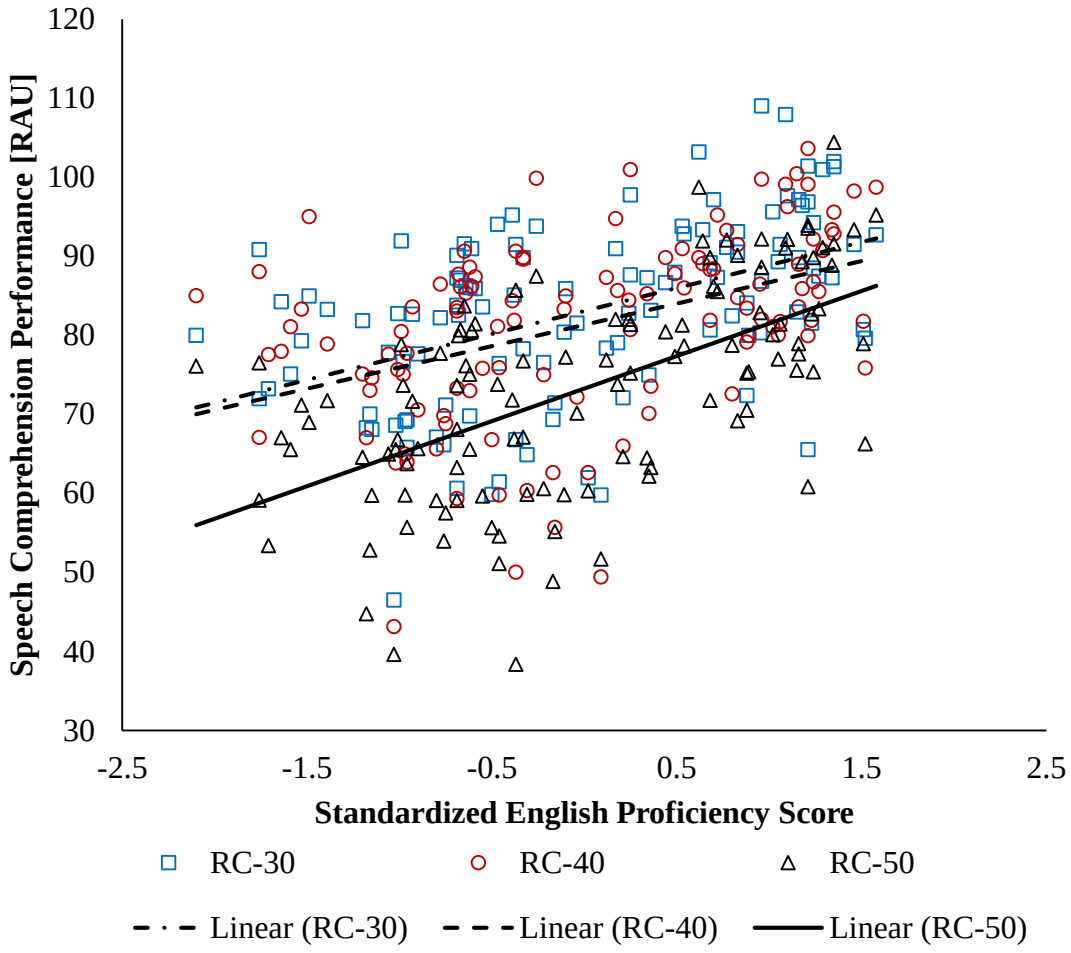


Figure 7.15 - Scatter plot of speech comprehension versus standardized English proficiency score across both Study 1 and 2 for each BNL condition (RC-30, RC-40 and RC-50). Linear regression lines were fitted to each BNL condition.

Table 7.3 - Summary of linear regression lines fitted to the relation between speech comprehension performance and English proficiency level across both Study 1 and 2 for each BNL

BNL	R^2_{adj}	b	SE b	β
RC-30	0.22	5.82	1.00	0.48***
RC-40	0.18	5.39	1.06	0.43***
RC-50	0.34	8.23	1.07	0.59***

Note: ***p < .001

Inference was also drawn from the post hoc analysis on the other significant two-way interaction between BNL X talker accent from all listeners. As shown in Figure 7.16, the performance deficit in foreign-accented speech comprehension was again greater in the RC-50 than the RC-30 BNL condition, $p < .001$, but not between RC-40 and RC-30 conditions ($p \geq .89$). Increased BNL worsened the performance decline in speech comprehension due to foreign accent only at the RC-50 condition.

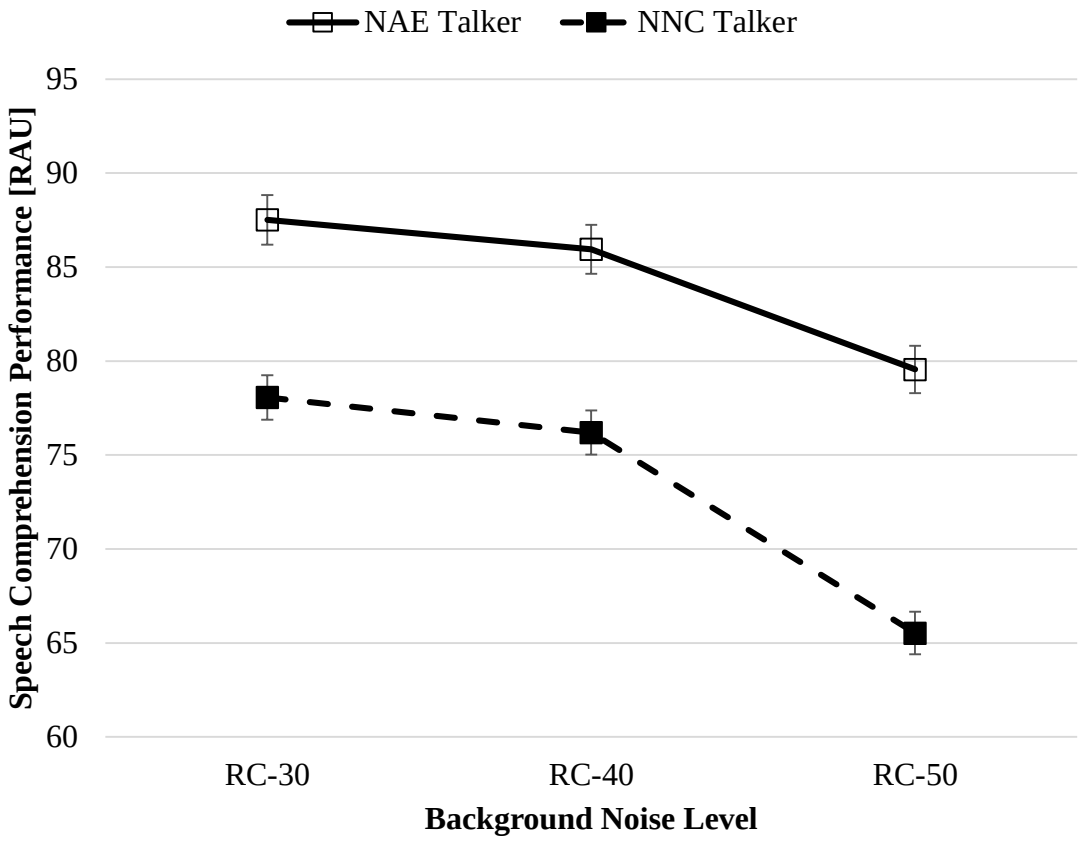


Figure 7.16 - Two-way interaction between BNL and talker accent on speech comprehension performance, adjusted at standardized English proficiency score at 0. Error bar indicates one standard error.

7.3.3 *Subjective Perceptions of Task Workload*

7.3.3.1 *NASA TLX Subscales*

A set of ANOVAs was applied to the individual subscales in the NASA TLX to examine the effect of talker accent in Section 7.3.1.3. The rest of the results are reported herein for BNL, RT, and listener groups (NAE vs. NNC vs. NNO). Results of pairwise comparisons using the Bonferroni correction for BNL using RC-30 as the reference level for individual subscales are listed in Table 7.4.

Mental Demand. There was only one significant main effect for BNL [F(2, 216) = 46.04, η_p^2 = 0.29, p < .001]. Pairwise comparisons reveal that listeners experienced significant increase in mental demand with each step of increase in the BNL conditions.

Physical Demand. There were significant main effects for BNL [F(2, 216) = 20.90, η_p^2 = 0.15, p < .001] and listener group [F(2, 108) = 15.75, η_p^2 = 0.21, p < 0.001]. Pairwise comparisons suggest significant difference in physical demand between RC-50 and the two lower BNL conditions, respectively. Listeners did not find the RC-40 condition to be more physically challenging than the RC-30 condition. The post hoc pairwise comparisons using Tukey's HSD test suggest that NAE listeners reported significantly lower physical demand than NNC (d = 0.96, p < .001) and NNO (d = 0.75, p < .001) listeners, while the non-native listener groups did not vary between each other (d = 0.14, p = .66).

Temporal Demand. There were again significant main effects for BNL [F(2,216) = 19.16, η_p^2 = 0.14, p < .001] and listener group [F(2, 108) = 13.61, η_p^2 = 0.19, p < .001]. Pairwise comparisons show that listeners experienced significantly stronger time pressure under the highest BNL of RC-50 than under RC-30 or RC-40. In addition, NAE listeners reported significantly lower temporal demand than the NNC (d = 0.66, p = .015) and NNO (d = 0.71, p < .001) listeners. The interaction of BNL X listener group was found to be significant [F(4,216) = 2.54, η_p^2 = 0.03, p = .04].

Effort. Significant main effect was found for BNL only [F(2, 216) = 56.68, η_p^2 = 0.34, p < 0.001]. Specifically, participants reported to have worked harder to accomplish the simultaneous tasks with increasing BNL. The significant two-way interactions were BNL X listener group [F(2, 216) = 3.31, η_p^2 = 0.04, p = .012] and talker accent X listener

group [F(2, 108) = 3.62, η_p^2 = 0.05, p = .03]. There was also a significant three-way interaction of BNL X talker accent X listener group [F(4, 216) = 3.65, η_p^2 = .05, p = .007].

Frustration. For this subscale, significant main effects included BNL [F(2, 216) = 73.32, η_p^2 = .40, p < .001], talker accent [F(1, 108) = 7.15, η_p^2 = 0.05, p = .009], and listener group [F(2, 108) = 6.99, η_p^2 = 0.10, p = .001]. While the frustration rating increased significantly with increasing BNL, it was also significantly higher for the non-native listeners [NNC vs. NAE listeners, d = 0.44, p = .003; NNO vs. NAE listeners, d = 0.74, p = .002]. Significant two-way interaction was found for talker accent X listener group [F(2, 108) = 3.55, η_p^2 = 0.04, p = .032] and BNL X talker accent [F(2, 216) = 4.71, η_p^2 = 0.03, p = .010].

Perceived Performance. There were significant main effects for BNL [F(2, 108) = 48.58, η_p^2 = 0.30, p < .001], RT [F(4, 432) = 5.54, η_p^2 = 0.04, p < .001], and talker accent [F(1, 108) = 8.20, η_p^2 = 0.06, p = .005]. The findings of the main effects were similar to the objective performance of speech comprehension, although listener group was non-significant in the perceived performance measure (p = .33). There were also significant two-way interactions for talker accent X listener group [F(2, 108) = 3.73, η_p^2 = 0.05, p = .027] and BNL X talker accent [F(2, 216) = 3.30, η_p^2 = 0.02, p = .039].

Table 7.4 - Pairwise comparisons of background noise level conditions for the NASA TLX subscales

	RC-30 vs. RC-40		**RC-30 vs. RC-50**		**RC-40 vs. RC-50**	
	p-value	**d**	**p-value**	**d**	**p-value**	**d**
Mental Demand	**0.036**	**0.24**	**<.001**	**0.78**	**<.001**	**0.64**
Physical Demand	0.562	0.12	**<.001**	**0.51**	**<.001**	**0.45**
Temporal Demand	0.314	0.15	**<.001**	**0.57**	**<.001**	**0.42**
Effort	**0.016**	**0.27**	**<.001**	**0.89**	**<.001**	**0.69**
Frustration	**0.014**	**0.27**	**<.001**	**0.27**	**<.001**	**0.88**
Perceived Performance	0.379	0.14	**<.001**	**0.75**	**<.001**	**0.71**

Note: Bonferroni corrections applied for the pairwise comparisons. Bold values indicate statistical significant results.

In summary, the threshold of significant perceptual degradation occurred between RC-30 and RC-40 for BNL for half of the subscales in NASA TLX, including mental demand, effort, and frustration ratings. The other subscales had significant degradation between RC-40 and RC-50, as shown in the above table. Similar to previous findings, subjective perceptions were generally not sensitive to RT, except for perceived performance. In comparison to NAE listeners, non-native listeners (both NNC and NNO) reported feeling the dual tasks as more physically challenging, under more time pressure, and more frustrating. Furthermore, the significant two-way interaction between BNL X listener groups for temporal demand and effort rating suggest that such perceptual

degradation due to BNL differed between listener groups, as shown in Figure 7.17 and Figure 7.18. Another significant two-way interaction between BNL X talker accent for frustration and perceived performance rating suggest the degradation due to talker foreign accent was more severe under higher BNL, as shown in Figure 7.19.

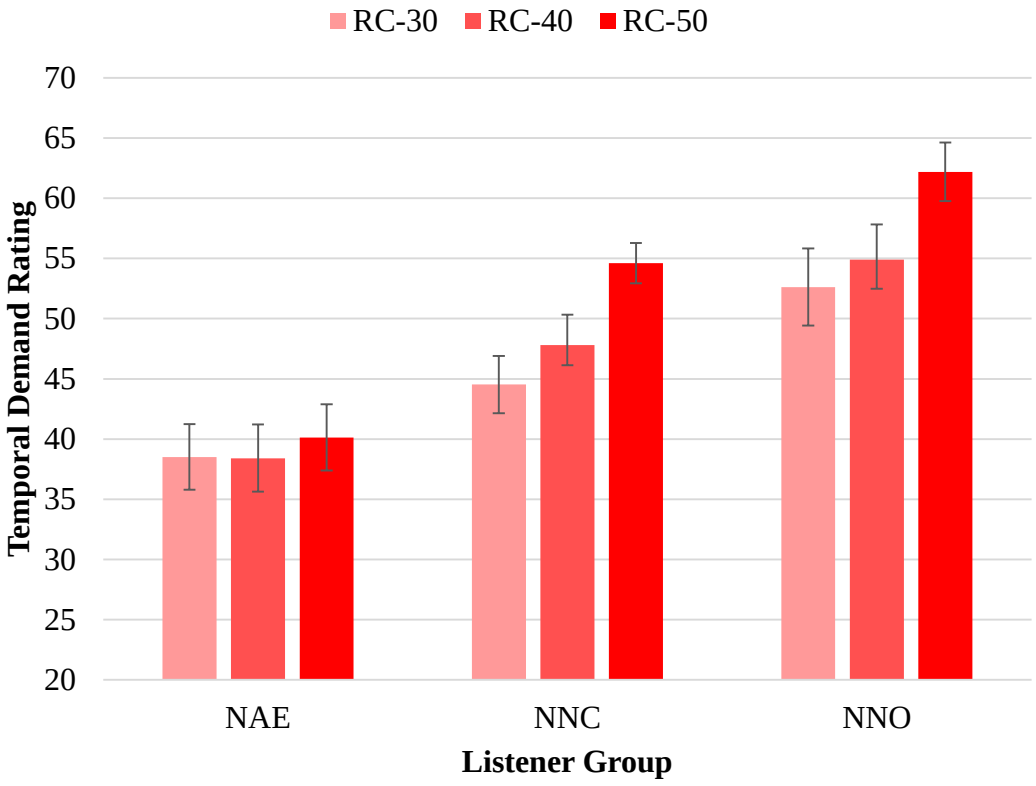


Figure 7.17 - Two-way interaction between BNL and listener group on temporal demand rating in NASA TLX. Error bar indicates one standard error.

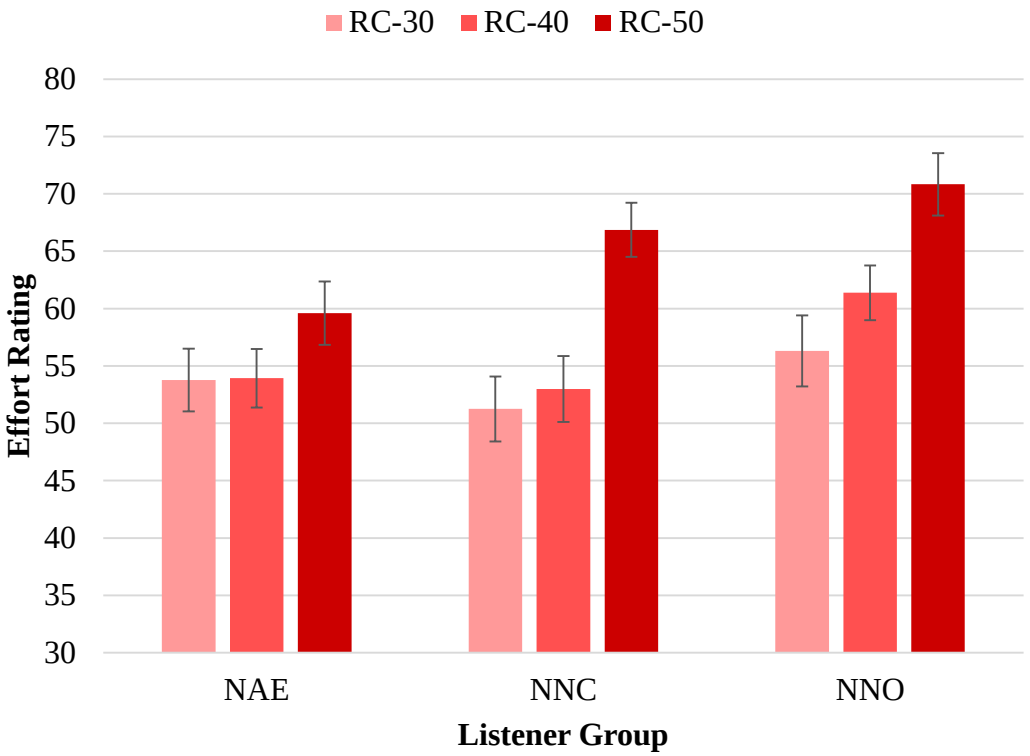


Figure 7.18 - Two-way interaction between BNL and listener group on effort rating in NASA TLX. Error bar indicates one standard error.

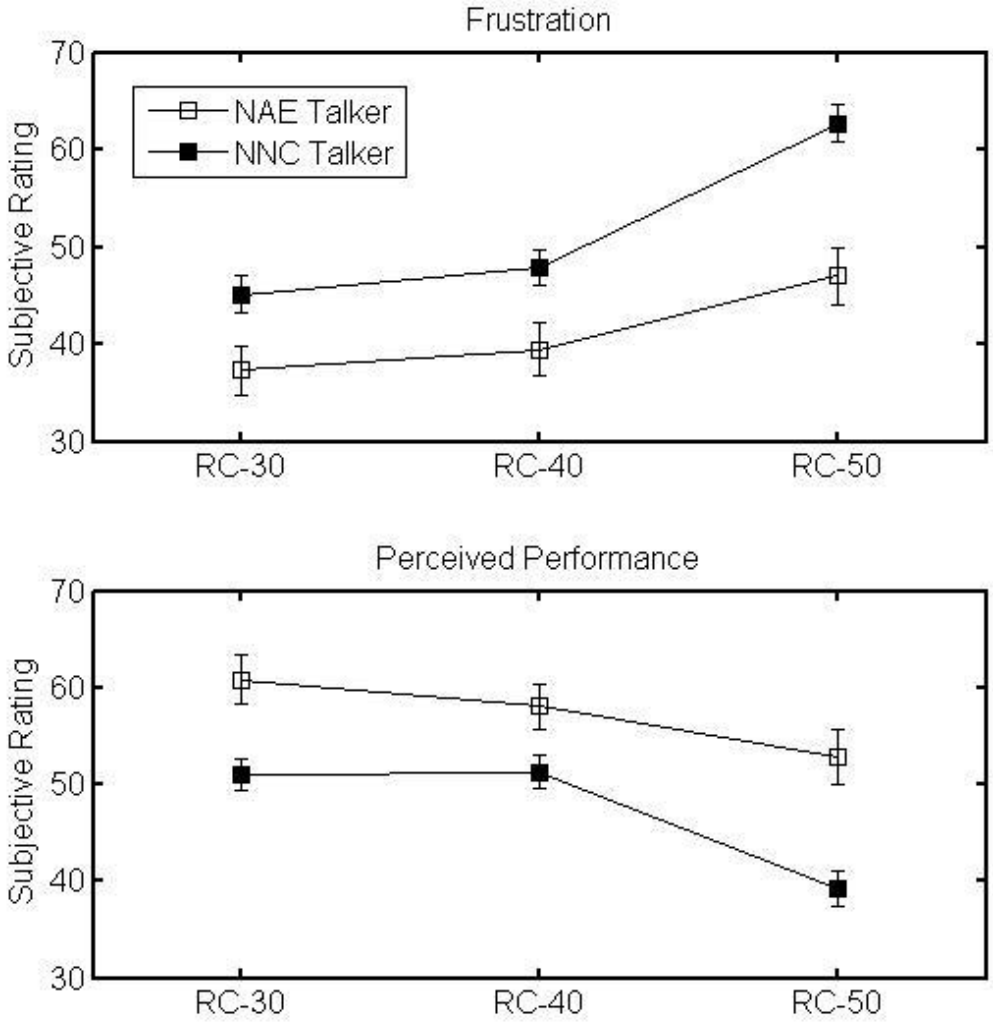


Figure 7.19 - Two-way interaction between BNL and talker accent for frustration and perceived performance ratings in NASA TLX. Error bar indicates one standard error.

7.3.3.2 *Relating Subjective Perception with Objective Performance under Acoustics*

In Section 7.3.2, the significant performance deficit in speech comprehension was identified beyond RC-30 for BNL and 0.6 second for RT. To further support these levels as the design thresholds, it was necessary to verify them against the perceived

performance obtained through NASA TLX. A similar approach from Study 1 was utilized to relate subjective perception with objective perofrmance.

The perceived performance measure was obtained at the end of each test of acoustic condition using the dual-tasks of speech comprehension and APR dot-tracing. To delineate the possibility that listeners providing ratings of the perceived performance based on both tasks, a partial correlation was performed to examine the relations among the two performance measures, while controlling for English proficiency level. As seen in Table 7.5, the partial correlation coefficients show that perceived performance was only significantly correlated with speech comprehension performance, but not with APR dot-tracing performance.

Table 7.5 - Coefficients of partial correlation between subjective perception and performance measures

Measure	Mean	SD	1	2
1. Perceived Performance (NASA TLX subscale)	47.93	20.49	-	-
2. Speech Comprehension Performance (in RAU)	79.2	15.21	0.37***	-
3. APR Dot-tracing Performance (in RPM)	4.34	2.54	0.03	0.18***

Note: N = 1725. Standardized English proficiency level as control variable.
*** $p < .001$

Similar with previous findings (see Study 1 in Chapter 5), listeners from both studies reported perceived performance solely based on the speech comprehension tasks. It can be concluded that the NASA TLX subscale of perceived performance is a measure of listeners' perception of their performance in the speech comprehension tasks.

In order to examine the effects of acoustics and talker accent, a mixed-design ANCOVA was fitted to the perceived performance measure. The within-subject independent variables were BNL and RT, while the between-subject variables were talker accent and English proficiency level. Results revealed significant main effects for BNL [$F(2, 224) = 47.80$, $\eta_p^2 = 0.29$, $p < .001$], RT [$F(4, 224) = 5.09$, $\eta_p^2 = 0.03$, $p = .001$], and talker accent [$F(1, 112) = 11.10$, $\eta_p^2 = 0.09$, $p = .001$]. There was a significant interaction between BNL X talker accent [$F(2, 224) = 4.11$, $\eta_p^2 = 0.03$, $p = .018$]. Interestingly, English proficiency level was not a significant predictor of perceived performance on speech comprehension tasks ($\eta_p^2 = 0.03$, $p \geq .069$), even though the actual comprehension performance was strongly dependent on it.

Listeners' perceived performance was sensitive to talker accent. Similar to their actual objective performance, listeners reported feeling less successful in completing the speech comprehension tasks when the speech was produced by NNC talkers than by NAE talkers, $d = 0.43$, $p = .001$, as seen in Figure 7.20.

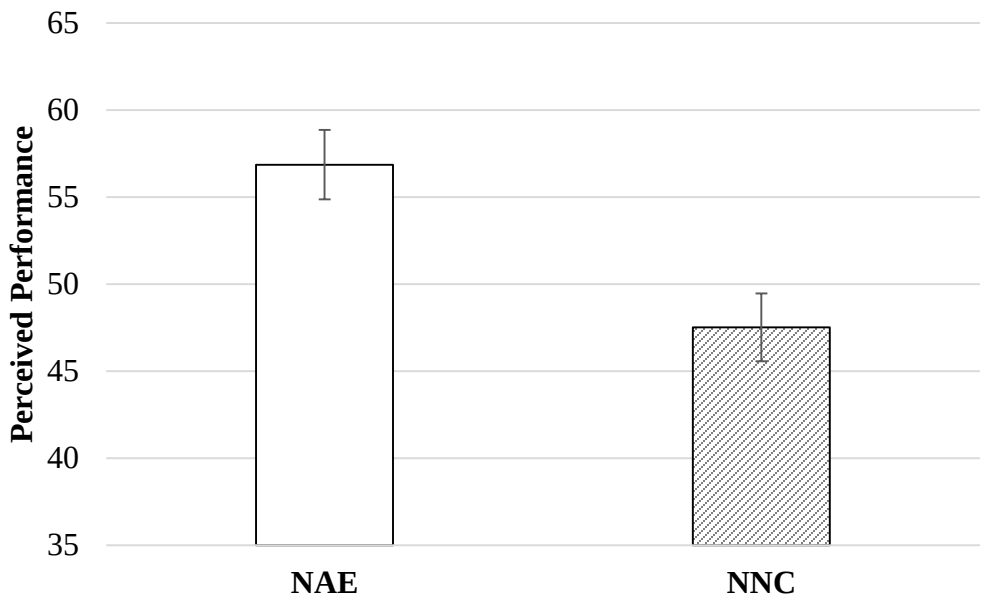


Figure 7.20 - Perceived comprehension performance of speech produced by native American English (NAE) talkers and native Mandarin Chinese talkers (NNC), adjusted at standardized English proficiency score at 0. Error bar indicates one standard error.

For the BNL main effect, planned comparisons using the RC-30 condition as the reference level suggested that listeners reported significantly lower perceived performance rating under the RC-50 ($d = 0.74$, $p < .001$) but not the RC-40 ($d = 0.14$, $p \geq .15$) condition. For RT, the higher RT scenarios were compared against the 0.4 second scenarios. Only the scenarios of 0.8 second ($d = 0.22$, $p = .012$) and 1.2 seconds ($d = 0.30$, $p = .02$) resulted in significantly lower perceived performance rating, but not for the 0.6 second ($d = 0.08$, $p \geq .39$) or 1.0 second ($d = 0.08$, $p \geq .41$) scenarios. The main effects of BNL and RT in this model are illustrated in Figure 7.21 and Figure 7.22, respectively.

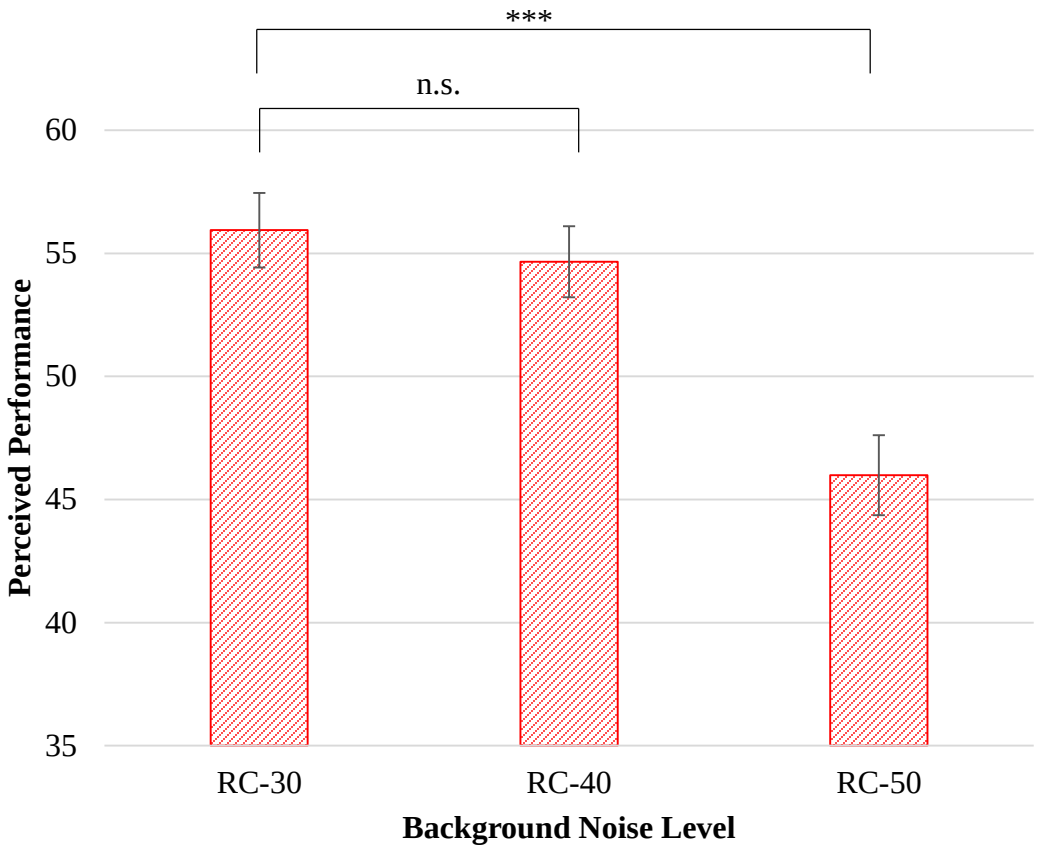


Figure 7.21 - Relation between perceived performance and background noise level, adjusted at standardized English proficiency score at 0. Error bar indicates one standard error.

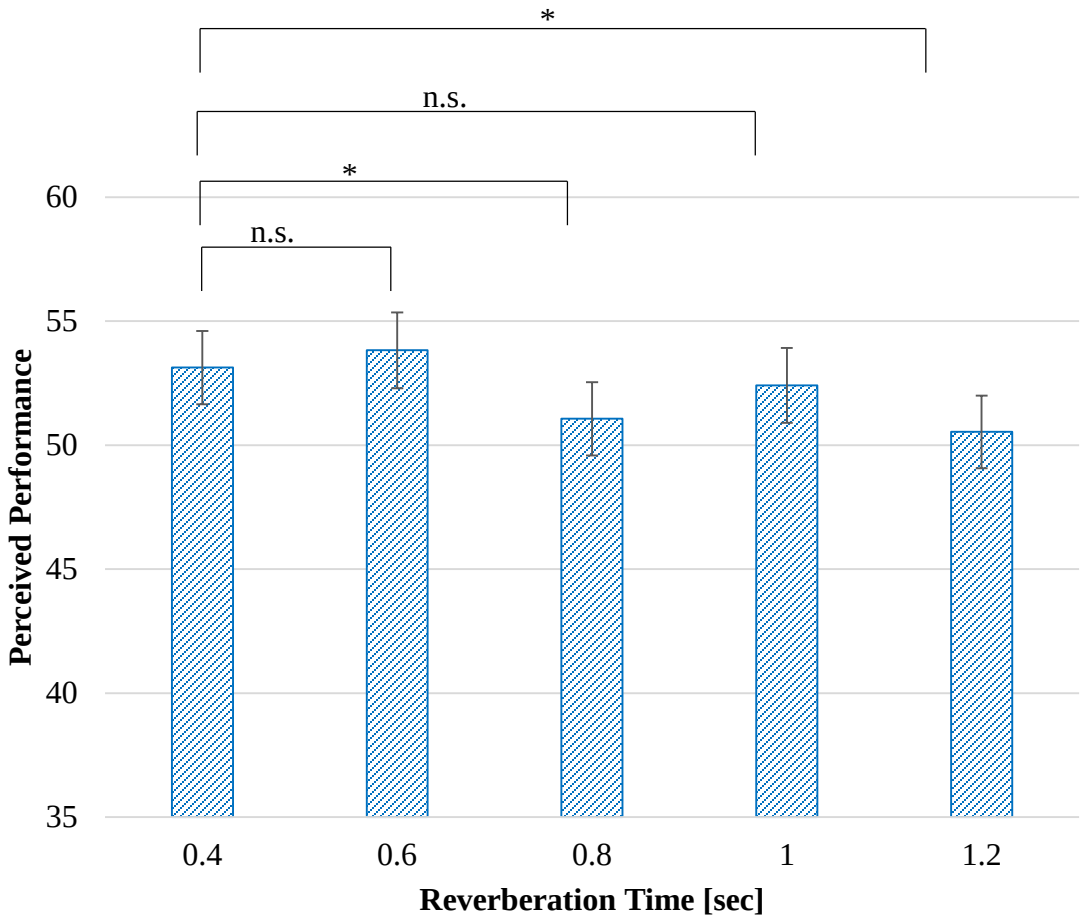


Figure 7.22 - Relation between perceived performance and reverberation time, adjusted at standardized English proficiency score at 0. Error bar indicates one standard error.

For the significant interaction between BNL and talker, planned comparisons using the RC-30 as the reference level did not reveal significant perception degradation between either RC-30 and RC-40 ($p = .077$) or RC-30 and RC-50 ($p = .21$). As seen in Figure 7.23, the significant degradation may exist between RC-40 and RC-50.

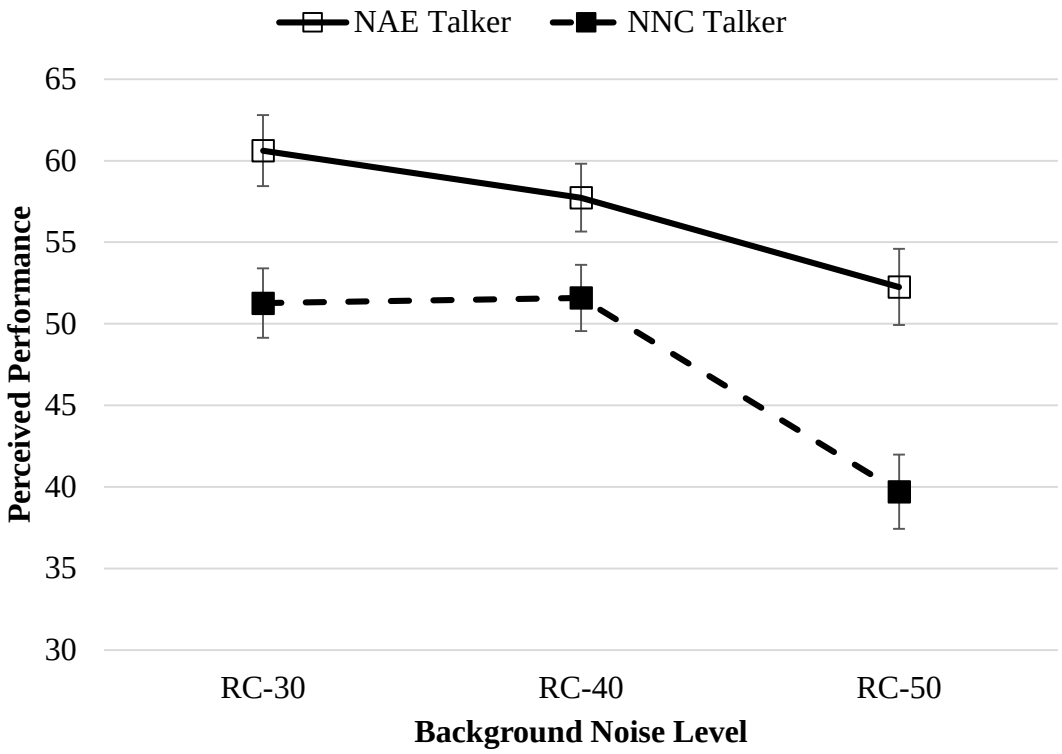


Figure 7.23 - Two-way interaction between BNL and talker accent on perceived performance rating, adjusted for standardized English proficiency score at 0. Error bar indicates one standard error.

7.4 Discussion on Confounding Factors

During the initial screen and main experiment, more variables were collected from listener participants than reported so far in the statistical analyses. Some variables, particularly those that were not descriptive, were worthwhile investigating for their abilities in confounding the observed effects of BNL and RT on listeners' speech comprehension performance. Screenings for potential confounding factors were most meaningful when performed on the combined dataset, which included data from all listener participants in this research. Although none of the potential confounders

discussed in this section was accepted, they were each screened based on the relevance to the research questions and the ability to improve the omnibus model. The omnibus model included BNL, RT and standardized English proficiency score as independent variables and speech comprehension performance as the dependent variable in an ANCOVA model. For some factors, the dataset was re-arranged to construct an omnibus regression model for the screening.

Adding independent variables (IVs) would always increase (or at least maintain) the total variance explained by the omnibus model in the dependent variable (DV). The stronger the strength of a unique IV, the more likely it would become statistically significant. The initial selection of IVs should be based on the research hypothesis and aim for parsimonies in the omnibus models. However, in the empirical screening of confounders in the steps listed above, the backward approach was adopted to provide an opportunity to amend the research hypothesis if strong evidences were identified from the statistical testing.

7.4.1 *Various Potential Confounders*

The results of the potential confounders are summarized in Table 7.6. The objective performance of speech comprehension was not affected by gender, handedness, or test chamber temperature. Although listeners provided self-report ratings of noise sensitivity in three domains of daily life, none of the sensitivity rating significantly predicted the speech comprehension performance beyond and above the acoustic factors and English proficiency level. Since the NoiSeQ-R utilized a 4-point ordinal scale, it is too early to conclude a relation between objective performance and baseline noise sensitivity. More investigation is needed for the relevance of noise sensitivity on

objective performance related to learning outcomes. Furthermore, the average time listeners took to respond to the test items during the speech comprehension tasks did not affect the performance either, which was expected if listeners had invested enough effort in the tasks.

Two confounding factors were found statistically significant when added to the omnibus model for test duration of the speech comprehension tasks and percent on-target of the APR dot-tracing task. However, the effect size of test duration was very small. It only significantly predicted half of a percent more of the overall variance in the speech comprehension performance. The inclusion of test duration in the omnibus model did not substantially change the results of the acoustic factors or English proficiency level. It was hence not included for further investigation. For the omnibus model on the APR dot-tracing task performance in RPM, percent on-target predicted much more than BNL, RT and English proficiency level all together. In this dissertation, the APR dot-tracing task served as a secondary distraction task. The performance on this task was less relevant to answering the research questions outlined in Chapter 1. For future work on dual-task paradigm using an adaptive dot-tracing task, it may be worthwhile to control for percent on-target if its performance is of interest.

Table 7.6 - Summary of confounder effects in omnibus model

Omnibus Regression Model			
Factor	**ΔF**	**Sign. Level**	**ΔR^2**
Temperature	2.94	.083	0.001
Test Duration	**10.96**	**.001**	**0.005**
Response Time	0.14	.71	< 0.001
% On-Target (DV = RPM)	**630.35**	**<.001**	**0.253**
Omnibus ANCOVA Model			
Factor	**ΔF**	**Sign. Level**	**η_p^2**
Gender	3.48	.065	0.03
Handedness	3.1	.08	0.03
NoiSeQ - Sleep	0.2	.65	0.002
NoiSeQ - Work	0.93	.34	0.008
NoiSeQ - Residential Surrounding	0.14	.71	0.001
NoiSeQ - Overall	0.08	.78	0.001

7.4.2 *Measures of English Proficiency*

Several alternative measures of English proficiency were investigated for their efficiency in predicting the speech comprehension performance in this dissertation. The alternative measures were individually included in an omnibus ANCOVA model to replace the standardized English proficiency score. The test statistics and effect sizes are summarized in Table 7.7.

Interestingly, all alternative measures of English proficiency identified in this dissertation were statistically significant and did not substantially change the results of other factors in the omnibus model. The self-report items in LEAP-Q for “English as first

language acquired" and "English as currently in dominance" shared similarly large effect sizes, suggesting equivalent predictability as a measure of English proficiency. The "English as first language acquired" item was utilized in this dissertation to categorize participants into different listener groups. Although not as strong a predictor as the composite scale of English proficiency tested in this dissertation, it provided plausible prediction and can be considered for future use if the English proficiency tests are not available. Also from the LEAP-Q was the "Month in English-dominant country" as an alternative ordinal instead of dichotomous measure of English dominance. But it was in fact less efficient in predicting the speech comprehension performance. Lastly, the additional dichotomous item of "ever dreamed in English" also provided some ability in explaining the comprehension performance. Taken together, English proficiency was best described by using the composite scale from the three tests administered during initial screening in this dissertation. The composite scale achieved high reliability and best predictability among all alternative measures in the omnibus model involving BNL and RT. Inclusion of other measures of English proficiency such as listener group, English dominance or immersion was not considered since they were redundant measures of the same construct.

Table 7.7 - Summary of alternative measures of English proficiency

Factor	F	Sign. Level	η_p^2
Standardized English Proficiency Score	56.91	< .001	0.34
English as first language acquired (Native vs. Non-native Listener)	35.41	<.001	0.24
English as currently in dominance	31.21	<.001	0.22
Month in English-dominant country	16.47	<.001	0.13
Ever dreamed in English	3.91	.011	0.10

7.5 Summary and Conclusions

By combining data from Study 1 and 2, this chapter examined the comprehensive effects of BNL, RT, and talker accent on the objective performance and subjective perception of speech comprehension by listeners from three groups: 1) native American English-speaking (NAE), 2) native Mandarin Chinese-speaking (NNC), and 3) other non-native English-speaking (NNO).

Previously found in speech intelligibility tasks by Bent and Bradlow (2003), the interlanguage benefit of matched accent was also identified in speech comprehension tasks that involve higher level of language processing. Non-native listeners who shared the same accent as the foreign talkers achieved better comprehension performance and were less negatively affected by both BNL and RT than their non-native counterparts with mismatched accent from the talkers. The matched accent not only provided advantages for non-native listeners on the actual speech comprehension performance, but also benefited them in the perception of task performance. While their non-native counterparts and native listeners reported feeling more frustrated and worse perceived

task performance when foreign accent was introduced, the matched-accent non-natives reported no significant difference on the two perception measures.

In addition to the effect of talker foreign accent, the effects of BNL and RT on the objective performance and subjective perception of speech comprehension were also carefully examined with individual listeners' English proficiency level controlled. Consistent with previous studies, the general trend of better performance and higher perception rating was found for lower BNL and shorter RT conditions within the range investigated in this dissertation. The design thresholds of these two acoustic metrics were identified based on the level beyond which significant performance deficit and perception degradation (as compared against the lowest BNL or RT) were observed. For speech comprehension performance, the design thresholds were identified at RC-30 BNL and 0.6 second RT, respectively. For the perceived performance rating of the comprehension tasks, the design thresholds were identified at RC-40 BNL and 0.6 second RT, respectively.

Furthermore, listeners experienced more negative impact of increasing BNL with speech produced by foreign-accented talkers than by native American English talkers. By comparing the design thresholds identified in Study 1, the addition of talker foreign accent required a more stringent design condition of BNL, which could be possibly as much as 10 dB lower.

In speech perception under realistic room acoustic conditions, there were concerns about increased BNL due to the slower decay of sounds levels of the running speech in an environment with long RT (Bradley *et al.*, 1999). Although physically related, the lack of statistical significant interaction disentangled BNL and RT in terms of

speech comprehension performance. It further suggests that the design of these two acoustic metrics should be conducted separately. The design level of BNL (from mechanical equipment only) or RT should not be regarded as compensation for each other. Instead, the design decision should be determined based on the classroom occupants, whether non-native English speakers are part of the talkers or listeners.

Chapter 8 – Conclusions

8.1 Summary of Findings and Conclusions

In this dissertation, the effects of background noise level (BNL), reverberation time (RT), and talker foreign accent on speech comprehension by native and non-native English-speaking listeners have been studied extensively. Using laboratory-controlled experiments, a total of 15 acoustic conditions comprised of three conditions of BNL (RC-30, 40, and 50) and five scenarios of RT (from 0.4 to 1.2 seconds) were created to simulate realistic classroom acoustic environments. To measure listeners' performance when exposed under the assorted acoustic conditions, a dual-task paradigm was adopted for testing speech comprehension and the adaptive pursuit rotor (dot-tracing) tasks simultaneously. The design criteria of BNL and RT were identified beyond which listeners began to experience significant performance deficit on the speech comprehension tasks. The listeners' objective performance of speech comprehension was further complemented by their self-report perception of task performance from the NASA Task Load Index (TLX). Good agreement was found between the objective performance and subjective perception measures.

In Study 1, listeners performed worse under higher BNL and longer RT in comprehending speech from native American English talkers. In general, BNL was more detrimental to listeners with lower English proficiency level. The design thresholds of classroom acoustics were identified at RC-40 BNL and 0.6 second RT, beyond which significant performance deficits were observed. When the speech was free from foreign accent, the detrimental effects of both BNL and RT were more pronounced for non-native listeners than for native listeners. Furthermore, while non-native listeners experienced

equivalently negative impacts of BNL and RT, native listeners were able to overcome such impact for RT but not for BNL. The perceived performance rating from NASA TLX showed similar trends, with the significant perception degradation occurring beyond RC-30 BNL and 0.6 second RT.

In Study 2, a similar trend of performance deficit under higher BNL and longer RT was observed when the same speech materials were produced by native Mandarin Chinese talkers, who shared similar and moderate degree of accentedness. Three groups of listeners were recruited, including native American English speakers (NAE), native Mandarin Chinese speakers (NNC), and non-native English-Chinese speakers (NNO). The interlanguage benefit of matched accent was observed where the NNC listeners, although least proficient in English among three groups, scored significantly higher on speech comprehension performance than the NNO listeners. The design thresholds of classroom acoustics were identified at RC-30 for BNL and 0.6 second RT.

Combining data from Study 1 and 2 enabled the investigation of the effect of talker foreign accent under assorted acoustic conditions. First, the interlanguage benefit of matched accent was further confirmed. It alleviated the negative impacts of BNL and RT for the NNC listeners on speech comprehension, who scored lowest on the English proficiency tests as a group. In addition, it also prevented the NNC listeners from feeling more frustrated and less successful in task completion, both of which were pronounced among NNO and NAE listeners. Second, BNL was even more detrimental when foreign accent was introduced. Using the comprehensive dataset, the design criteria were again identified from speech comprehension performance at RC-30 BNL and 0.6 second RT. And these were also supported by the subjective perception of task performance from the

NASA TLX. Interestingly, the interaction between BNL and RT was never found to be statistically significant, suggesting independence between the two acoustic metrics on objective performance and subjective perception. In other words, meeting or even exceeding the requirements in one acoustic metric would not be able to compensate deficiencies in the other metric. The designs of BNL and RT should be carried out separately.

In conclusion, room acoustic design should be conscious of the linguistic diversity among occupants in the classroom. Depending on whether non-native English speakers exist among listeners and talkers, more stringent acoustic requirements may be necessary to attain optimal speech comprehension performance. From the findings in this dissertation, the recommended design thresholds of BNL and RT are summarized in Table 8.1.

Table 8.1 - Design guidelines of BNL and RT depending on the English nativeness of talker and listener occupants in the classroom

	Native English *Talkers* Only	**Both Native and Non-native English *Talkers***
Native English *Listeners* Only	BNL ≤ RC-40 (48 dBA) RT ≤ 1.2 second	BNL ≤ RC-30 (38 dBA) RT ≤ 1.2 second
Both Native and Non-native English *Listeners*	BNL ≤ RC-40 (48 dBA) RT ≤ 0.6 second	BNL ≤ RC-30 (38 dBA) RT ≤ 0.6 second

8.2 Future Work

This dissertation identified design guidelines of BNL and RT to supplement the existing classroom acoustics standard using a more relevant measure of speech comprehension to represent learning outcomes and a more representative sample of occupants involved in the classroom activities. Future work can still be completed to further improve the classroom acoustic design guidelines.

First, *in situ* testing in an actual classroom is necessary to generalize conclusions from this work, which used strictly controlled laboratory conditions. Although simulated to closely approximate realistic classroom environments, test conditions in this investigation were only created for a single listener position in the classroom (i.e., 4 m in front of the talker). Even though they may result in very similar physical measurements, the conclusions of BNL and RT should be verified at other listener positions, particularly in the back of the room with lower resulting SNR and on the side of the room with lower interaural cross-correlation due to the proximity to a reflecting surface. Second, further investigation may involve testing even lower levels and finer intervals of BNL and RT. The testing of lower levels of BNL (i.e., below RC-30 or 38 dBA) and RT (i.e., below 0.4 second) can confirm whether there exists additional benefits on speech comprehension performance. The design guidelines will also benefit from using even finer intervals well within the just-noticeable-difference (JND) of the acoustic test conditions to identify thresholds of performance deficit, such as intervals of less than 3 dBA in BNL and of 0.1 second in RT. Third, only general guidelines of BNL and RT are recommended in this work. These recommendations can be studied further by investigating their effects of spectral and temporal masking on speech due to different talker characteristics such as

gender (male vs. female), age (young vs. elderly), foreign accent (e.g., degree of phonetic similarity between English and the foreign language), and speech style (e.g., clear vs. conversational). As a result of this work using native Mandarin Chinese talkers to produce foreign-accented speech, the effect of foreign accent is of particular interest to study another foreign language with similar phonetic characteristics with English.

In addition, the test material in measuring speech comprehension in this dissertation was limited to trivial knowledge, such as casual conversations and simple informative paragraphs. Realistic activities in a classroom, though, involve learning new concepts, which was not included in the scope of this research. Furthermore, talkers in both studies recorded the speech materials in ideal acoustic environments with very low ambient noise and free from distractions. Room acoustic effects did not contribute to the deterioration in their speech production, which could occur in the interaction between talker and listener in a realistic classroom. Three directions for future work are summarized below.

1) How do realistic room acoustic conditions affect knowledge gain during lecture-style learning?

Performances of both speech intelligibility (previous research) and speech comprehension (current research) were found to be impeded under adverse acoustic environments, setting a trajectory of research into investigating even higher level of information processing during learning. To provide more evidence to further improve the classroom acoustics standards, the next step of investigation can be oriented to examine

knowledge gain (e.g., conceptual and procedural) obtained via oral instructions in realistic room acoustic environments.

2) What is the role of realistic room acoustics on the top-down versus bottom-up processes during speech comprehension?

In comprehending speech in reverberant environments, the advantage that native English-peaking listeners have over non-native listeners seems to suggest that the top-down process may compensate for the degraded speech signals. If the two processes in speech comprehension can be separated, the effect of reverberation on the individual processes can be studied further to provide implications on designing room acoustics for populations with special education needs perhaps even beyond non-native English-speaking listeners.

3) How do realistic room acoustic conditions affect speech production and ultimately speech comprehension by non-native English-speaking listeners?

Classroom learning is an interactive process involving both the talker and listener simultaneously in the room acoustics environment. It may be worth investigating the mediating effect of talkers' speech pattern on the effects of room acoustics that further contribute to speech comprehension by the listeners, particularly when both parties are non-native English speakers. A conceptual illustration of such relations is included in Figure 8.1.

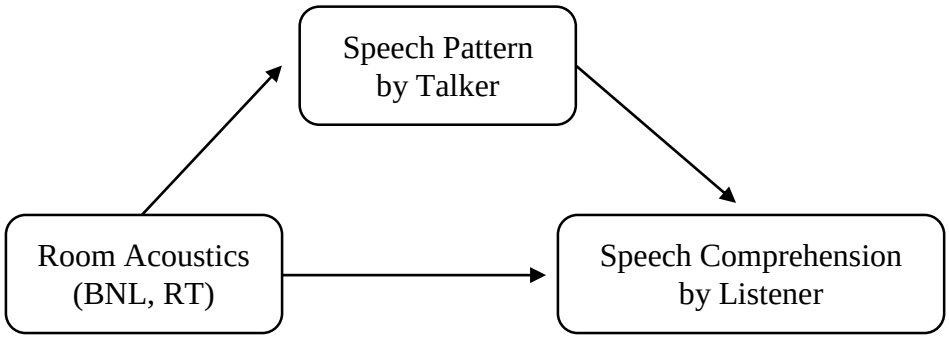


Figure 8.1 - Conceptual illustration of effects of room acoustics on the interactive process of speech production and comprehension

References

American National Standard Institute (**2010**). "ANSI/ASA S12.60 Performance Criteria, Design Requirements, and Guidelines for Schools, Part 1: Permanent Schools."

American National Standard Institute (**2012**). "ANSI/ASA S3.5-1997 (r2012), Methods for the calculation of the speech intelligibility index."

Aud, S., Hussar, W., Planty, M., Snyder, T., Bianco, K., Fox, M., Frohlich, L., Kemp, J., and Drake, L. (**2010**). "The condition of education 2010 (NCES 2010-028)," (US Department of Education, Institute of Education Sciences, Washington, DC).

Bamford, J., and Wilson, I. (**1979**). "Methodological considerations and practical aspects of the BKB sentence lists," Speech-hearing tests and the spoken language of hearing-impaired children, 148-187.

Beaman, C. P., and Holt, N. J. (**2007**). "Reverberant auditory environments: the effects of multiple echoes on distraction by 'irrelevant'speech," Applied Cognitive Psychology **21**, 1077-1090.

Bench, J., Kowal, Å., and Bamford, J. (**1979**). "The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children," British Journal of Audiology **13**, 108-112.

Bent, T., and Bradlow, A. R. (**2003**). "The interlanguage speech intelligibility benefit," J. Acoust. Soc. Am. **114**, 1600.

Bent, T., Kewley-Port, D., and Ferguson, S. H. (**2010**). "Across-talker effects on non-native listeners' vowel perception in noise," J Acoust Soc Am **128**, 3142-3151.

Bradley, J. S. (**1986**). "Speech intelligibility studies in classrooms," J. Acoust. Soc. Am. **80**, 846.

Bradley, J. S. (**2011**). "Review of objective room acoustics measures and future needs," Applied Acoustics **72**, 713-720.

Bradley, J. S., Reich, R. D., and Norcross, S. G. (**1999**). "On the combined effects of signal-to-noise ratio and room acoustics on speech intelligibility," J. Acoust. Soc. Am. **106**, 1820.

Bradley, J. S., and Sato, H. (**2008**). "The intelligibility of speech in elementary school classrooms," J. Acoust. Soc. Am. **123**, 2078.

Bradley, J. S., Sato, H., and Picard, M. (**2003**). "On the importance of early reflections for speech in rooms," J. Acoust. Soc. Am. **113**, 3233.

Bradlow, A. R., and Alexander, J. A. (**2007**). "Semantic and phonetic enhancements for speech-in-noise recognition by native and non-native listeners," J. Acoust. Soc. Am. **121**, 2339-2349.

Bradlow, A. R., and Bent, T. (**2002**). "The clear speech effect for non-native listeners," J. Acoust. Soc. Am. **112**, 272.

Bradlow, A. R., Kraus, N., and Hayes, E. (**2003**). "Speaking Clearly for Children With Learning Disabilities: Sentence Perception in Noise," Journal of Speech, Language, and Hearing Research **46**, 80-97.

Bradlow, A. R., and Pisoni, D. B. (**1999**). "Recognition of spoken words by native and non-native listeners: Talker-, listener-, and item-related factors," J. Acoust. Soc. Am. **106**, 2074.

Bray, J. H., and Maxwell, S. E. (**1985**). *Multivariate analysis of variance* (Sage).

Bronzaft, A. L. (**1981**). "The effect of a noise abatement program on reading ability," Journal of Environmental Psychology **1**, 215-222.

Bronzaft, A. L., and McCarthy, D. P. (**1975**). "The effect of elevated train noise on reading ability," Environment and Behavior.

Clark, C., Martin, R., Van Kempen, E., Alfred, T., Head, J., Davies, H. W., Haines, M. M., Barrio, I. L., Matheson, M., and Stansfeld, S. A. (**2006**). "Exposure-Effect Relations between Aircraft and Road Traffic Noise Exposure at School and Reading Comprehension The RANCH Project," American Journal of Epidemiology **163**, 27-37.

Cohen, J. (**1977**). *Statistical power analysis for the behavioral sciences (rev* (Lawrence Erlbaum Associates, Inc).

Cohen, J. (**1992**). "A power primer," Psychological bulletin **112**, 155.

Cohen, S., Glass, D. C., and Singer, J. E. (**1973**). "Apartment noise, auditory discrimination, and reading ability in children," Journal of experimental social psychology **9**, 407-422.

Cutler, A., Weber, A., Smits, R., and Cooper, N. (**2004**). "Patterns of English phoneme confusions by native and non-native listeners," J. Acoust. Soc. Am. **116**, 3668.

Daneman, M., and Carpenter, P. A. (**1980**). "Individual differences in working memory and reading," Journal of verbal learning and verbal behavior **19**, 450-466.

Downey, R., Farhady, H., Present-Thomas, R., Suzuki, M., and Van Moere, A. (**2008**). "Evaluation of the usefulness of the Versant for English test: A response," Language Assessment Quarterly **5**, 160-167.

Ellis, N., and Hennelly, R. (**1980**). "A bilingual word ‐ length effect: Implications for intelligence testing and the relative ease of mental calculation in Welsh and English," British Journal of Psychology **71**, 43-51.

Erdfelder, E., Faul, F., and Buchner, A. (**1996**). "GPOWER: A general power analysis program," Behavior research methods, instruments, & computers **28**, 1-11.

Evans, G. W., Bullinger, M., and Hygge, S. (**1998**). "Chronic noise exposure and physiological response: A prospective study of children living under environmental stress," Psychological science **9**, 75-77.

Faul, F., Erdfelder, E., Buchner, A., and Lang, A.-G. (**2009**). "Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses," Behavior research methods **41**, 1149-1160.

Faul, F., Erdfelder, E., Lang, A.-G., and Buchner, A. (**2007**). "G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences," Behavior research methods **39**, 175-191.

Field, A. (**2009**). *Discovering statistics using SPSS* (Sage publications).

Field, A., and Hole, G. J. (**2002**). *How to design and report experiments* (Sage).

Flege, J. E., Yeni-Komshian, G. H., and Liu, S. (**1999**). "Age constraints on second-language acquisition," Journal of memory and language **41**, 78-104.

Griefahn, B. (**2008**). "Determination of noise sensitivity within an internet survey using a reduced version of the Noise Sensitivity Questionnaire," J. Acoust. Soc. Am. **123**, 3449-3449.

Haines, M. M., Stansfeld, S. A., Brentnall, S., Head, J., Berry, B., Jiggins, M., and Hygge, S. (**2001**). "The West London Schools Study: the effects of chronic aircraft noise exposure on child health," Psychological medicine **31**, 1385.

Hair, J. F., Tatham, R. L., Anderson, R. E., and Black, W. (**2006**). *Multivariate data analysis* (Pearson Prentice Hall Upper Saddle River, NJ).

Hak, C., and Wenmaekers, R. (**2013**). "Room in Room Acoustics: Using Convolutions to find the Impact of a Listening Room on Recording Acoustics," International Symposium on Room Acoustics.

Hart, S. G. (**2006**). "NASA-task load index (NASA-TLX); 20 years later," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Sage Publications), pp. 904-908.

Hart, S. G., and Staveland, L. E. (**1988**). "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research," Human mental workload **1**, 139-183.

Hodgson, M., and Nosal, E.-M. (**2002**). "Effect of noise and occupancy on optimal reverberation times for speech intelligibility in classrooms," J. Acoust. Soc. Am. **111**, 931.

Hǿjen, A., and Flege, J. E. (**2006**). "Early learners' discrimination of second-language vowels," J. Acoust. Soc. Am. **119**, 3072.

Hornsby, B. W. Y. (**2004**). "The Speech Intelligibility Index: What is it and what's it good for?," The Hearing Journal **57**, 10-17.

Houtgast, T., and Steeneken, H. (**1984**). "A multi-language evaluation of the RASTI-method for estimating speech intelligibility in auditoria," Acta Acustica united with Acustica **54**, 185-199.

Hoyle, R. H., Harris, M. J., and Judd, C. M. (**2002**). *Research methods in social relations* (Thomson Learning).

Hygge, S., Boman, E., and Enmarker, I. (**2003**). "The effects of road traffic noise and meaningful irrelevant speech on different memory systems," Scandinavian Journal of Psychology **44**, 13-21.

Hygge, S., Evans, G. W., and Bullinger, M. (**2002**). "A prospective study of some effects of aircraft noise on cognitive performance in schoolchildren," Psychological Science **13**, 469-474.

Hygge, S., and Kjellberg, A. (**2010**). "Special issue on noise, memory and learning," Noise and Health **12**, 199.

Iglehart, F. (**2009**). "Combined effects of classroom reverberation and noise on speech perception by students with typical and impaired hearing," in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings* (Institute of Noise Control Engineering), pp. 650-657.

Institute for International Education (**2012**). "Open doors 2012 fast facts."

International Electrotechnical Commission (**2003**). "IEC 60268-16: Sound system equipment-Part 16: Objective rating of speech intelligibility by speech transmission index," IEC, Switzerland.

Judd, C. M., McClelland, G. H., and Ryan, C. S. (**2011**). *Data analysis: A model comparison approach* (Routledge).

Kennedy, S. M., Hodgson, M., Edgett, L. D., Lamb, N., and Rempel, R. (**2006**). "Subjective assessment of listening environments in university classrooms: Perceptions of students," J. Acoust. Soc. Am. **119**, 299.

Klatte, M., Hellbrück, J., Seidel, J., and Leistner, P. (**2010a**). "Effects of classroom acoustics on performance and well-being in elementary school children: A field study," Environment and Behavior **42**, 659-692.

Klatte, M., Lachmann, T., and Meis, M. (**2010b**). "Effects of noise and reverberation on speech perception and listening comprehension of children and adults in a classroom-like setting," Noise & health **12**, 270-282.

Knecht, H. A., Nelson, P. B., Whitelaw, G. M., and Feth, L. L. (**2002**). "Background noise levels and reverberation times in unoccupied classrooms: Predictions and measurements," American Journal of Audiology **11**, 65.

Levine, T. R., and Hullett, C. R. (**2002**). "Eta squared, partial eta squared, and misreporting of effect size in communication research," Human Communication Research **28**, 612-625.

Ljung, R., and Kjellberg, A. (**2009**). "Recall of spoken words presented with a prolonged reverberation time," Build Acoust **16**, 301-312.

Luce, P. A., and Pisoni, D. B. (**1998**). "Recognizing spoken words: The neighborhood activation model," Ear and hearing **19**, 1.

Mackay, I. R., and Flege, J. E. (**2004**). "Effects of the age of second language learning on the duration of first and second language sentences: The role of suppression," Applied Psycholinguistics **25**, 373-396.

Marian, V., Blumenfeld, H. K., and Kaushanskaya, M. (**2007**). "The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language

profiles in bilinguals and multilinguals," Journal of Speech, Language and Hearing Research **50**, 940.

Matheson, M., Clark, C., Martin, R., van Kempen, E., Haines, M., Barrio, I. L., Hygge, S., and Stansfeld, S. (**2010**). "The effects of road traffic and aircraft noise exposure on children's episodic memory: The RANCH Project," Noise and Health **12**, 244.

Morset, L. H. (**2004**). "WinMLS 2004," Morset Sound Development, Trondheim, Norway, 400-404.

Muñoz-Sandoval, A. F., Cummins, J., Alvarado, C. G., and Ruef, M. L. (**1998**). *Bilingual verbal ability tests: Comprehensive manual* (Riverside Pub.).

Munro, M. J. (**1998**). "The effects of noise on the intelligibility of foreign-accented speech," Studies in Second Language Acquisition **20**, 139-154.

National Science Board (**2012**). "Science and Engineering Indicators 2012," Arlington VA: National Science Foundation (NSB 12-01).

Nelson, P., Kohnert, K., Sabur, S., and Shaw, D. (**2005**). "Classroom noise and children learning through a second language: Double Jeopardy?," Language, Speech, and Hearing Services in Schools **36**, 219.

Nelson, P. B., and Soli, S. (**2000**). "Acoustical barriers to learning: Children at risk in every classroom," Language, Speech, and Hearing Services in Schools **31**, 356.

Nouri, H. (**2000**). *Effect size for ANOVA designs* (Sage).

Nunnally, J. C., Bernstein, I. H., and Berge, J. M. t. (**1967**). *Psychometric theory* (McGraw-Hill New York).

Perham, N., Banbury, S., and Jones, D. M. (**2007**). "Do realistic reverberation levels reduce auditory distraction?," Applied Cognitive Psychology **21**, 839-847.

Richardson, J. T. (**2011**). "Eta squared and partial eta squared as measures of effect size in educational research," Educational Research Review **6**, 135-147.

Rogers, C. L., Dalby, J., and Nishi, K. (**2004**). "Effects of noise and proficiency on intelligibility of Chinese-accented English," Language and speech **47**, 139-154.

Rogers, C. L., Lister, J. J., Febo, D. M., Besing, J. M., and Abrams, H. B. (**2006**). "Effects of bilingualism, noise, and reverberation on speech perception by listeners with normal hearing," Applied Psycholinguistics **27**, 465.

Rogers, C. L., and Lopez, A. S. (**2008**). "Perception of silent-center syllables by native and non-native English speakers," J Acoust Soc Am **124**, 1278-1293.

Ronsse, L. M., and Wang, L. M. (**2010**). "AB-10-C037: Effects of noise from building mechanical systems on elementary school student achievement."

Ronsse, L. M., and Wang, L. M. (**2013**). "Relationships between unoccupied classroom acoustical conditions and elementary student achievement measured in eastern Nebraska," J Acoust Soc Am **133**, 1480-1495.

Sandrock, S., Schutte, M., and Griefahn, B. (**2007**). "The reliability of the noise sensitivity questionnaire in a cross-national analysis," Noise and Health **9**, 8.

Schutte, M., Marks, A., Wenning, E., and Griefahn, B. (**2007a**). "The development of the noise sensitivity questionnaire," Noise and Health **9**, 15.

Schutte, M., Sandrock, S., and Griefahn, B. (**2007b**). "Factorial validity of the noise sensitivity questionnaire," Noise and Health **9**, 96.

Sherbecoe, R. L., and Studebaker, G. A. (**2004**). "Supplementary formulas and tables for calculating and interconverting speech recognition scores in transformed arcsine units," International Journal of Audiology **43**, 442-448.

Shi, L. F. (**2009**). "Normal-hearing English-as-a-second-language listeners' recognition of English words in competing signals," International Journal of Audiology **48**, 260-270.

Shield, B., Greenland, E., and Dockrell, J. (**2010**). "Noise in open plan classrooms in primary schools: A review," Noise and Health **12**, 225.

Shield, B. M., and Dockrell, J. E. (**2008**). "The effects of environmental and classroom noise on the academic attainments of primary school children," J Acoust Soc Am **123**, 133-144.

Srinivasan, N. K. (**2010**). "The Perception of Natural, Cell Phone, and Computer-Synthesized Speech During The Performance Of Simultaneous Visual-Motor Tasks."

Steeneken, H. J. M., and Houtgast, T. (**1980**). "A physical method for measuring speech‐transmission quality," J. Acoust. Soc. Am. **67**, 318-326.

Studebaker, G. A. (**1985**). "A "rationalized" arcsine transform," Journal of Speech, Language and Hearing Research **28**, 455.

Takayanagi, S., Dirks, D. D., and Moshfegh, A. (**2002**). "Lexical and talker effects on word recognition among native and non-native listeners with normal and impaired hearing," Journal of Speech, Language and Hearing Research **45**, 585.

Tiller, D. K., Wang, L. M., Musser, A., and Radik, M. (**2010**). "AB-10-017 Combined Effects of Noise and Temperature on Human Comfort and Performance (RP-1128)," ASHRAE Transactions **116**, 522.

United States Access Board (**2014**). "Adoption of Acoustic Standards."

Valente, D. L., Plevinsky, H. M., Franco, J. M., Heinrichs-Graham, E. C., and Lewis, D. E. (**2012**). "Experimental investigation of the effects of the acoustical conditions in a simulated classroom on speech recognition and learning in children," J Acoust Soc Am **131**, 232-246.

Woodcock, R. W., McGrew, K., and Mather, N. (**2001a**). *Woodcock-Johnson tests of achievement* (Itasca, IL: Riverside Publishing).

Woodcock, R. W., McGrew, K. S., and Mather, N. (**2001b**). *Woodcock-Johnson III tests of cognitive abilities* (Riverside Pub.).

Wróblewski, M., Lewis, D. E., Valente, D. L., and Stelmachowicz, P. G. (**2012**). "Effects of reverberation on speech recognition in stationary and modulated noise by school-aged children and young adults," Ear and hearing **33**, 731.

Xie, H., Kang, J., and Tompsett, R. (**2011**). "The impacts of environmental noise on the academic achievements of secondary school students in Greater London," Applied Acoustics **72**, 551-555.

Yang, W., and Bradley, J. S. (**2009**). "Effects of room acoustics on the intelligibility of speech in classrooms for young children," J Acoust Soc Am **125**, 922-933.

Zahorik, P. (**2002**). "Assessing auditory distance perception using virtual acoustics," J. Acoust. Soc. Am. **111**, 1832-1846.

Appendix A – Listening Chamber

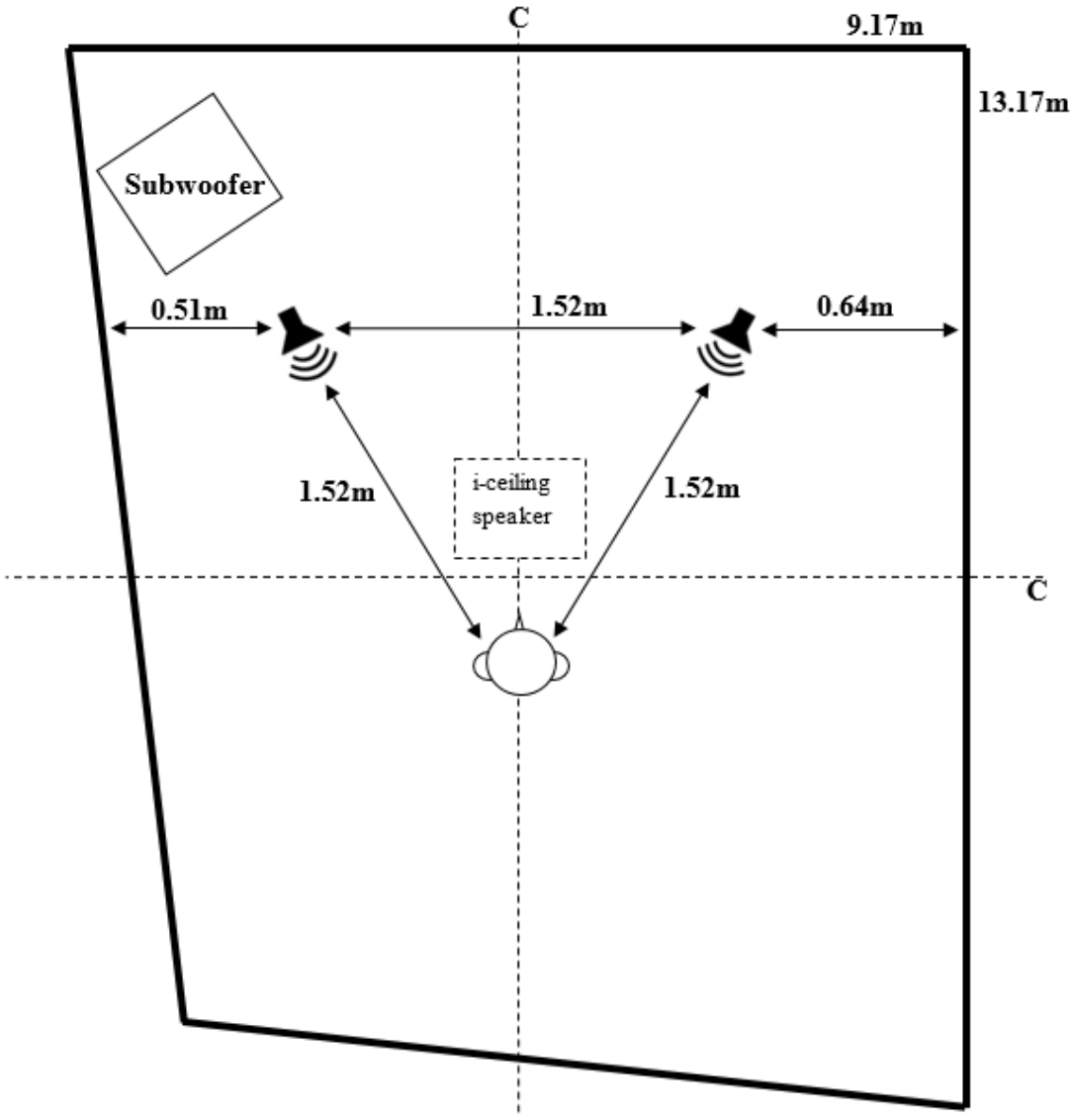


Figure A.1 - Floor plan layout of listening chamber

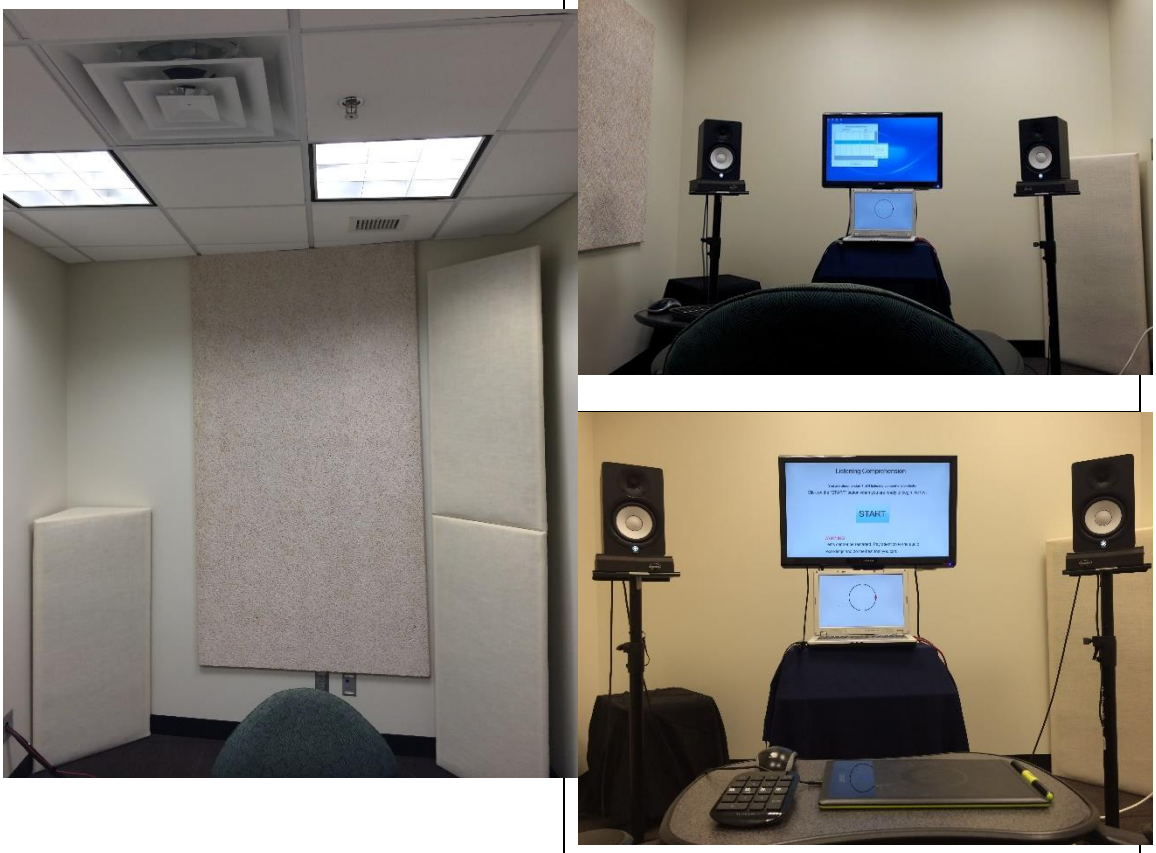
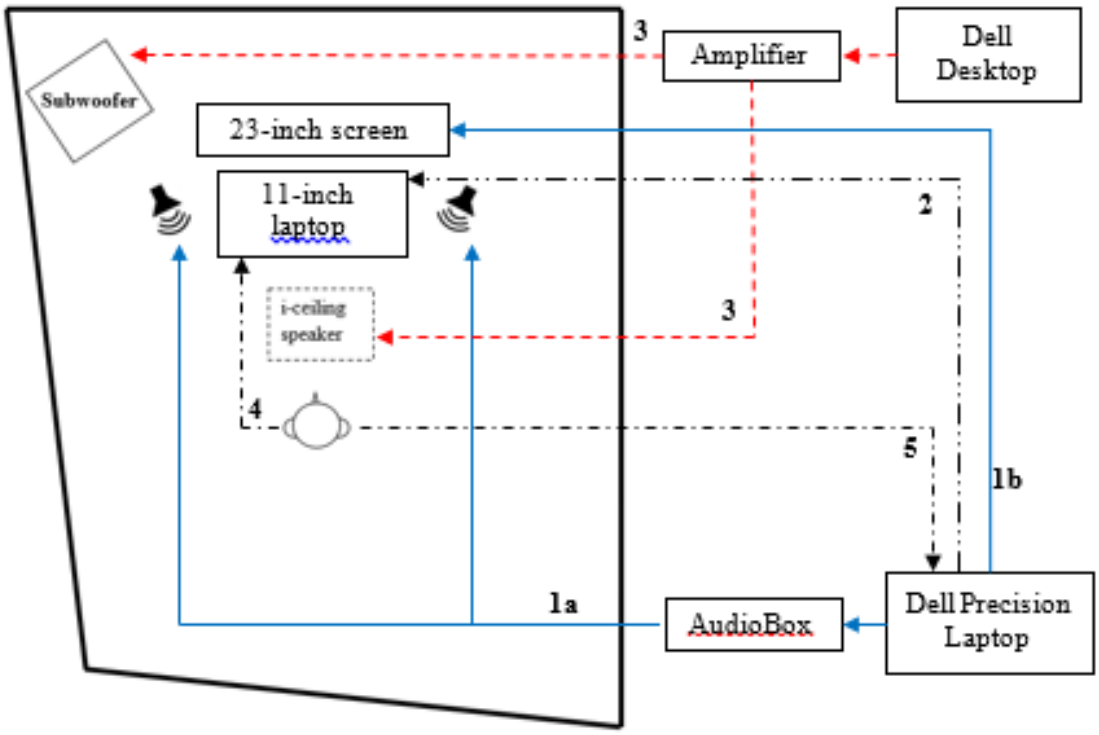


Figure A.2 – Listening chamber front wall view (upper right), back wall view (left), and listener participant view during main experiment (lower right)



Note

Signal 1: Speech comprehension program delivery (1a: audio, 1b: visual)

Signal 2: Control start/end for APR tracing task

Signal 3: Background noise playback

Signal 4: Participant response on APR dot-tracing

Signal 5: Participant response on speech comprehension program

Figure A.3 – Schematics of test program and equipment connections

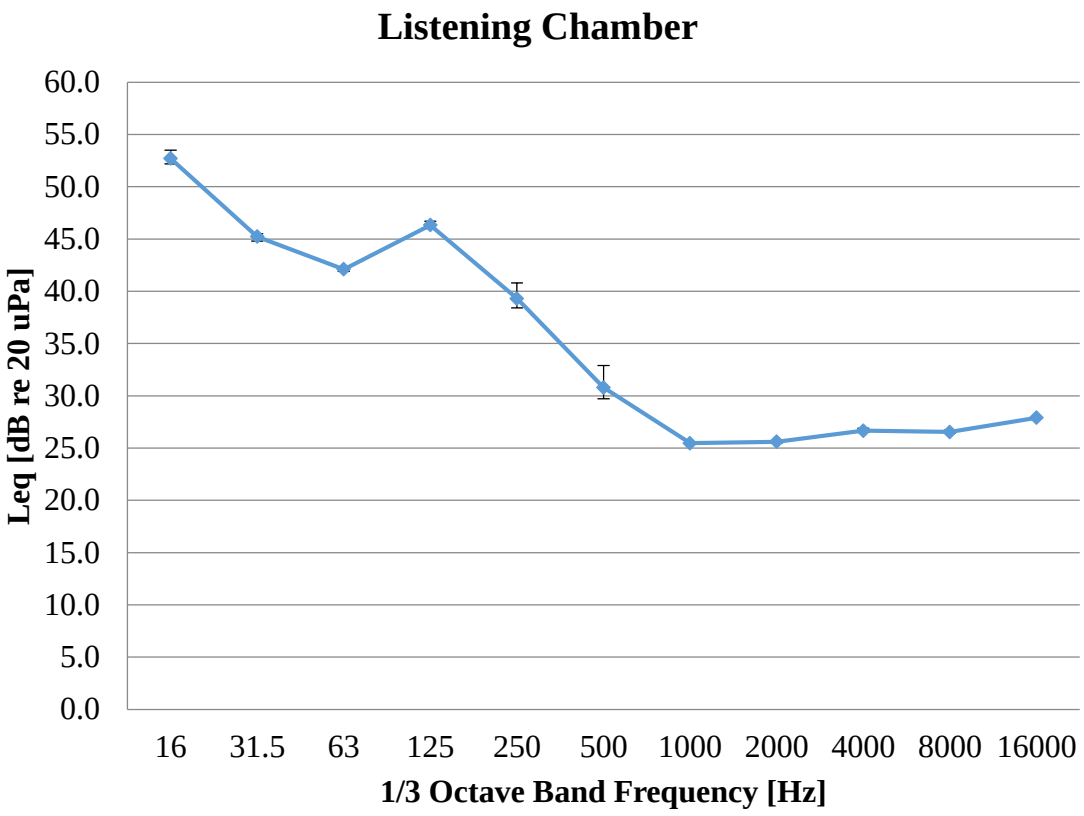


Figure A.4 – Ambient background noise level in listening chamber. Error bars indicate range of values from three measurements.

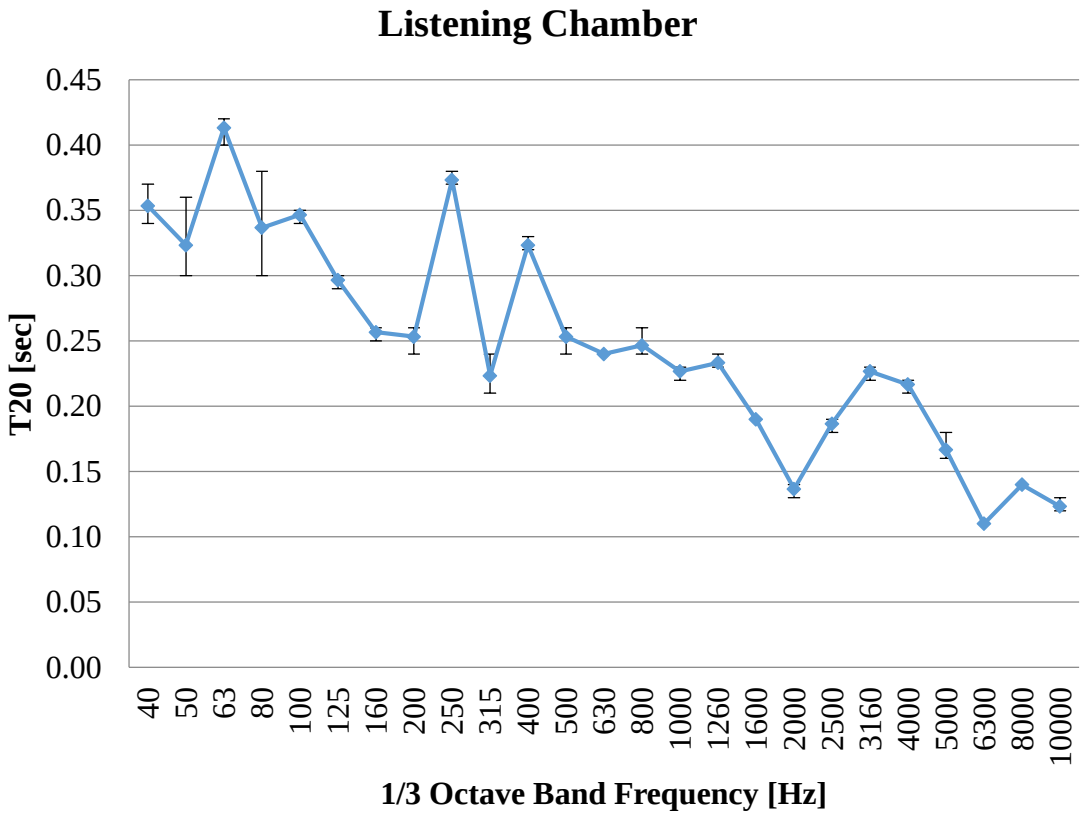


Figure A.5 – Ambient reverberation time in listening chamber. Error bars indicate range of values from three measurements.

Appendix B – Sound Booth

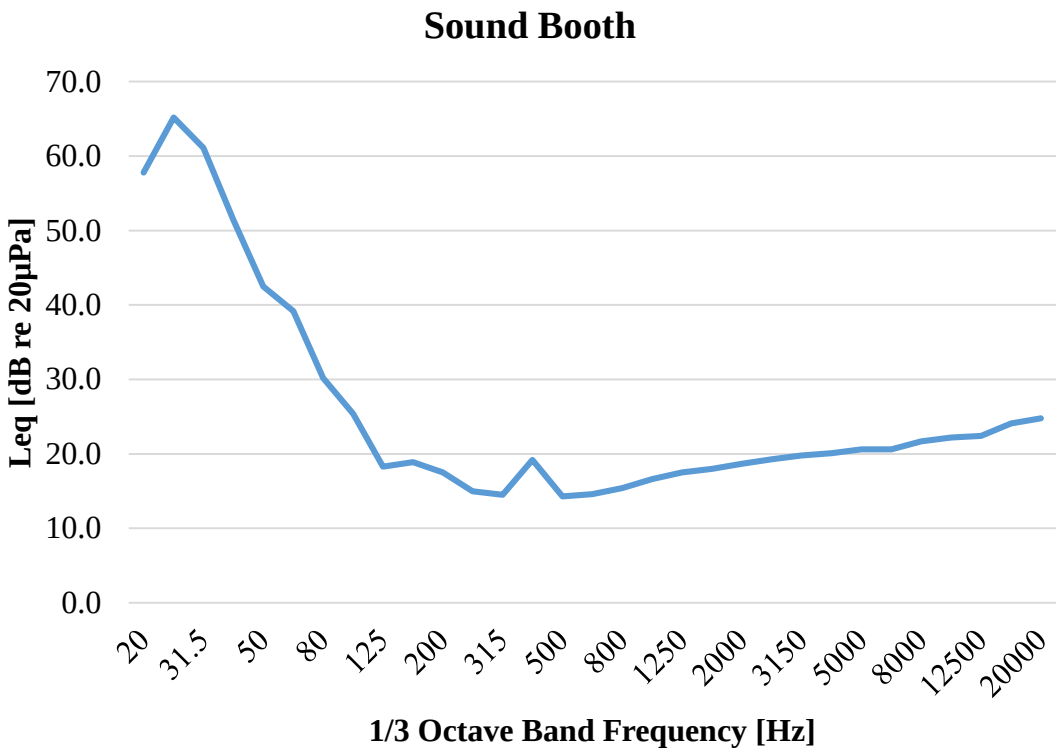


Figure B.1 – Ambient background noise level in sound booth for speech material recording in Study 2

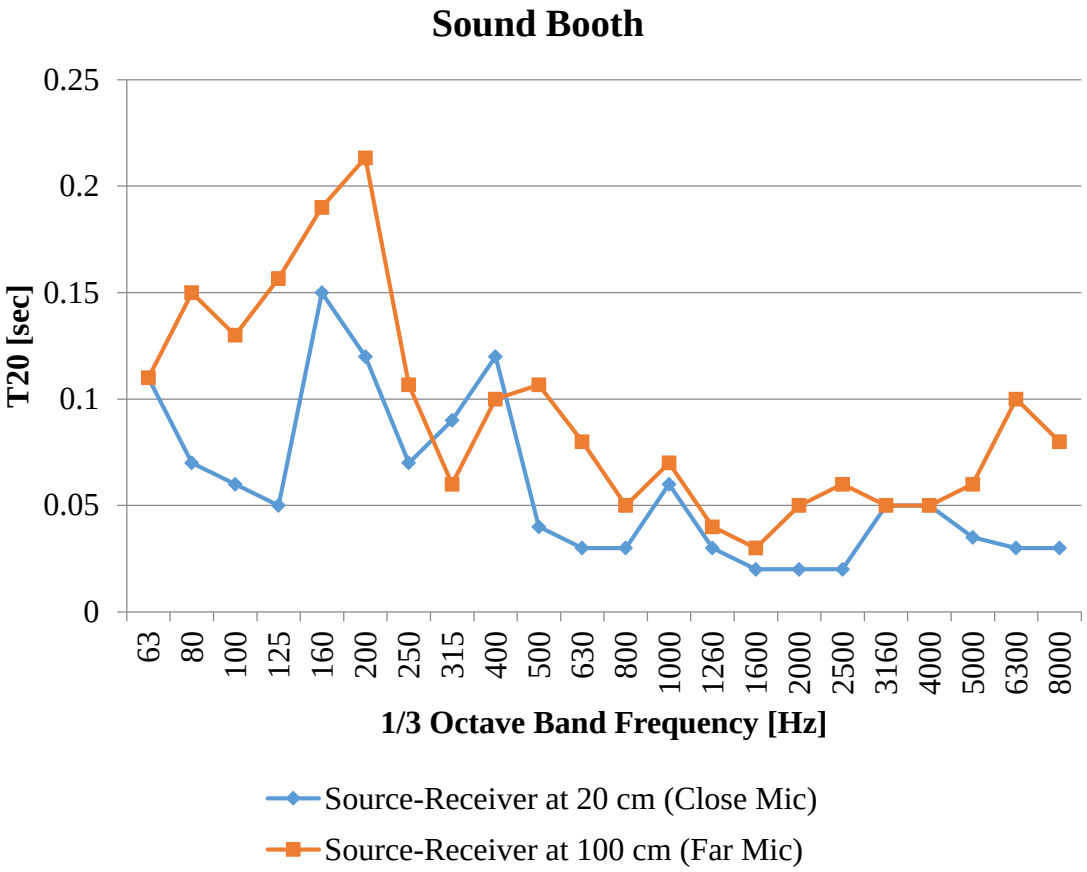


Figure B.2 – Ambient reverberation time in sound booth for speech recording in Study 2 under two source-receiver configurations

Appendix C – Surveys and Questionnaires

Admin Use
Participant ID _____

Participant Demographic Survey – Part 1

Please answer the following questions by marking the appropriate box(es) or filling in the blanks. This survey is intended ONLY for gathering demographic information of participants. Please be assured that your eligibility in participating in this study WILL NOT be affected by answers on this survey.

1a. Age: _____ 1b. Gender: ☐ Male ☐ Female

2. Do you consider yourself to be Hispanic/Latino?

☐ Yes. ☐ No.

2a. In addition, select one or more of the following racial category to describe yourself:

☐ American Indian or Alaska Native ☐ Asian ☐ White

☐ Black or African American ☐ Native Hawaiian or Pacific Islander

☐ Other, please indicate _____

3. Please list all the languages you know **in order of acquisition** (your native language first):

1	2	3	4	5

4. Please rank the above languages **in order of dominance**:

1	2	3	4	5

5. When choosing to read a text available in all your languages, in what percentage of cases would you choose to read it in each of your languages? Assume that the original was written in another language, which is unknown to you.
(*Your percentages should add up to 100%)*

List language here:					
List percentage here:					

6. When choosing a language to speak with a person who is equally fluent in all your languages, what percentage of time would you choose to speak each language? Please report percent of total time.
(*Your percentages should add up to 100%)*

List language here:					
List percentage here:					

Survey continues on next page ~~~

Participant Demographic Survey – Part 1

7a. Have you ever dreamed in English?

☐ Yes. ☐ No. ☐ I would rather not answer.

7b. What language do you use to count numbers? _____

8. Age when you …

8a. began acquiring *English:*	8b. became fluent in *English:*	8c. began reading in *English:*	8d. became fluent reading in *English:*

9. Please list the number of years and months you spent in each language environment:

	Years	Months
A country where English is spoken		
A family where English is spoken		
A school and/or working environment where English is spoken		

10. Please circle your *level of **proficiency*** in speaking, understanding, reading and writing from a scale of 0 to 10.

0 – none 1 – very low 2 – low 3 – fair
4 – slightly less than adequate 5 – adequate 6 – slightly more than adequate
7 – good 8 – very good 9 – excellent 10 – perfect

10a. Speaking

0 1 2 3 4 5 6 7 8 9 10

10b. Understanding spoken language

0 1 2 3 4 5 6 7 8 9 10

10c. Reading

0 1 2 3 4 5 6 7 8 9 10

10d. Writing

0 1 2 3 4 5 6 7 8 9 10

11. Have you ever taken or studied for the following tests? Check all that apply.

☐ TOEFL ☐ TOEIC ☐ GRE ☐ SAT ☐ ACT

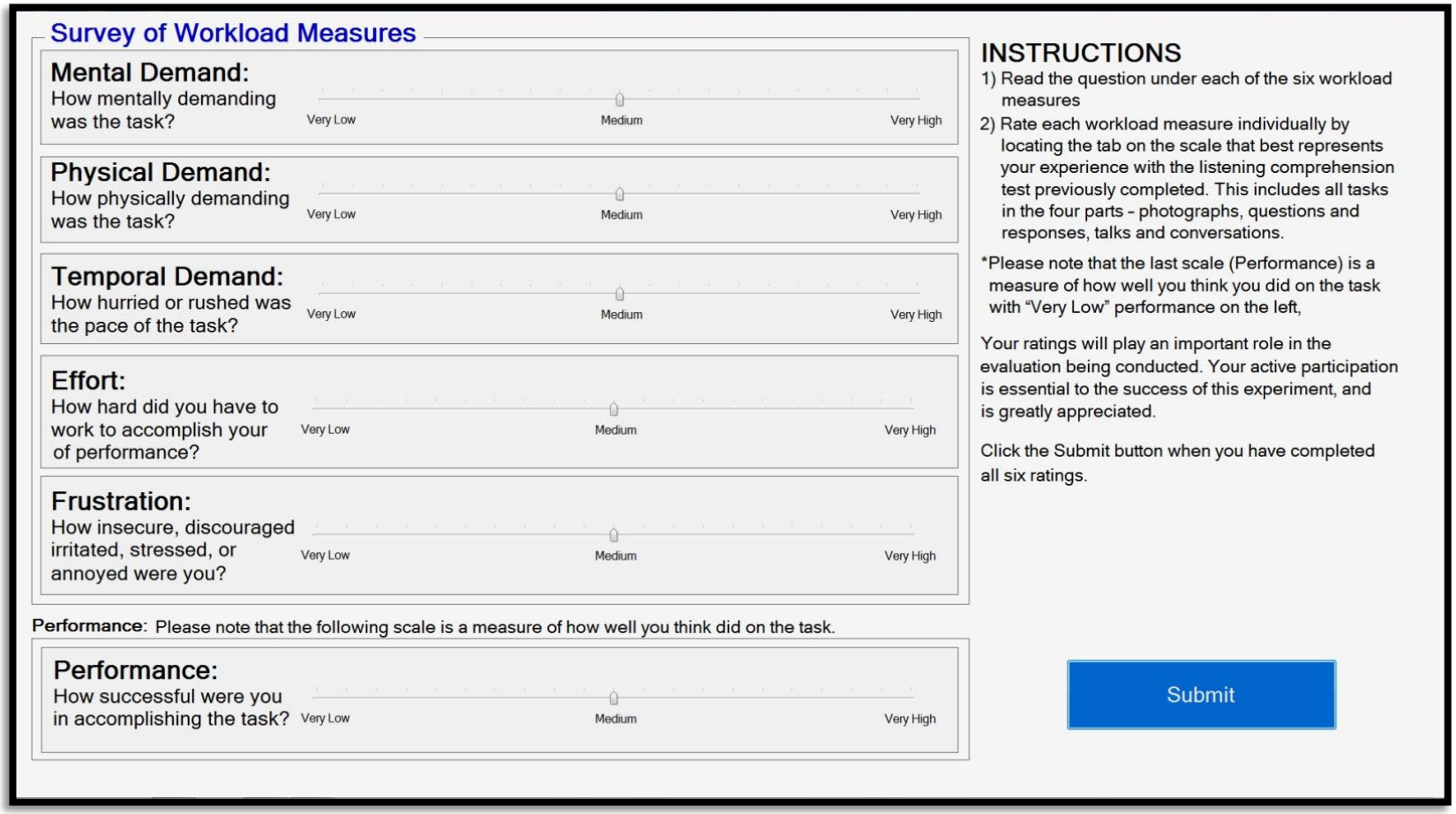
~~~ Survey continues on next page ~~~

Participant Demographic Survey – Part 2

Please rate each statement in order. Please do not skip any questions. If possible, imagine yourself in each situation and response accordingly without spending too much time considering if you agree or disagree with a given statement. We are looking for your personal opinions. There are no correct or incorrect responses.

		Strongly agree	Slightly agree	Slightly disagree	Strongly disagree
1	I need an absolutely quiet environment to get a good night's sleep.	1	2	3	4
2	I need quiet surroundings to be able to work on new tasks.	1	2	3	4
3	When I am at home, I habituate to noise quickly.	1	2	3	4
4	I become very agitated if I can hear someone talking while I am trying to fall asleep.	1	2	3	4
5	I am very sensitive to neighborhood noise.	1	2	3	4
6	When people around me are noisy I don't get on with my work.	1	2	3	4
7	I am sensitive to noise.	1	2	3	4
8	My performance is much worse in noisy places.	1	2	3	4
9	I do not feel well rested if there has been a lot of noise the night before.	1	2	3	4
10	It would not bother me to live in a noisy street.	1	2	3	4
11	For a quiet place to live I would accept other disadvantages.	1	2	3	4
12	I need peace and quiet to do difficult work.	1	2	3	4
13	I can fall asleep even when it is noisy.	1	2	3	4

~~~End of Survey ~~~


Survey of Workload Measures
Mental Demand:
How mentally demanding was the task?
Very Low
Medium
Very High
Physical Demand:
How physically demanding was the task?
Very Low
Medium
Very High
Temporal Demand:
How hurried or rushed was the pace of the task?
Very Low
Medium
Very High
Effort:
How hard did you have to work to accomplish your of performance?
Very Low
Medium
Very High
Frustration:
How insecure, discouraged irritated, stressed, or annoyed were you?
Very Low
Medium
Very High
Performance: Please note that the following scale is a measure of how well you think did on the task.
Performance:
How successful were you in accomplishing the task?
Very Low
Medium
Very High
INSTRUCTIONS
1) Read the question under each of the six workload measures
2) Rate each workload measure individually by locating the tab on the scale that best represents your experience with the listening comprehension test previously completed. This includes all tasks in the four parts - photographs, questions and responses, talks and conversations.
*Please note that the last scale (Performance) is a measure of how well you think you did on the task with "Very Low" performance on the left,
Your ratings will play an important role in the evaluation being conducted. Your active participation is essential to the success of this experiment, and is greatly appreciated.
Click the Submit button when you have completed all six ratings.
Submit

Appendix D – Native Language Profile of Listeners in Both Studies

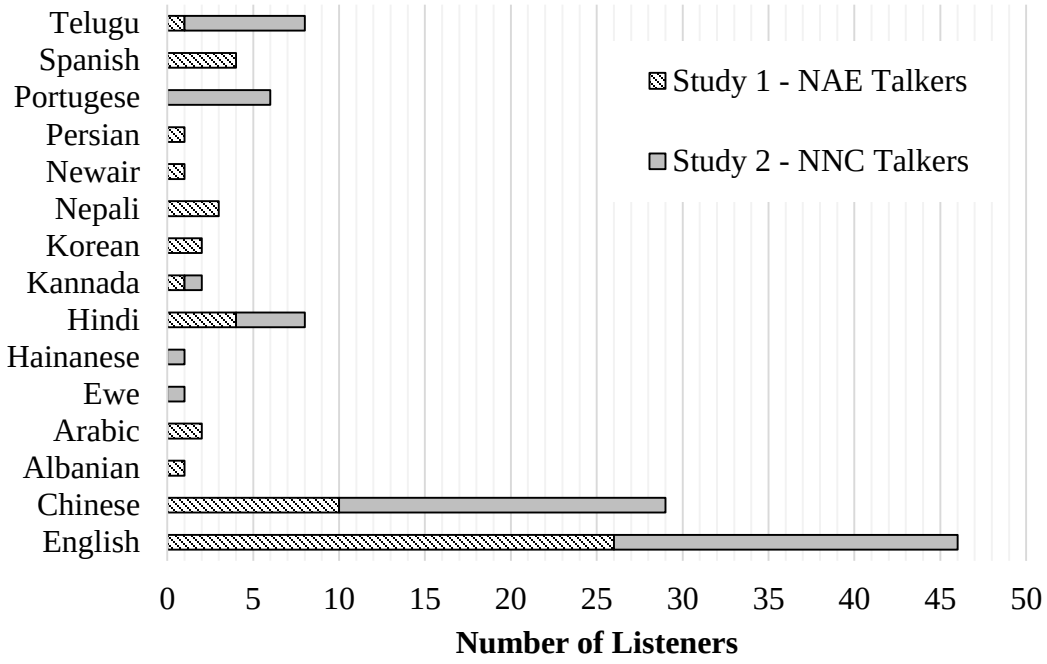


Figure D.1 – Native language profile of listeners

Appendix E – Select List of BKB-R Sentences

1) The children dropped the bag
2) The dog came back
3) The floor looked clean
4) She found her purse
5) The fruit is on the ground
6) Mother got a saucepan
7) They washed in cold water
8) The young people are dancing
9) The bus left early
10) The ball is bouncing very high
11) Father forgot the bread
12) The girl has a picture book
13) The boy forgot his book
14) A friend came for lunch
15) The match boxes are empty
16) He climbed his ladder
17) The family bought a house
18) The jug is on the shelf
19) The ball broke the window
20) They are shopping for cheese
21) The pond water is dirty
22) They heard a funny noise
23) The police are clearing the road
24) The bus stopped suddenly
25) The book tells a story
26) The young boy left home
27) They are climbing the tree
28) She stood near her window
29) The table has three legs
30) A letter fell on the floor
31) The five men are working
32) The shoes were very dirty
33) They went on a vacation
34) The baby broke his cup
35) The lady packed her bag
36) The dinner plate is hot
37) A dish towel is by the sink
38) She looked in her mirror
39) The good boy is helping
40) They followed the path
41) The kitchen clock was wrong
42) Someone is crossing the road
43) The mailman brought a letter
44) They are riding their bicycles
45) He broke his leg
46) The milk was by the front door
47) The shirts are hanging in the closet
48) The chicken laid some eggs
49) The orange was very sweet
50) He is holding his nose
51) The new road is on the map
52) She writes to her brother
53) The football player lost a shoe
54) The three girls are listening
55) The coat is on a chair
56) The train is moving fast
57) The child drank some milk
58) The janitor used a broom
59) The ground was very hard
60) The buckets hold water

Note: Sentences adopted from Bent and Bradlow (2003; appendix).