Winter 12-21-2023

Characterizing the Relationship between Accented Speech Intelligibility and Listening Effort

Mel Mallard
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Department of Psychological & Brain Sciences

Characterizing the Relationship between Accented Speech Intelligibility and Listening Effort

by
Mel Mallard

A thesis presented to
Washington University in St. Louis
in partial fulfillment of the
requirements for the degree
of Master of Arts

December 2023
St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

Acknowledgements

Thank you to my Master's Examination Committee: Kristin Van Engen, Mitchell Sommers, and Julie Bugg. Thank you to our undergraduate research assistants for their truly indispensable work and enthusiasm. Thank you to the National Science Foundation for supporting this work under Grant No. 2146993. And thank you to the Department of Psychological & Brain Sciences at Washington University in St. Louis for the opportunity to pursue this degree.

Mel Mallard

Washington University in St. Louis

December 2023

ABSTRACT OF THE THESIS

Characterizing the Relationship Between Accented Speech Intelligibility and Listening Effort

by

Mel R. Mallard

Master of Arts in Psychological & Brain Sciences

Washington University in St. Louis, 2023

Dr. Kristin J. Van Engen

Unfamiliar accents can make speech communication difficult, both by reducing speech intelligibility and by increasing the effort listeners must put forth to understand speech. However, these two constructs, while related, are independent: for example, two 100% intelligible utterances may require different amounts of effort to accurately process.

To better characterize the relationship between intelligibility and effort, this study presents speakers of four intelligibility levels (one natively-accented English speaker, and three Mandarin-accented English speakers) within a dual-task paradigm (featuring a vibrotactile secondary task) to measure listening effort. We found a negative nonlinear relationship between intelligibility and effort, with the steepest slope between the native speaker and the highly intelligible Mandarin-accented speaker, and the shallowest slope between the two least intelligible Mandarin-accented speakers. These results suggest a local plateau in effort that arises relatively soon after intelligibility begins falling.

Chapter 1: Introduction & Background

Speech communication is a complex, but often automatic, process. The acoustic signal must be sensed, processed, and interpreted, all with speed and ease. However, if the signal is in some way degraded, distorted, ambiguous, or unfamiliar, this automatic process can become a more intentional and effortful task. Whether struggling to understand a friend at a loud party, or a professor mumbling behind a distant lectern, most have felt the strain induced by such challenges. What's more, these situations can impair one's ability to perform concurrent tasks: one is unlikely to fully remember the story told at the busy party or take effective notes during the frustrating lecture. In this way, we all are experienced with the challenge of decreased speech intelligibility and its consequence of increased listening effort.

1.1 Intelligibility & Listening Effort

Speech intelligibility generally refers to the proportion of speech that a listener can accurately identify. It can often be conceptualized along a psychometric function, where the x-axis represents some feature of the speech signal (e.g. decibels, the strength of competing noise, the number of audio channels), and the y-axis represents keyword identification accuracy. Though there is great variation in the functions themselves, the general trend is that as the speech signal increasingly departs from baseline, intelligibility decreases, following a linear or sigmoidal function (see MacPherson & Akeroyd, 2014, for a systematic survey).

In contrast, listening effort is a psychological concept. Listening effort is the recruitment of additional cognitive resources (such as attention and working memory) towards a difficult listening task. According to the "effortfulness hypothesis", as an acoustic challenge becomes increasingly difficult, more cognitive resources must be recruited to the listening task in order to

maintain adequate performance (Broadbent, 1958; McCoy et al., 2005; McGarrigle et al., 2014; Rabbitt 1968; Sarampalis et al., 2009, Pichora-Fuller et al., 2016).

This effortfulness hypothesis is situated within the psychological framework of limited cognitive resources, such as those initially described by Broadbent (1958) and Kahneman (1973). Operating under this understanding of the interplay between cognitive resources and cognitive demand, deploying listening effort is a form of resource reallocation. According to Kahneman's (1973) Capacity Model of Attention, the recruitment of additional cognitive resources towards listening would necessarily deplete the resources available for other cognitive tasks. This theoretical supposition is evidenced by many studies reporting impaired performance for processes that compete for resources during a challenging listening task (see Peelle, 2018, for a review). Examples include poorer recall for speech heard with hearing loss, hearing aids, and in noise (Lunner et al., 2009; McCoy et al., 2005; Sarampalis et al., 2009), increased physiological markers of stress (cognitive load, task demand) for speech heard with competing-talkers or in noise (Mackersie & Cones, 2011; Seeman & Sims, 2015; Zekveld et al., 2010), and increased subjective measures of effort, frustration, and task difficulty for speech in noise (Seeman & Sims, 2015, Mackersie & Cones, 2011).

It should be noted that increased listening effort can be elicited by acoustic challenge even when speech is highly intelligible (Fraser et al., 2010; Mackersie & Cones, 2011; Rabbitt, 1968;). As proposed in a prominent model of listening effort, the Ease of Language Understanding model (ELU; Rönnberg et al., 2008), whenever an acoustic "mismatch" occurs between the speech signal and the listener's stored internal representations, effort must be used to rectify the mismatch. This characterizes effort as a process that resolves ambiguity, increasing the proportion of the speech signal that can be accurately perceived. Consequently, this also

means that even fully intelligible speech can require effort, and that two fully intelligible utterances may have required different amounts of effort to perceive.

Listening effort therefore can be conceptualized as a productive phenomena – used to maintain intelligibility – rather than a cognitive state comorbid to acoustic challenge (see Francis & Love, 2019, for further discussion). Considering the adaptive goal of listening effort alongside its cognitive costs, it theoretically follows that the deployment of effort would be sensitive (though not commensurate) to task difficulty. In other words, maximal effort would not be expended if it is not needed, and would be withdrawn when unnecessary or ineffective. These patterns indeed are contained within the literature thus presented. Specifically, these findings generally describe either a monotonic (effort remains maximal) or peaked (effort withdraws) or relationship.

1.2 Speech Perception of Non-native Accents

An unfamiliar accent can also be a form of acoustic challenge (Mattys et al., 2012), since non-native speech deviates from native norms at both the segmental and suprasegmental level (Munro & Derwing, 1995). Within the ELU's acoustic mismatch model (Rönnberg et al., 2008), this causes the listener to experience increased ambiguity and difficulty mapping speech sounds onto stored representations, which would require extra effort to resolve. This has been empirically reflected in: listeners' impaired accented speech recognition scores (Bent and Bradlow, 2003; Ferguson et al., 2010; Munro and Derwing, 1995), increased self-reported effort (Munro and Derwing, 1995; Schmid and Yeni-Komshian, 1999), different patterns of brain activity as measuring by event-related potentials (ERPs; Romero-Rivas, 2015), and increased task-evoked pupil dilation during accented speech processing, even when the speech is fully intelligible (Brown et al. 2020; McLaughlin & Van Engen, 2020).

Listening to accented speech therefore appears to elicit increased listening effort in much the same way as other forms of acoustic challenge (Peelle & Van Engen, 2014); however, since the nature of the challenge is distinct, the relationship between intelligibility and listening effort for accented speech may also be distinct. The effort of processing accented speech across several levels of intelligibility has been investigated relatively little, though some studies have broached the topic: Bradlow & Bent (2008) investigated perceptual adaptation for four Mandarin-accented speakers, and Wilson & Spaulding (2010) investigated processing speed for four Korean-accented speakers.

Most intriguing, Porretta and Tucker (2019) used monosyllabic words from four Mandarin-accented speakers and analyzed peak pupil dilation in response to the mean intelligibility of the word. This study provides important insight that intelligibility and listening effort appear to follow the peaked trend (with higher effort exerted at 50% intelligibility than at 100% and 0%). However, this study also raises further questions. Porretta and Tucker (2019) used monosyllabic words, and presented their inferential models predicting the extreme binned levels of 0%, 50%, and 100% intelligibility. This design leaves room for further investigation using more naturalistic stimuli; namely, longer utterances situated mostly in the upper ranges of intelligibility. It also leaves room for further specification and nuance in characterizing this relationship. For constructing a relationship between accented speech intelligibility, there are several parameters of interest. These include but are not limited to: the degree of deviance necessary to elicit an initial increase in effort, the peak of the function (height, sharpness), the change in intelligibility that elicits the largest change in effort, and the behavior of the tail.

Clearly, more research needs to be done on how the intelligibility of non-native accent impacts the psychological process of effortful listening, especially for utterances longer than

monosyllabic words, and across a more fine-grained and ecologically relevant range of intelligibility. This is the goal of the present study. Given our knowledge of the relationship between intelligibility and listening effort for other forms of acoustic challenge and the knowledge that non-native accent is an acoustic challenge in a league of its own, we aimed to illuminate this relationship.

We sought to present native monolingual English speaking participants with several speakers of differing intelligibility, and measure the listening effort associated with each speaker. Our sample of speakers is mostly concentrated in the upper ranges of intelligibility because L2 English speakers residing in the United States with enough proficiency to enroll in a recording session are unlikely to have such low pronunciation accuracy as to be mostly unintelligible. Therefore, this is a novel (compared to Poretta & Tucker, 2019), ecologically valid, and theoretically appropriate approach, as it allows us to take a finer-grained look at the range of intelligibility most likely to appear in real-world listening scenarios.

1.3 The Dual-Task Paradigm

Many methodologies have been used to experimentally measure listening effort. Popular methods include: subjective measures such as self-reports of effort, frustration, performance; behavioral measures such as dual-tasking and recall paradigms; and physiological measures of changes in heart rate, skin conductance, pupil dilation, and brain activity (see McGarrigle et al., 2014 for a white paper on these methods and their theoretical underpinnings, and Strand et al., 2018 for their sensitivity and validity).

For our study, we chose the dual-task paradigm (see Gagné et al., 2017, for a review of dual-task listening effort studies). In this paradigm, participants complete two simultaneous tasks: a primary task and a secondary task. The primary task contains the manipulation of interest

(e.g., listening to speech and repeating what was heard), whereas the secondary task is a separate or unrelated task (e.g., recalling the highest number spoken, or responding to a visual probe). Participants are explicitly instructed to prioritize the primary task and perform it to the best of their ability, and performance on the secondary task is taken as an inverse index of effort of the primary task. The theoretical assumptions of this paradigm come from Kahneman's (1973) Capacity Model of Attention, which describes cognitive resources, primarily attention, as limited and finite. (Recall that this model is foundational in defining the effortfulness hypothesis for listening effort.) Within this model, the effort put forth towards two simultaneous tasks cannot exceed the participants' individual attentional capacities. This means that so long as both tasks are easy and require minimal resources, they both can be performed optimally, but when the prioritized primary task becomes sufficiently difficult, there are limited cognitive resources available to be directed towards the secondary task, and performance there will suffer. In this way, impaired performance in the secondary task can index the effort being directed towards the primary task.

This paradigm can dissociate the performance in the primary task from the effort that is required to complete said task - a participant performing at 100% accuracy in a listening task may require different levels of effort to achieve that level of accuracy. This dovetails with the listening effort literature that speech intelligibility is dissociable from listening effort (Winn & Teece, 2021).

Another consideration when choosing the dual-task paradigm was the ongoing discourse in defining the construct of listening effort and its best measures (Pichora-Fuller et al., 2016; McGarrigle, 2014; Strand et al., 2018). The dual-task paradigm is well-situated within the fundamentals of cognition (cognitive resources are limited) and listening effort (effort is costly).

Additionally, this paradigm is noteworthy in that it results in online behavioral data reflecting the difficulty of completing a task during a listening challenge. In the real world, this could be listening to an accented professor while taking notes, deciphering the announcements on an airport PA system while trying to navigate to an unfamiliar gate, or listening to a crackling phone call while driving a car. Even if the general construct of "listening effort" is somewhat broadly defined, the paradigm itself provides ecologically valid data on how listening challenge affects our ability to perform concurrent tasks.

The sensitivity of the secondary task is dependent on many features, including modality, difficulty, complexity, and relevance to the main task (Picou & Ricketts, 2014; Gagné et al., 2017; Strand et al., 2018). For this reason, we targeted a secondary task that would be the most flexible across multiple studies, so we would not be forced to switch or adapt the task, as this would increase the difficulty in synthesizing results. For example, a visual or visual-motor task (e.g., a flashing light or moving target) is incompatible with studies of audiovisual speech integration; semantic secondary tasks (such as word category identification) are implemented with respect to the specific lexical items being presented, constraining expansion into new sets of audio stimuli; and auditory secondary tasks may alter speakers' intelligibility (see Rogers et al., 2004, and Wilson & Spaulding, 2010, for the effect of noise on accented speech perception). Therefore, we chose a vibrotactile secondary task. It occupies an entirely separate modality than both speech and vision, and can be readily applied to diverse stimuli. This has been used for studies on listening effort for, among others, audiovisual speech (Fraser et al., 2010; Brown & Strand, 2019), noise reduction (Sarampalis et al., 2009), and aging (Gosselin & Gagné , 2011).

Our paradigm is based largely on that used by Brown and Strand (2019) for investigating audiovisual listening effort. Particularly, that study used three possible vibrotactile stimuli: a

short (100 ms), medium (150 ms) or long (250 ms) vibration presented to the participants' index finger. During experiment building, our in-lab pilots revealed poor accuracy for distinguishing these durations (perhaps an artifact of our particular vibrational motor), and so we opted for the durations of 100 ms, 200ms, and 350 ms.

1.4 Hypotheses

Taken all together, we therefore hypothesize that as speaker intelligibility decreases, listening effort increases, and our behavioral index of effort – reaction times on the secondary task – will increase. We also anticipate the possibility that intelligibility may reach a point so low that listeners withdraw their effort from the listening task, and thus improve their performance on the secondary task, resulting in decreased reaction times. Of primary interest, however, is the relationship that appears between intelligibility and listening effort. Though a peaked shape has been identified in other listening effort literature, and touched upon for non-native accent in Poretta and Tucker (2019), we are interested in specifics such as the slope, point(s) of inflection, and general shape (e.g. peaked, monotonic, linear). Our study aims to be a first look towards potentially identifying these features.

Chapter 2: Intelligibility Pilot Study

Prior to the main experiment, we piloted the intelligibility of 32 male L2 speakers of English to determine the L1 accent for which we had the most suitable stimuli. The goal was to construct the broadest range of intelligibility possible, with the speakers distributed somewhat equidistant across that range.

2.1 Methods

All procedures were approved by the Washington University in St. Louis Institutional Review Board prior to experimentation.

Participants

We collected complete online datasets from 234 participants, resulting in a final sample of $N = 202$ participants after exclusions. Three separate pilots were conducted, each with unique participants. The final samples were: Mandarin $N = 104$, Turkish $N = 90$, and Hindi-Russian-Korean $N = 8$. All participants for the Mandarin and Turkish pilots were undergraduate students from Washington University in St. Louis, recruited through the SONA psychology research participation portal, and compensated with 0.5 credits for 30 minutes of participation. Participants for the Hindi-Russian-Korean pilot were recruited through Prolific and compensated with $10 for 30 minutes of participation.

Participant eligibility was determined through the Study Questionnaire administered at the end of the study. All eligible participants were aged 18-35 years old, self-identify as "native" speakers of English (acquired English before the age of 3), with self-reported normal hearing, did not report the target language(s) in the "known languages" free-response, listened to the stimuli using earbuds or over-ear-headphones, and did not select "yes" for the question "Is there any

reason that your data should be excluded from this experiment?"

Materials

Stimuli

The auditory stimuli consisted of 120 sentences from the Hearing In Noise Test (HINT; Nilsson et al., 1994) set, read aloud by male non-native speakers of English. HINT sentences were developed for the use of measuring speech intelligibility. An example sentence is "She's drinking from her own cup". The recordings were downloaded from the ALLSSTAR corpus (Bradlow, n.d.), which is available on the online repository SpeechBox (https://speechbox.linguistics.northwestern.edu/). Using Praat, the recordings were segmented into individual sentence-length audio files, and equalized for RMS amplitude.

Three separate pilots were conducted, consisting of, respectively: 10 L1 Mandarin speakers; 10 L1 Turkish speakers; and four speakers each of L1 Hindi, Russian, and Korean.

Hardware & Software

These pilots were conducted on Gorilla, an online behavioral science experiment builder (https://gorilla.sc/). Participants completed the study on their own personal computers, listening to the stimuli using earbuds or over-ear headphones.

Procedure

Upon being directed to the experiment page on Gorilla, participants were required to read an informed consent sheet (no signature required due to the project's exempt IRB status). They were then provided instructions: for each trial, they hear a sentence and must type what they hear into the free response box.

For the Mandarin and Turkish pilots, 100 HINT sentences were chosen. Each participant heard all 100 sentences once, 10 per speaker, in fully randomized order. For the Hindi-Russian-

Korean pilot, each participant heard all 120 sentences once, 10 per speaker. For this pilot, we blocked by accent (block order counterbalanced), with stimuli fully randomized within each respective block.

In each pilot, counterbalancing ensured that we had data for each speaker producing each sentence. Note however that there were four audiofiles that did not exist (e.g., Turkish speaker 004 saying sentence 51). These missing files (three for Turkish speakers, one for Hindi) were replaced with "filler" files of a different HINT sentence produced by the same speaker.

For the Mandarin and Turkish pilots there was a mid-experiment break. For the Hindi-Russian-Korean pilot, there was a break between each block. On the break screen, participants were reminded of their 1 hour time limit and told to continue when they were ready.

2.2 Data

Before any analyses, participants' data were removed if they failed our eligibility criteria detailed in the Participant section (2.1 Methods). There were 32 total excluded participants. Final counts for the exclusion criteria are as follows: self-reported non-native English speaker (2 Mandarin, 1 Turkish, 0 Hindi-Russian-Korean), self-reported familiarity with the target language (8 Mandarin, 0 Turkish, 0 Hindi-Russian-Korean), self-reported non-normal hearing (0 Mandarin, 2 Turkish, 0 Hindi-Russian-Korean), failure to use earbuds or over-ear headphones (5 Mandarin, 3 Turkish, 1 Hindi-Russian-Korean), an average performance below 3SD of the overall mean (4 Mandarin, 0 Turkish, 0 Hindi-Russian-Korean), and requested self-exclusion (3 Mandarin, 3 Turkish, 0 Hindi-Russian-Korean).

The responses were scored using the HINT scoring system (Nilsson et al., 1994). In this system, each word is scored as correct or incorrect and the "score" for the sentence is reported as a "proportion correct". For example, if the target sentence is "Father forgot the bread" and the

participant types “Father bought the bread”, that trial will score three correct words and one incorrect word, and earn a final proportion correct of 0.75.

This scoring was performed using the autoscore package (Borrie et al., 2019) in R. The autoscore() function compares the strings in a rubric column to the strings in a participant response column, and counts the number of correct and incorrect matches, and reports the proportion correct per sentence.

To accurately perform this grading, prior to using the package, the typed responses – when appropriate – must be adjusted to align with the rubric. For example, if the rubric sentence features a compound word such as “milkman”, any participant responses of “milk man” must be collapsed into a single word (and vice versa, such as a sentence featuring “every day” where participant responses of “everyday” must be separated). Additionally, one feature of the HINT scoring system is that function words often have acceptable substitutions, such as verb tense (“is/was”, “have/had”, “has/had”, “are/were”), and articles (“a/the”). For example, the audio stimulus “A new road is on the map" has the following allowances: “(A/the) new road (is/was) on (a/the) map" (Nilsson et al., 1994). (These substitutions are outlined for each specific sentence, as they are not uniformly permitted: “They *had* some cold cuts for lunch” cannot be “They *have* some cold cuts for lunch”; and “They *walked* across *the* grass” cannot be “They *walk* across the grass” or “They walked across *a* grass”.) Autoscore itself has an “a/the” equivalence setting which we implemented. To handle all other cases of participant substitutions, we created R code to perform an editing procedure, where eligible substitutions are replaced with the original target word, only for the sentences where they are permitted. Meaning, if the speaker said “A new road is on the map” – a sentence where “was” is an appropriate substitution – the code will check any participant responses for that sentence, and replace any “was” with “is”, so

that the response earns the appropriate points when compared to the rubric column. This code will be uploaded to this project's public OSF page upon project completion.

Note that because of our particular interest in Mandarin Chinese, after the first round of data analysis, we added an additional data processing step where we manually corrected typos (such as "sisterr", "quick;y", "al lnight"), to achieve the most accurate measure of intelligibility as possible. Typos were not corrected if the participant's intent was not fully apparent. Correcting typos improved all Mandarin speakers' average intelligibility by less than 0.02. No other languages' data are typo-corrected.

2.3 Results

The mean intelligibility score across listeners is given for each speaker in Table 1.

Table 1. *Mean intelligibility of piloted speakers. Arranged from most to least intelligible.*

	Speaker									
L1 Accent	1	2	3	4	5	6	7	8	9	10
Mandarin	0.955	0.944	0.937	0.934	0.922	0.917	0.912	0.899	0.865	0.829
Turkish	0.974	0.973	0.972	0.971	0.968	0.963	0.961	0.582	0.537	0.53
Hindi	0.986	0.974	0.954	0.915						
Russian	0.959	0.952	0.941	0.94						
Korean	0.991	0.93	0.895	0.87						

2.4 Discussion

In accordance with our goal, we chose Mandarin Chinese for our experimental stimuli. We chose the speakers with the following piloted intelligibility: 0.955, 0.899, 0.829. These speakers represent the largest range that also had a favorable distribution. While this range is quite high considering a 0%-100% scale, our piloting so far indicates that this is a somewhat

representative range of accented English speech produced by L2 speakers at universities in the U.S.

Turkish was a poor candidate due to its bimodal distribution, and both Hindi and Russian had small ranges. Korean, however, was a good candidate due to its broad range and favorable distribution. Therefore, Korean has been chosen as the candidate target language for our follow-up study, which will repeat our main experiment using speakers of a different L1.

We are currently collecting in-lab measures of intelligibility for these Korean-accented speakers to verify their adequacy for the follow-up study. So far, 14 participants have been collected, and three were excluded due to eligibility criteria outlined in the Participant section of the main experiment (3.2 Methods). This preliminary dataset $N = 11$ currently reflects the following speaker intelligibilities: 0.992, 0.985, 0.938, 0.925. At this moment, these speakers do not appear to be appropriate candidates for a study replicating the Mandarin speakers. Data collection is ongoing.

Chapter 3: Experiment

The public pre-registration form for this project is available on the Open Science Foundation website at osf.io/ef2bx. This document includes a power analysis and justification for sample size, exclusion criteria, and planned data analyses. Any deviations from this pre-registration are addressed within this document.

3.1 Design

The study was a within-subjects dual-task paradigm featuring a primary listening task and a secondary vibrotactile task. Additionally, a control group (listening task only) was included to perform a between-groups comparison of listening task performance. That is, to determine whether the addition of the secondary task significantly affected performance on the primary task.

For the primary listening task, each participant listened to a sentence that they repeated aloud after the trial ended. Speakers were blocked into four blocks of 30 sentences, with self-paced breaks for participants after each block.

For the secondary vibrotactile task, there was one vibrotactile stimulus per listening task trial. After the audiofile began, a randomized wait-interval passed, and then a small motor held in the participant's non-dominant hand vibrated – the participant immediately indicated the vibration's duration (short, medium, long) using their dominant hand on the keyboard.

3.2 Methods

All procedures were approved by the Washington University in St. Louis Institutional Review Board prior to experimentation.

Participants

Presented here is our initial dataset of $N = 108$ (control group $n = 28$ and dual-tasking group $n = 80$). Participants were undergraduate students from Washington University in St. Louis, recruited through the SONA psychology research participation portal, and compensated with 1 credit for 1 hour of participation.

All eligible participants were monolingual speakers of English from monolingual English households, with clinically normal hearing, and with minimal exposure to Mandarin Chinese and its accent. Eligibility was determined via a verbal pre-screening prior to the experiment, in addition to a Demographic & Language History Questionnaire (DLHQ) conducted at the very end of the experiment. The DLHQ included questions about general demographics (e.g. age, handedness), and the known languages and accents of both the participants and their parents. For their three most proficient languages, participants rated their proficiency in understanding, reading, writing, and listening, using 10-point Likert scales. They also reported the age at which they learned each language, and the geographic location and setting in which this learning took place.

For the first criteria of being a monolingual speaker from a monolingual English household, the following constituted disqualifying responses: any L1 that is not English; acquired a non-English language prior to age 3; acquired a parent's language prior to age 10; acquired a parent's language and two scales in that language score 2 or above; parent has non-native accent (regional is ok); participant has non-native accent (regional is ok); proficiency for any non-English language met or exceeded a total score of 32 (out of 40) Likert points – one point is added to their total if there is a verbal report of an immersion experience. For the criteria of having minimal exposure to Mandarin Chinese and its accent, the following constituted

disqualifying responses: reported knowledge of Mandarin Chinese; verbally confirmed having extensive exposure to a person (e.g., family member, friend, roommate, professor) who speaks English with a Mandarin Chinese accent. This latter response was weighted in terms of extent, frequency, and recency. For example, having a lightly Mandarin-accented group member for an in-class activity 2 years ago was not a disqualification.

Materials

Stimuli

Speech Stimuli.

Each participant heard four male speakers: a native monolingual English speaker and three L2 English speakers with Mandarin Chinese accents. These Mandarin-accented speakers' intelligibilities were determined during piloting: "high" (96%), "medium" (90%), and "low" (83%).

The 120 HINT sentences were divided into 30-sentence blocks. We created four counterbalances in order to distribute the four speakers across the four sentence-blocks. I.e., in the first counterbalance Speaker A produced sentences 1-30, while in the second counterbalance Speaker A produced sentences 31-60, etc. In this way, the experiment collected data for all 480 total audiofiles, but each participant heard all 120 sentences only once, at 30 sentences per speaker. Both the block order and within-block sentence order were randomized for each participant.

For the practice phase, we used six sentences from *The Little Prince*. These were spoken by a different native English speaker in the SpeechBox repository. These audiofiles were segmented and leveled to scale intensity to 65 dB using the same Praat script as the experimental stimuli.

Tactile Task Stimuli.

The vibrotactile stimulation had six levels of wait-interval (100-600ms in 100ms steps), and three levels of duration (short 100ms, medium 200ms, and long 350ms). These parameters were fully randomized per trial per participant. The intention in varying these parameters was to prevent the task from becoming too predictable (e.g., if the vibrations were always presented at 500 ms into the trial), which would decrease the necessary cognitive resources to perform the task and thus reduce sensitivity to the changing demands of the primary listening task.

The wait interval and the durations were chosen such that even in the most extreme case of the longest wait interval (600ms) and longest vibrational duration (350ms), the stimulation from the secondary task was always presented in its entirety during the primary listening task, of which the shortest duration is 1.02 seconds.

Hardware and Software

Audiometry was conducted using a GSI Pello audiometer. Participants' verbal responses to the listening task were recorded using an Olympus WS-853 digital voice recorder. The experiment was run using PsychoPy version 2022.1.4 on an iMac with macOS Monterey version 12.4. The audio stimuli were presented via DT-100 headphones through an Aphex Headpod 4 headphone amplifier. The vibrotactile stimuli were delivered using a custom apparatus made of a small direct current vibration motor, which was controlled by PsychoPy through a LabJack U3 USB DAQ.

Procedure

Upon arrival in the laboratory, participants were greeted and given an informed consent sheet. After indicating verbal consent (no signature required, due to the study's exempt status as per the IRB), pure tone audiometry (1,000 Hz, 2,000 Hz, and 4,000 Hz) was conducted in a

sound-attenuated booth.

After this, participants began the experimental task. They first performed a supervised practice block (12 trials) of the touch-task, where they were presented with a vibrotactile stimulus in their nondominant hand, responded as quickly and accurately as possible with the correct key (j for short vibrations, k for medium, and l for long) on the keyboard using their dominant hand, and then received visual feedback on the accuracy of their response. If they scored 80% or above they advanced to the next phase. If they scored under 80% they returned to the start of the practice to try again. They were able to try as many times as needed to attain the 80% threshold required for participation. Then the audio recorder was turned on, and participants performed a supervised practice of the dual-task (12 total trials). The researcher provided necessary reminders and feedback on the participants' performance, such that by the end of the practice, they are ready to perform the experiment.

Participants were then left alone in the testing room to perform the task. There were four blocks of 30 sentences, with a rest screen between each block. These self-paced rest screens included a reminder to prioritize the listening task.

The trial structure was as follows. First, a white fixation cross appeared on the screen, and an audiofile began playing. Then, after a randomized wait interval, the handheld motor vibrated for one of three possible durations. The participant responds on the keyboard as quickly and accurately as possible. When the audiofile completed some moments after, the white fixation cross remained on screen for two additional seconds (to allow more time to respond on the keyboard, if need be). Then the white fixation cross turned green, and participants repeated aloud the sentence that they just heard, as clearly as possible. They then moved on to the next trial, self-paced, by pressing the spacebar.

After completing 120 trials, the computer prompted the participant to retrieve the researcher. They were then given a Demographic & Language History Questionnaire and a debriefing document. Participants were encouraged to ask questions about the forms and the experiment.

Control Group

The control group's procedure was identical, except: 1) the buzzer has been deactivated, 2) the touch-task practice has been removed, and 3) any instructions relating to the touch-task have been altered (e.g., on the rest screen reminder, we removed the reminder to "prioritize the listening task"). That is, they just listened to and repeated the target sentences.

3.3 Data

Participants' verbal responses were transcribed by undergraduate research assistants and scored using the procedures outlined in the piloting section (2.2 Data). Transcription scoring, data processing, and analyses were all conducted in RStudio 2022.12.0+353 "Elsbeth Geranium" Release for Windows, using R version 4.1.

Before any analyses, 46 participants were removed for failing our eligibility criteria as assessed by the DLHQ (see Participants section of 3.2 Methods).

Data Loss

Key presses made during the wait period (between the end of the audiofile and the fixation cross turning green) were logged with inaccurate RTs due to a programming error.. Therefore, all data for the trials with RTs exceeding the audiofile length were removed (3514 trials). The remaining data therefore represent the subset of trials wherein effort was recorded during online speech processing (7046 trials). In sum, we removed a little less than a third of our total trials, and the remaining data ultimately represent the most theoretically relevant trials.

The data loss affected downstream portions of our data processing. Because of our departure from the pre-registration, here we present the intended data exclusion process and note which criteria were altered.

Trial-Level Exclusion

Pre-registered: "Individual vibrotactile trials will be excluded from a participant's dataset during analysis if they meet any of the following criteria: incorrect response; RT exceeds three median absolute deviations (MADs) below or above that participant's median RT for that condition."

Amendments: The latter criteria was removed, as the constrained RT range due to data loss made this redundant.

Participant Exclusion

Pre-registered: "A participant's entire dataset will be excluded from analysis if they meet any of the following exclusion criteria: vibrotactile task accuracy falls three standard deviations below the condition mean; speech task accuracy falls three standard deviations below the group (e.g. control, dual-task) mean; mean RT for a condition falls three standard deviations below the grand mean for that condition."

Removal of Incomplete Datasets

Pre-registered: "We have defined a complete dataset as meeting the following criteria: Eligible responses are recorded for all speech stimuli, with allowance for one missed response per talker…At least 50% of vibrotactile trials are able to be included in an RT analysis… Incomplete dataset may arise from participant actions or from technical difficulties. In either case, the dataset in its entirety will be replaced."

Amendments: Due to the number of trials impacted by the data loss, for this iteration of the project, we are not able to remove datasets for being incomplete. However, in the upcoming months we will rerun the experiment (with the programming error fixed), replacing all 80 of the participants in the dual-task group, and continuing data collection until the target sample size of $n = 100$ is reached.

Summary

Data from 165 participants were collected: 46 participants were ineligible, one participant's entire dataset had to be excluded when trials were removed due to data loss, and one participant was removed for exceeding our stopping rule in the control condition. Of the remaining 117 eligible participants, we made nine additional exclusions: four due to low listening task accuracy, and six due to low touch task accuracy (one participant was flagged for both). Therefore, we achieved a final $N = 108$, with the control group $n = 28$ and the dual-taskers $n = 80$. After that, 562 individual trials were removed for inaccurate touch task responses, leaving a total of 3360 control trials and 6008 dual-task trials.

3.4 Results

For all of our analyses, Speaker is coded as a factor with Native as the reference group. When appropriate, we follow up by rotating the reference level to investigate specific pairwise comparisons. Trial Number is scaled to assist with model convergence.

Speaker Intelligibility by Group

A fundamental assumption of the dual-task paradigm is that participants are properly prioritizing the listening task. To validate this assumption, we compared the dual-taskers' listening task performance to that of the control group (listening-only). This is visualized in

Figure 1. If performance between these two groups statistically differs, then we must conclude that the presence of the secondary tactile task impacts performance on the primary listening task.

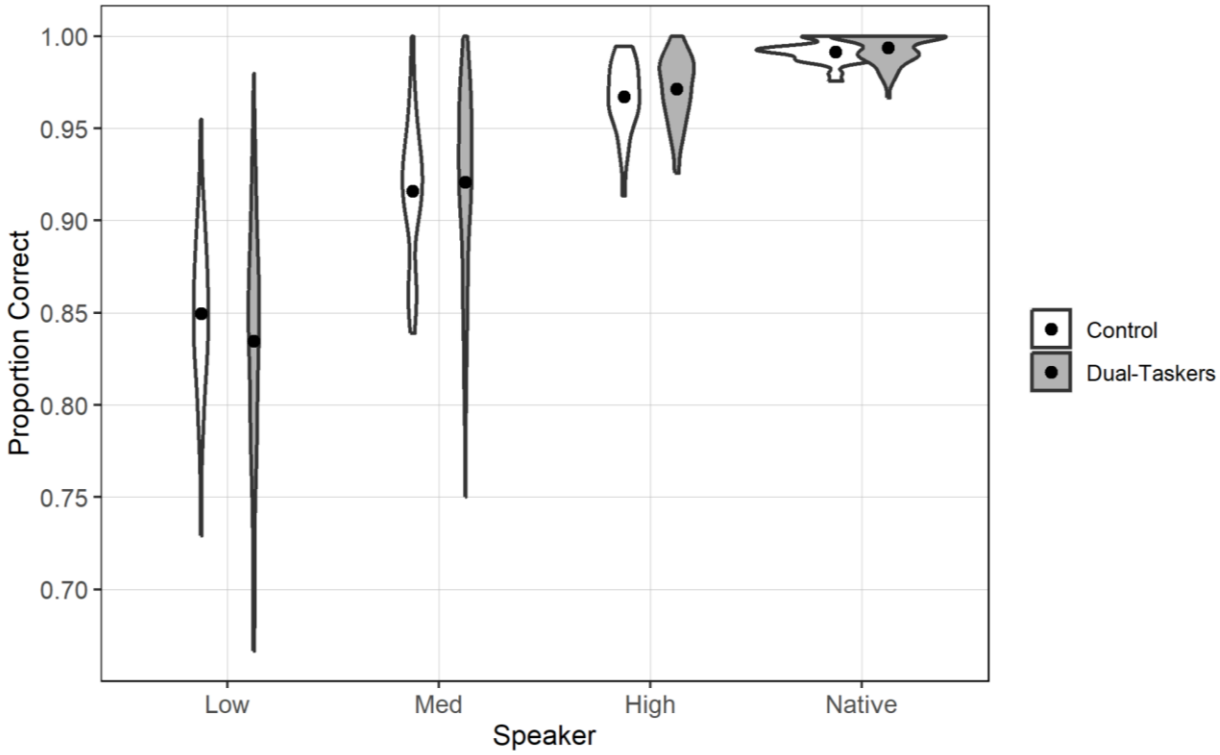

Figure 1. *Speaker Intelligibility by group. Mean intelligibilities as collected from the dual-taskers are: Native 0.99, High 0.97, Medium 0.92, Low 0.84.*

This was assessed using generalized linear regression via glmer(), followed by likelihood ratio testing to verify the appropriateness of our pre-registered model. We modeled the outcome variable of word repetition accuracy as a grouped binomial. This means that for each trial, accuracy is coded as two counts: number of correctly and incorrectly identified words.We included random intercepts for Participant and HINT sentence in each model, and then tested whether our fixed effects improved model fit.

The addition of Counterbalance to the null (no fixed effects) model did not significantly improve model fit ($\chi 2$ (3) = 7.03, p = .071). Adding Group to the null model did not significantly improve model fit ($\chi 2$ (1) = 0.15, p = .696).

The individual additions of Speaker ($\chi 2$ (3) = 2971.4, p < .001) and Trial Number ($\chi 2$ (1) = 15.43, p < .001) to the null model each significantly improved model fit, and including both fixed effects fit better than either of the single-effect models ($\chi 2$ (1) = 20.26, p < .001; $\chi 2$ (3) = 2976.20, p < .001). Model fit was not improved by adding a fixed effect for Counterbalance or Group ($\chi 2$ (3) = 5.19, p = 0.158; $\chi 2$ (1) = 0.0042, p = 0.9486). A model that also included an interaction effect for Speaker and Trial Number did not fit better than the model without the interaction ($\chi 2$ (3) = 1.20, p = 0.754).

Thus, our final model includes the two fixed effects of Speaker and Trial Number. The fixed effect of Speaker demonstrates that each of the non-native speakers are less intelligible than the Native speaker, while the fixed effect of Trial Number indicates listening task improvement over the course of the experiment (adaptation). Rotating the reference Speaker reveals all speakers to have significantly different intelligibility levels.

Vibrotactile Accuracy by Speaker

Figure 2 shows the mean vibrotactile task accuracy across participants per speaker.

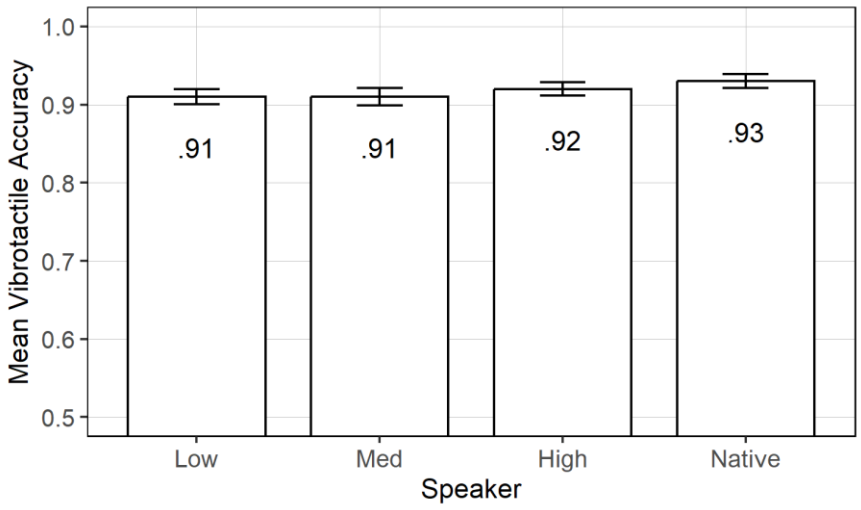

Figure 2. *Mean vibrotactile task accuracy per Speaker, calculated by averaging by-participant condition means; error bars represent the standard error of the mean.*

To determine if vibrotactile accuracy changes depending on Speaker, we used generalized linear mixed effects modeling via glmer(), followed by likelihood ratio testing to verify the appropriateness of our pre-registered model.

We modeled accuracy (0 for inaccurate, 1 for accurate) per trial. When attempting to include our pre-registered random intercepts for Participant and HINT sentence, the model fit was singular. We removed the random intercept of HINT sentence.

The addition of Trial Number did not significantly improve model fit ($\chi 2$ (1) = 2.30, p = .130), indicating that participants were not significantly changing in accuracy across the duration of the task.

The addition of Speaker to the null model did significantly improve model fit ($\chi 2$ (3) = 8.40, p = .039). We therefore modeled accuracy using a fixed effect for Speaker and a random intercept for Participant. Rotating the reference level reveals that the Medium and Low speakers both have significantly lower accuracy than Native. These results, while statistically significant, are minimal.

RT by Speaker

The RT data is visualized in Figure 3 below.

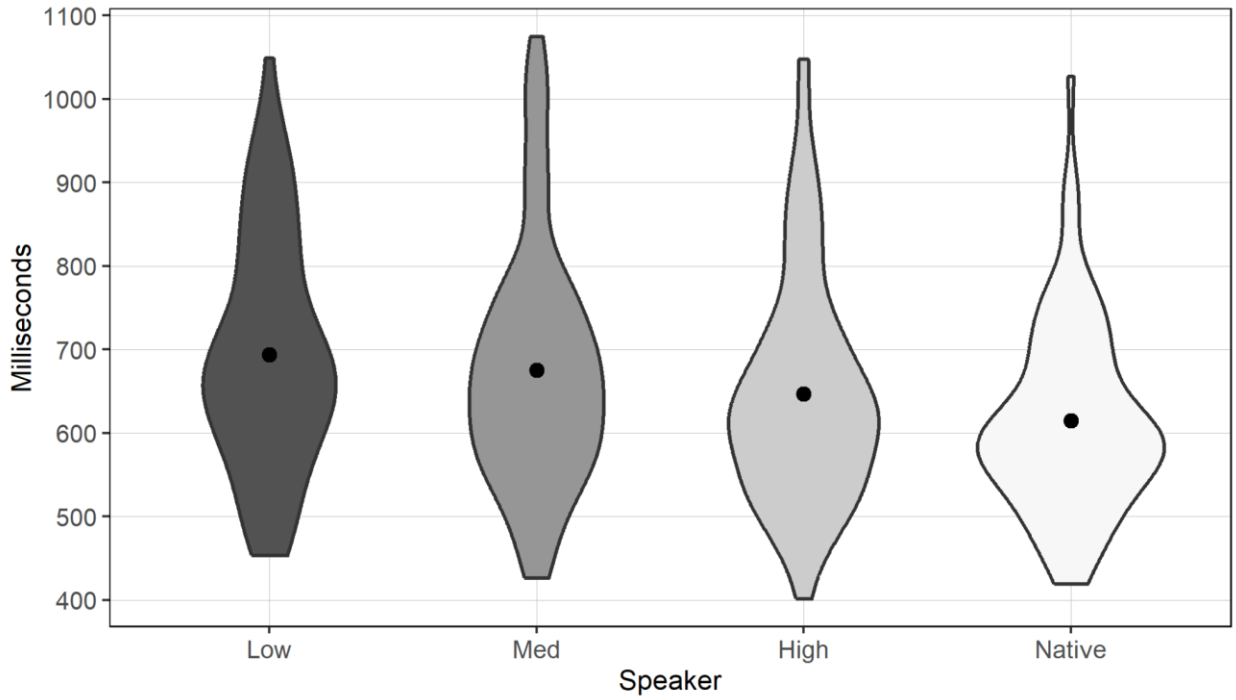

Figure 3. *Mean reaction time per speaker. The dots represent the grand speaker mean, whereas the distributions are of the by-participant condition means. The mean values are: Native 614.30, High 647.01, Med 675.47, Low 693.58.*

To determine how RT changes depending on Speaker, we used linear mixed effects modeling via lmer(), followed by likelihood ratio testing to verify the appropriateness of our pre-registered model. We modeled RTs per trial and included random intercepts for Participant and HINT sentence.

The addition of Counterbalance to the null model did not significantly improve model fit ($\chi2$ (3) = 5.37, p = .147), verifying that our counterbalancing procedures did not add an experimental confound.

The individual additions of fixed effects for Speaker ($\chi2$ (3) = 111.64, p < .001) and Trial Number ($\chi2$ (1) = 68.80, p < .001) each significantly improved model fit. The model with

both of these fixed effects fit better than either of the single-effect models ($\chi 2$ (3) = 56.71, p < .001; $\chi 2$ (3) = 99.55, p < .001). The addition of a fixed effect for Counterbalance did not improve model fit ($\chi 2$ (3) = 5.28, p < .153). The model where Speaker and Trial Number interact fits significantly better than the non-interaction model ($\chi 2$ (3) = 28.17, p < .001).

Therefore, our final model predicts RTs using the interacting fixed effects of Speaker and Trial Number, and the random effects of Participant and HINT sentence[1]. The model output with the Native reference level is shown in Table 2; each Speaker and Speaker*Trial level has significantly different RT estimates than Native. Trial number itself is not significant.

Table 2. *Modeling reaction time with Native reference*

		Reaction Time	
Predictors	*Estimates*	*CI*	*p*
(Intercept)	605.73	573.34 – 638.11	**<0.001**
Speaker [High]	80.18	48.48 – 111.87	**<0.001**
Speaker [Med]	126.84	93.28 – 160.41	**<0.001**
Speaker [Low]	115.59	85.00 – 146.18	**<0.001**
Trial Number	8.50	-12.39 – 29.39	0.425
Speaker [High] * TN	-58.85	-90.95 – -26.76	**<0.001**
Speaker [Med] * TN	-87.98	-122.07 – -53.89	**<0.001**
Speaker [Low] * TN	-58.64	-91.49 – -25.80	**<0.001**
Random Effects			
σ^2	32676.01		
$\tau_{00\ HINT}$	323.17		
$\tau_{00\ PID}$	12429.93		
ICC	0.28		
$N_{\ PID}$	80		
$N_{\ HINT}$	120		
Observations	6008		
Marginal R^2 / Conditional R^2	0.026 / 0.299		

[1] Our pre-registration also included block order, but we did not include this factor since it is redundant with both Speaker and Trial Number included.

Table 3 provides the model output with Medium as the reference level for Speaker rather than Native, which allows us to better understand relationships among performance on the non-native speakers.

Table 3. *Modeling reaction time with Medium reference*

	Reaction Time		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	732.57	697.96 – 767.17	**<0.001**
Speaker [Native]	-126.84	-160.41 – -93.28	**<0.001**
Speaker [High]	-46.67	-79.90 – -13.43	**0.006**
Speaker [Low]	-11.25	-43.37 – 20.86	0.492
TN	-79.48	-103.99 – -54.97	**<0.001**
Speaker [Native] * TN	87.98	53.89 – 122.07	**<0.001**
Speaker [High] * TN	29.12	-4.94 – 63.18	0.094
Speaker [Low] * TN	29.33	-5.23 – 63.89	0.096
Random Effects			
σ^2	32676.01		
$\tau_{00\ HINT}$	323.17		
$\tau_{00\ PID}$	12429.93		
ICC	0.28		
$N_{\ PID}$	80		
$N_{\ HINT}$	120		
Observations	6008		
Marginal R^2 / Conditional R^2	0.026 / 0.299		

Specifically, we see that RT estimates for the High and Medium speakers are significantly different from one another, but the Medium and Low are not. Furthermore, the Speaker*Trial interactions for High and Low are not significantly different from the Medium speaker. In summary, the model estimates reveal that while trial number does not impact participants' RTs when listening to the Native speaker, RTs decrease over the course of listening to the non-native speakers. Notably, this effect is similar across all non-native speakers.

Listening Effort For Intelligible Speech as a Function of Speaker

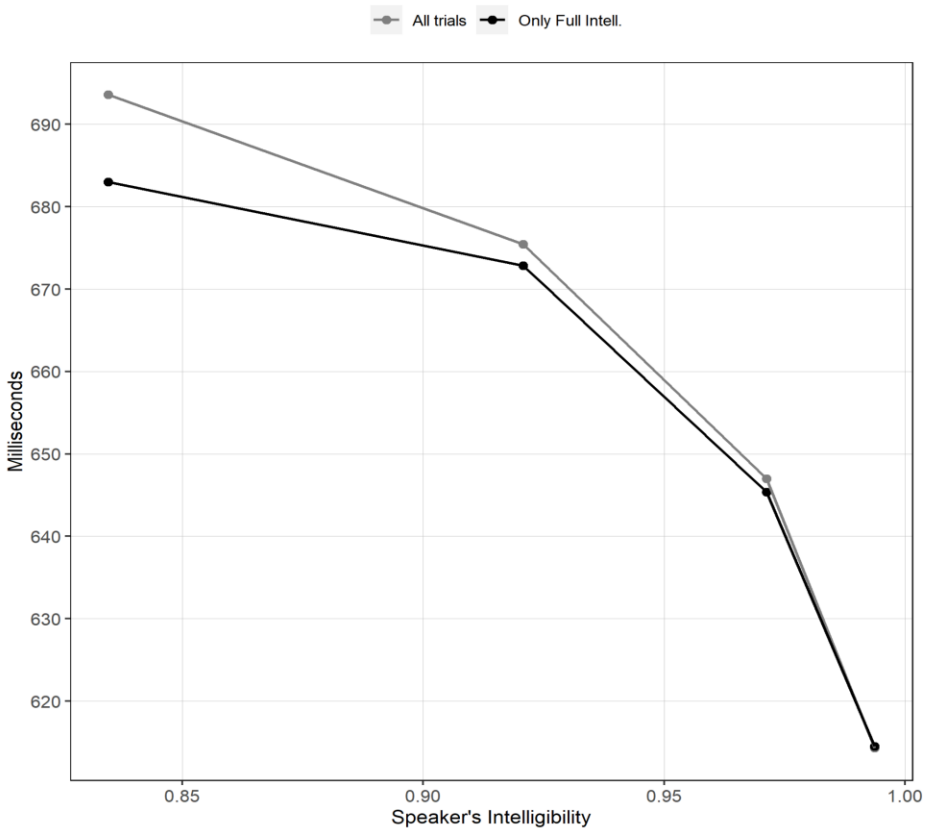

Figure 4. *Mean reaction time as a function of speaker intelligibility, plotted for both the full dataset (6008 trials) and the subset of fully intelligible trials (4842 trials). Calculated by averaging by-participant condition means.*

For this exploratory analysis, we used only the subset of trials that were fully intelligible (i.e., where participants correctly repeated every word of the sentence). We used linear mixed effects modeling via lmer(), followed by likelihood ratio testing to verify the appropriateness of our pre-registered model. The means are shown in Figure 4 above.

We modeled the outcome variable of RT with random effects for Participant and HINT sentence. The addition of Counterbalance ($\chi2$ (3) = 5.25 p = .155) did not significantly improve model fit. The addition of Trial Number ($\chi2$ (1) = 51.22 p < .001) did significantly improve model fit. Most importantly, the addition of Speaker ($\chi2$ (3) = 76.18 p < .001) did significantly

improve model fit.

Including both the fixed effects of Speaker and Trial Number improved model fit compared to either of the single-effect models ($\chi 2$ (1) = 44.87 $p < .001$; $\chi 2$ (3) = 69.83 $p < .001$). The addition of Counterbalance ($\chi 2$ (3) = 4.98 $p = .174$) did not significantly improve model fit. Allowing the fixed effects of Speaker and Trial Number to interact significantly improved model fit further ($\chi 2$ (3) = 21.97 $p < .001$).

Therefore, our final model predicts RTs using the interacting fixed effects of Speaker and Trial Number, and the random effects of Participant and HINT sentence. This is the same model that was derived when modeling the full set of RTs. The model estimates demonstrate that fully intelligible trials demonstrate the same trend of RTs increasing as speaker intelligibility decreases (Table 4). Furthermore, rotating the reference Speaker reveals the same patterns as in the complete dataset: 1) the Medium and Low speaker do not significantly differ in RT, and 2) while all non-native speakers have significant interactions with Trial Number when using the Native reference, these interactions are statistically similar amongst the three non-native speakers.

Table 4. *Modeling reaction time of fully intelligible trials with Native reference.*

		Reaction Time	
Predictors	*Estimates*	*CI*	*p*
(Intercept)	608.44	576.18 – 640.70	**<0.001**
Speaker [High]	71.44	38.02 – 104.85	**<0.001**
Speaker [Med]	117.18	80.94 – 153.41	**<0.001**
Speaker [Low]	117.40	82.44 – 152.37	**<0.001**
TN	5.35	-16.38 – 27.08	0.630
Speaker [High] * TN	-53.44	-87.84 – -19.04	**0.002**
Speaker [Med] * TN	-78.99	-116.19 – -41.78	**<0.001**
Speaker [Low] * TN	-69.74	-107.81 – -31.67	**<0.001**
Random Effects			
σ^2	31253.82		
$\tau_{00\ HINT}$	378.86		
$\tau_{00\ PID}$	11830.42		
ICC	0.28		
$N_{\ PID}$	80		
$N_{\ HINT}$	120		
Observations	4842		
Marginal R^2 / Conditional R^2	0.023 / 0.298		

3.5 Discussion

This study aimed to characterize the relationship between accented speech intelligibility and listening effort. Specifically, we sought to measure listening effort across a range of speakers with varying intelligibility. We first piloted speakers to identify an appropriate set of stimuli, and then implemented a dual-task paradigm to gather behavioral data indexing online listening effort. Our paradigm was successfully implemented in that primary task performance was not impaired by the secondary task, and the RTs reflect that effort generally increased as intelligibility decreased, though effort reached a plateau between the Medium (92%) and Low (83%) speakers. This trend holds even when analyzing only fully intelligible trials.

The dual-task paradigm assumes that the secondary task is a measure of *residual* cognitive resources, leftover after the primary task has been fully attended to. Verifying our paradigm (by showing comparable listening task performance for control and dual-tasking groups) was a necessary first step. If the dual-taskers had performed significantly worse, they may not have been effectively prioritizing the primary listening task.

Since primary task performance did not differ across control and test groups, we are justified in investigating our RTs. We had hypothesized a negative relationship, such that RTs would increase as the speaker intelligibility decreases. In particular, we wanted to characterize the shape of the trend (e.g., peak-shaped versus monotonic versus linear). Visually, our results appear strongly nonlinear, with the largest increase in effort occurring between the Native and High speakers, even though these speakers are the most similar in terms of intelligibility. In contrast, the smallest increase in effort occurred between the Medium and Low speaker, even though these have the least similar intelligibility. This nonlinear effect is so pronounced that the Medium and Low speaker are statistically identical in terms of effort.

Presuming that this local plateau is not a statistical artifact (see Limitations), this finding invites some speculation. Without data for speakers with intelligibilities below our Low speaker, we cannot determine if our local plateau is 1) a global plateau/peak, 2) a local plateau within a monotonic trend, or 3) a local plateau separate from a global peak. At the least, we have ruled out the possibility that a global peak occurs between the speaker intelligibilities of 99% to 84%.

Therefore, the most interesting trend in this data is the steep slope of change in effort from Native to High, in other words, from natively-accented to non-natively-accented, compared to the shallower slopes amongst the non-natively-accented. *This* pattern of effort is nearly categorical in nature, and was identified for both the full dataset and the smaller set of fully intelligible trials. One interpretation is that some noteworthy portion of effort is dedicated to performing (or perhaps, maintaining) a perceptual state change to be successfully receptive to new pronunciations. (See Zheng & Samuel, 2020, for the competing theories of phonemic boundary recalibration versus relaxation.)

Another interpretation is that this trend is driven not by intelligibility per se, but that effort is more closely related to some feature of accent that is somewhat independent from intelligibility. In this way, the feature could vary between speakers in a way that is not indexed when ranking speakers by intelligibility level. Munro and Derwing (1995), for example, found that ratings of accent strength are at least partially independent from intelligibility. Under this account, it could be that our Medium and Low speakers, though nearly 10% different in intelligibility, differ only by, say, 1% in some other phonemic or prosodic quality that is more closely tied to effort.

Comparison with Existing Literature

Overall, our RTs are faster than those in similar dual-task paradigms using a vibrotactile secondary task. In particular, the study we largely based our design on (Brown & Strand, 2019) saw an average RT of 1,045 ms in a 91% intelligible condition (10 dB SNR) and 1,222 ms in a 47% intelligible condition (-4 dB SNR). However, we believe this may reflect task difficulty more than effort. Average vibrotactile accuracy for their Easy and Hard conditions were 79% and 75% respectively, whereas our accuracies per speaker were 93%, 92%, 91%, 91%. The relative ease of our task is likely a contributing factor for why our RTs are comparatively quick.

However, we believe that the ease of the task may not necessarily imply a lack of sensitivity. Wu et al. (2016) employed two dual-task paradigms: one with an "Easy" secondary task of reacting to a visual probe, and one with a "Hard" secondary task of performing a Stroop task. The primary listening task in each paradigm was a speech recognition task using HINT sentences, presented at various SNRs so as to construct a psychometric function. Across the range of SNRs, the RTs for the Hard task visually range from around 800-1,1000ms, and in the Easy task range from around 300-400ms. However, Wu et al. note that the *shapes* of the two curves are highly similar when rescaled for peak height. Despite the difference in task difficulty and relative RT, the two tasks recorded the same general relationship between SNR and effort.

Our exploratory investigation of the RTs in fully intelligible trials expands on previous literature which has also shown increased effort in fully intelligible trials by speakers with different accents (Brown et al. 2020; McLaughlin & Van Engen 2020), as measured by pupillometry. Our results reveal that the increased effort that is deployed for processing fully intelligible non-natively accented speech differs across levels of baseline speaker intelligibility. (It is also of interest that our relatively cruder behavioral paradigm is sensitive enough to detect

this trend.) This finding is well-situated within the framework that listening effort is functional: what we've managed to measure in the fully intelligible subset is the effort that was put forth to *maintain* full intelligibility for each speaker.

Limitations and Future Directions

Data loss is likely to have disproportionately impacted longer RTs, which means that the traditionally right-skewed tail has been amputated. In interpreting our data, this would result in overly-conservative estimates of effort. It is possible that the similarity in effort for our Medium and Low speakers is therefore the result of RT suppression.

In addition, this dataset is currently underpowered, due not only to data loss but also to sample size. We pre-registered 100 dual-tasking participants, which was intended to produce around 12,000 trials. Our current dataset consists of 6008 trials. Further, those 12,000 trials were powered to detect RT differences of roughly 25 ms. We believed this to be a conservative estimate, based on the RTs in previous literature. In particular, the RT difference in Brown and Strand (2019) was 69 ms. However, we have much smaller mean differences between conditions than anticipated (32.71 ms, 28.46 ms, 18.11 ms), and do not have the proper power to assess them. Perhaps with greater power, Medium and Low would be statistically different.

Thus, our next step is to rerun the experiment with the coding error fixed. We have already corrected the code, and in-lab piloting has confirmed that this has resolved the data loss. A new full dataset of 100 dual-taskers, with all RTs intact, will result not only in more power, but in RTs that will likely be more similar to those in the existing literature.

As discussed earlier (2.4 Discussion), we are also already collecting speaker intelligibility data in pursuit of iterating upon the present study using another L1 accent (Korean). This project has been publicly pre-registered on OSF (osf.io/kbu46).

An ancillary project has also been publicly pre-registered on OSF (osf.io/nrp49), designed to assess the temporal sensitivity of the present study. We felt it important to verify that the time course of speech processing does not significantly impact effort as measured early in a sentence versus late in a sentence. The project uses the audio stimuli from the Low speaker in the present study, and manipulates within-participants whether the secondary task occurs "early" or "late" in the audio stimulus. There are two counterbalanced groups, such that for each audiofile, one group performs an "early" trial, and the other performs a "late" trial. If no significant differences are found for early trial versus late trial RTs, then the temporal effect (if any) is too small to have impacted the results of the present study. Further, the results of that project can inform how to proceed with secondary task timing when comparing RTs arising from speakers with different speaking speeds, which is a potential confound.

Further Considerations

The most compelling further considerations concern the richness of the dataset. First, presenting more Mandarin-accented speakers would reduce extrapolation within the curve, leading to more precise estimates of trend parameters. If possible, it would be particularly informative to find speakers with lower intelligibility, expanding the range of our dataset. Second, collecting data using more languages – for both L1 accent and L2 utterances – to qualify the generalizability of our findings. Third, measuring listening effort for these same stimuli using different methodologies (e.g. pupillometry, subjective ratings, even other dual-tasks). This is an important consideration, as our conclusions about the strength of our effect might be limited by the sensitivity of this particular dual-task. All together, these immediate steps are essential for drawing robust conclusions about the relationship between accented speech intelligibility and listening effort.

Additionally, one promising avenue of investigation is of the speakers themselves. For example: there is only a 5% difference between our High and Medium speakers, but they elicit a significant difference in effort; in comparison, there is nearly a 10% difference between our Medium and Low speakers, but they elicit a non-significant difference in effort. There are many potential explanations for this, one being: perhaps this effect is driven not by the speaker's intelligibility, but on some feature(s) related to the speaker themselves (e.g., prosody). An investigation of the properties of the speaker's utterances could illuminate variables that are more directly related to effort than intelligibility per se. This would be a monumental step forward towards addressing current discourse on how accent relates to speech perception and processing.

References

Anderson Gosselin, P., & Gagné, J.-P. (2011). Older adults expend more listening effort than young adults recognizing speech in noise. *Journal of Speech, Language, and Hearing Research: JSLHR*, *54*(3), 944–958. https://doi.org/10.1044/1092-4388(2010/10-0069)

Bent, T., & Bradlow, A. R. (2003). The interlanguage speech intelligibility benefit. *The Journal of the Acoustical Society of America*, *114*(3), 1600–1610. https://doi.org/10.1121/1.1603234

Borrie, S. A., Barrett, T. S., & Yoho, S. E. (2019). Autoscore: An open-source automated tool for scoring listener perception of speech. *The Journal of the Acoustical Society of America*, *145*(1), 392. https://doi.org/10.1121/1.5087276

Bradlow, A. R. (n.d.). ALLSSTAR: Archive of L1 and L2 scripted and spontaneous transcripts and recordings. Retrieved from https://speechbox.linguistics.northwestern.edu/#!/?goto=allsstar

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*(2), 707–729. https://doi.org/10.1016/j.cognition.2007.04.005

Broadbent, D. E. (1958). The effects of noise on behaviour. In D. E. Broadbent, *Perception and communication.* (pp. 81–107). Pergamon Press. https://doi.org/10.1037/10037-005

Brown, V. A., McLaughlin, D. J., Strand, J. F., & Van Engen, K. J. (2020). Rapid adaptation to fully intelligible nonnative-accented speech reduces listening effort. *Quarterly Journal of Experimental Psychology*, *73*(9), 1431–1443. https://doi.org/10.1177/1747021820916726

Brown, V. A., & Strand, J. F. (2019). About face: Seeing the talker improves spoken word recognition but increases listening effort. *Journal of Cognition*, *2*(1), 44. https://doi.org/10.5334/joc.89

Ferguson, S. H., Jongman, A., Sereno, J. A., & Keum, K. A. (2010). Intelligibility of foreign-accented speech for older adults with and without hearing loss. *Journal of the American Academy of Audiology*, *21*(3), 153–162. https://doi.org/10.3766/jaaa.21.3.3

Francis, A. L., & Love, J. (2020). Listening effort: Are we measuring cognition or affect, or both? *WIREs Cognitive Science*, *11*(1). https://doi.org/10.1002/wcs.1514

Fraser, S., Gagné, J.-P., Alepins, M., & Dubois, P. (2010). Evaluating the effort expended to understand speech in noise using a dual-task paradigm: The effects of providing visual speech cues. *Journal of Speech, Language, and Hearing Research*, *53*(1), 18–33. https://doi.org/10.1044/1092-4388(2009/08-0140)

Gagné, J.-P., Besser, J., & Lemke, U. (2017). Behavioral assessment of listening effort using a dual-task paradigm. *Trends in Hearing*, *21*, 2331216516687287. https://doi.org/10.1177/2331216516687287

Kahneman, D. (1973). *Attention and effort*. Prentice-Hall.

Lunner, T., Rudner, M., & Rönnberg, J. (2009). Cognition and hearing aids. *Scandinavian Journal of Psychology*, *50*(5), 395–403. https://doi.org/10.1111/j.1467-9450.2009.00742.x

Mackersie, C. L., & Cones, H. (2011). Subjective and psychophysiological indexes of listening effort in a competing-talker task. *Journal of the American Academy of Audiology*, *22*(2), 113–122. https://doi.org/10.3766/jaaa.22.2.6

MacPherson, A., & Akeroyd, M. A. (2014). Variations in the slope of the psychometric functions for speech intelligibility: A systematic survey. *Trends in Hearing*, *18*, 233121651453772. https://doi.org/10.1177/2331216514537722

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes*, *27*(7–8), 953–978. https://doi.org/10.1080/01690965.2012.705006

McCoy, S. L., Tun, P. A., Cox, L. C., Colangelo, M., Stewart, R. A., & Wingfield, A. (2005). Hearing loss and perceptual effort: downstream effects on older adults' memory for speech. *The Quarterly Journal of Experimental Psychology. A, Human Experimental Psychology*, *58*(1), 22–33. https://doi.org/10.1080/02724980443000151

McGarrigle, R., Munro, K. J., Dawes, P., Stewart, A. J., Moore, D. R., Barry, J. G., & Amitay, S. (2014). Listening effort and fatigue: what exactly are we measuring? A British Society of Audiology Cognition in Hearing Special Interest Group "white paper." *International Journal of Audiology*, *53*(7), 433–440. https://doi.org/10.3109/14992027.2014.890296

McLaughlin, D. J., & Van Engen, K. J. (2020). Task-evoked pupil response for accurately recognized accented speech. *The Journal of the Acoustical Society of America*, *147*(2), EL151–EL156. https://doi.org/10.1121/10.0000718

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, *45*(1), 73–97. https://doi.org/10.1111/j.1467-1770.1995.tb00963.x

Nilsson, M., Soli, S. D., & Sullivan, J. A. (1994). Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise. *The Journal of the Acoustical Society of America*, *95*(2), 1085–1099. https://doi.org/10.1121/1.408469

Peelle, J. E. (2018). Listening effort: How the cognitive consequences of acoustic challenge are reflected in brain and behavior. *Ear & Hearing*, *39*(2), 204–214. https://doi.org/10.1097/AUD.0000000000000494

Pichora-Fuller, M. K., Kramer, S. E., Eckert, M. A., Edwards, B., Hornsby, B. W. Y., Humes, L. E., Lemke, U., Lunner, T., Matthen, M., Mackersie, C. L., Naylor, G., Phillips, N. A., Richter, M., Rudner, M., Sommers, M. S., Tremblay, K. L., & Wingfield, A. (2016). Hearing impairment and cognitive energy: The Framework for Understanding Effortful Listening (FUEL). *Ear and Hearing*, *37 Suppl 1*, 5S-27S. https://doi.org/10.1097/AUD.0000000000000312

Picou, E. M., & Ricketts, T. A. (2014). The effect of changing the secondary task in Dual-Task Paradigms for measuring Listening effort. *Ear and Hearing*, *35*(6), 611. https://doi.org/10.1097/AUD.0000000000000055

Porretta, V., & Tucker, B. V. (2019). Eyes wide open: Pupillary response to a foreign accent varying in intelligibility. *Frontiers in Communication*, *4*. https://doi.org/10.3389/fcomm.2019.00008

Rabbitt, P. M. A. (1968). Channel-capacity, intelligibility and immediate memory. *Quarterly Journal of Experimental Psychology*, *20*(3), 241–248. https://doi.org/10.1080/14640746808400158

Rogers, C. L., Dalby, J., & Nishi, K. (2004). Effects of noise and proficiency on intelligibility of Chinese-accented English. language and speech, 47(2), 139–154. https://doi.org/10.1177/00238309040470020201

Romero-Rivas, C., Martin, C. D., & Costa, A. (2015). Processing changes when listening to foreign-accented speech. *Frontiers in Human Neuroscience, 9*. https://doi.org/10.3389/fnhum.2015.00167

Rönnberg, J., Lunner, T., Zekveld, A., Sörqvist, P., Danielsson, H., Lyxell, B., Dahlström, Ö., Signoret, C., Stenfelt, S., Pichora-Fuller, M. K., & Rudner, M. (2013). The Ease of Language Understanding (ELU) model: Theoretical, empirical, and clinical advances. *Frontiers in Systems Neuroscience, 7*. https://doi.org/10.3389/fnsys.2013.00031

Sarampalis, A., Kalluri, S., Edwards, B., & Hafter, E. (2009). Objective measures of listening effort: Effects of background noise and noise reduction. *Journal of Speech, Language, and Hearing Research: JSLHR*, *52*(5), 1230–1240. https://doi.org/10.1044/1092-4388(2009/08-0111)

Schmid, P. M., & Yeni-Komshian, G. H. (1999). The effects of speaker accent and target predictability on perception of mispronunciations. *Journal of Speech, Language, and Hearing Research*, *42*(1), 56–64. https://doi.org/10.1044/jslhr.4201.56

Seeman, S., & Sims, R. (2015). Comparison of psychophysiological and dual-task measures of listening effort. *Journal of Speech, Language, and Hearing Research*, *58*(6), 1781–1792. https://doi.org/10.1044/2015_JSLHR-H-14-0180

Strand, J. F., Brown, V. A., Merchant, M. B., Brown, H. E., & Smith, J. (2018). Measuring listening effort: Convergent validity, sensitivity, and links with cognitive and personality measures. *Journal of Speech, Language, and Hearing Research*, *61*(6), 1463–1486. https://doi.org/10.1044/2018_JSLHR-H-17-0257

Van Engen, K. J., & Peelle, J. E. (2014). Listening effort and accented speech. *Frontiers in Human Neuroscience*, *8*. https://doi.org/10.3389/fnhum.2014.00577

Wilson, E. O., & Spaulding, T. J. (2010). Effects of noise and speech intelligibility on listener comprehension and processing time of Korean-accented English. *Journal of Speech, Language, and Hearing Research*, *53*(6), 1543–1554. https://doi.org/10.1044/1092-4388(2010/09-0100)

Winn, M. B., & Teece, K. H. (2021). Listening effort is not the same as speech intelligibility score. *Trends in Hearing*, *25*, 233121652110276. https://doi.org/10.1177/23312165211027688

Wu, Y.-H., Stangl, E., Zhang, X., Perkins, J., & Eilers, E. (2016). Psychometric functions of Dual-Task paradigms for measuring listening effort. *Ear and Hearing*, *37*(6), 660–670. https://doi.org/10.1097/AUD.0000000000000335

Zekveld, A. A., & Kramer, S. E. (2014). Cognitive processing load across a wide range of listening conditions: Insights from pupillometry. *Psychophysiology*, *51*(3), 277–284. https://doi.org/10.1111/psyp.12151

Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2010). Pupil response as an indication of effortful listening: The influence of sentence intelligibility. *Ear and Hearing*, *31*(4), 480–490. https://doi.org/10.1097/AUD.0b013e3181d4f251

Zheng, Y., & Samuel, A. G. (2020). The relationship between phonemic category boundary changes and perceptual adjustments to natural accents. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(7), 1270–1292. https://doi.org/10.1037/xlm0000788