NORTHWESTERN UNIVERSITY

Learning to Decipher Speech in Noise and the Impact of Sleep on Learning

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Psychology

By

Adrianna Bassard

EVANSTON, ILLINOIS

September 2023

ABSTRACT

Conversation is an important part of human life. Given globalization and the numerous languages around the world, it is increasingly likely that we will be communicating with others speaking in their second language (L2) rather than their first language (L1). In these situations, communication may require more effort. However, people can become attuned to different speakers and understand others quite well, even in suboptimal listening environments. Previous research has shown that exposure to L2 speech from many talkers of one language background can give the listener experience with a variety of sounds to gain a more generalized understanding of speech sounds. This information allows listeners to better understand new talkers, even if they are of novel talkers from different language backgrounds. In Experiment 1, we explored how brief training in single- vs multiple-talker conditions affected later performance. We found that participants were able to improve with training and that this training allowed for later generalization to novel talkers when tested roughly 11 hours later. Interestingly, the ability to generalize to new talkers depended on individual speaker intelligibility as well as the number of talkers experienced during training. Experiment 2 built on these findings employing a technique known as targeted memory reactivation to explore the role of sleep in this generalization process after training on a low-intelligibility talker. Based on our results, we hypothesize that long and undisturbed sleep may support this type of generalization learning. Overall, this research adds to a growing literature on speech perception helping us better understand the nuances of human communication.

ACKNOWLEDGEMENTS

I would first like to thank my advisor, Dr. Ken A. Paller for his guidance during my time in graduate school. I could not have completed this accomplishment without support from you and my other dissertation committee members, Dr. Ann Bradlow, Dr. Marcia Grabowecky, and Dr. Paul Reber. I would also like to thank my wonderful collaborator Dr. Cat Han. You were instrumental in seeing this dissertation through and I couldn't have asked for a more enjoyable final year of graduate school. Dr. Annie Yetzer and Dr. Elizabeth Dworak, thank you for always rooting for me – I'm happy I can now say I have joined the PhD club.

To my mother, Kathy Cruz Zumhingst, thank you for your unwavering support during the most joyous times and the most difficult times. I would not be where I am today without you, your sacrifices, and your love. Thank you to my father, Earl Bassard, a proud U.S. Marine, for your support in this journey. Also thank you for passing down your drive and your writing skills to me. To my sisters, Casie Arnold (my role model), Ariel Bassard (my best friend), and Katrina (the best big sister ever), thank you for keeping me focused and for the heart to hearts. I felt like I was near you all, even when I was hours away.

I would also like to thank some of my dearest friends and biggest cheerleaders: Tayler Gutierrez, Esq., Stella Auer, Madisen Hursey, Broderick Hollins, Maddy Dynes, Alexis Huron, and Chelsea Robinson. I will always remember this time fondly because of you and I am forever grateful for your friendship.

TABLE OF CONTENTS

ABSTRACT

ACKNOWLEDGEMENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: GENERAL INTRODUCTION

- **1.1 Memory consolidation during sleep**
- **1.2 Targeted memory reactivation**
- **1.3 Memory reactivation for declarative and procedural tasks**
- **1.4 Influence of sleep and sleep manipulation on the perception of auditory speech**

CHAPTER 2: EXPERIMENT 1

- **2.1 Introduction**
- **2.2 Methods**
  - 2.2.1 Participants
  - 2.2.2 Stimuli
  - 2.2.3 Behavioral task
  - 2.2.4 Experimental design
  - 2.2.5 Behavioral analysis
- **2.3 Results**
  - 2.3.1 Training
  - 2.3.2 Training vs test vs Control
  - 2.3.3 Generalization
- **2.4 Discussion**

CHAPTER 3: EXPERIMENT 2

- **3.1 Introduction**
- **3.2 Methods**
  - 3.2.1 Participants
  - 3.2.2 Stimuli
  - 3.2.3 Behavioral tasks

3.2.4 Afternoon nap

3.2.5 Experimental design

3.2.6 Behavioral analysis

3.2.7 EEG recording and analysis

3.3 Results

3.3.1 Training

3.3.2 Training vs test

3.3.3 Generalization

3.3.4 Delayed test

3.3.5 Describing sleep

3.3.6 Effect of sleep on performance

3.4 Discussion

CHAPTER 4: IN-DEPTH STATISTICAL ANALYSIS

4.1 Introduction

4.2 Methods

4.2.1 Data

4.2.2 Model building

4.3 Results

4.3.1 Chapter 2: Experiment 1 online model

4.3.2 Chapter 3: Experiment 2 in-lab model

4.3.3 Experiment 1 and 2 combined data model

4.3.4 Testing model assumptions

4.4 Discussion

CHAPTER 5: GENERAL DISCUSSION

5.1 Thesis overview and findings

5.2 Future directions

5.3 Conclusion

Chapter 6: REFERENCES

Chapter 7: APPENDICES

LIST OF FIGURES

Figure 2.1 Experiment 1 study design

Figure 2.2 Transcription training performance separated by training group

Figure 2.3 Average transcription performance at training and test for different training types

Figure 2.4 Performance on novel and trained talkers for different training types

Figure 2.5 Test performance on novel and trained groups separated by exposure group

Figure 2.6 Test performance for each talker

Figure 3.1 Experiment 2 study design

Figure 3.2 Training performance based on TMR group

Figure Training and test performance and difference scores by TMR group

Figure 3.4 Mean performance for participants depending on group and talker

Figure 3.5 Control group performance

Figure 3.6 Correlation of TMR duration by improvement

Figure 3.7 Correlation of time asleep and sleep disruption on improvement

Figure 4.1 Experiment 1 MLM hierarchy

Figure 4.2 Experiment 2 MLM hierarchy

Figure 4.3 MLM hierarchy for combined data

Figure 4.3 Visualization of Experiment 1 MLM variables

Figure 4.4 Visualization of Experiment 2 MLM variables

Table 4.6 Models tested to fit combined data

Figure 4.5 Significant MLM statistics for combined data

Figure 4.6 Linearity of variance Experiment 1

Figure 4.7 Homogeneity of variance Experiment 1

Figure 4.8 Plot of standardized residuals Experiment 1

Figure 4.9 Linearity of variance Experiment 2

Figure 4.10 Homogeneity of variance Experiment 2

Figure 4.11 Plot of standardized residuals Experiment 2

Figure 4.12 Linearity of variance combined data

Figure 4.13 Homogeneity of variance combined data

Figure 4.14 Plot of standardized residuals combined data

List of Tables

Table 2.1 Transcription performance at training and test separated by group

Table 2.2 Test performance for each talker

Table 2.3 Significant pairwise comparisons based on talker

Table 3.1 Performance at training and test based on TMR group

Table 3.2 Test performance for each talker

Table 3.3 Sleep measures and statistics between groups

Table 3.4 Between group sleep statistics

Table 3.5 Statistics on correlations between sleep measures and improvement

Table 3.6 Statistics on correlations between sleep measures and improvement for the cued group

Table 4.1 Possible variables included in MLM

Table 4.2 Models tested to fit Experiment 1 data

Table 4.3 Significant MLM statistics for Experiment 1

Table 4.4 Models tested to fit Experiment 2 data

Table 4.5 Significant MLM statistics for Experiment 2

Table 4.6 Models tested to fit combined data

Table 4.7 Significant MLM statistics for combined data

Table S1. Experiment 1 Test performance for each training group by talker

CHAPTER 1

GENERAL INTRODUCTION

1.1 Memory Consolidation During Sleep

Humans learn many things each day and sleep each night. Although day and night are idiomatically as different as it gets, we now suspect that daily learning and sleep are not unconnected. How sleep contributes to the acquisition of new information is currently under active investigation. Indeed, the literature reviewed below shows that much progress has been made in addressing this question in recent years.

Damage to the medial temporal region of the brain, and in particular to the hippocampus, results in serious memory impairments, especially for declarative or explicit memory which is broadly considered to support the recall of facts and events (Squire & Zola, 1996). Studies of the amnesia that results from such brain damage have been immensely informative for our understanding of human memory. These brain areas play a critical role for information to be stored as long-lasting memories (Eichenbaum, 2004; Squire & Wixted, 2011). Whereas patients with hippocampal damage have trouble forming new, long-term declarative memories, remote memory seems to be intact. This phenomenon is explained by the standard system consolidation model, which suggests that the hippocampus rapidly stores new information which is slowly, over time, stored in the neocortex (McClelland, McNaughton, & O'Reilly, 1995). In fact, the hippocampus is thought to guide the reorganization of memory to create a more stable representation in the neocortex (Squire, Genzel, Wixted, & Morris, 2015).

Neural mechanisms underlying memory storage and consolidation have been studied extensively in animal models (O'keefe & Nadel, 1978). A branch of research important for this was understanding the role of the hippocampus for navigation in rodents. In one experiment,

Pavlides and Winson (1989) found that cells in the hippocampus that fired during a wake navigation task exhibited increased firing during sleep compared to those cells that did not fire during wake learning. This experiment supported the idea that previously learned information is processed during sleep. It was not clear, however, whether the cells fired in an organized fashion. To examine this further, Wilson and McNaughton (1994) recorded place cells in the rodent hippocampus during a maze-learning task that was reinforced with a food reward. They were curious about whether the same neural ensembles that fired during wake were also active during sleep in a coherent fashion. They found that ensembles of cells that fired together as the rodents completed the spatial task also fired together during sleep. They hypothesized that this so-called hippocampal replay is a sign of memory consolidation and important for the process of memory transfer from the hippocampus to the neocortex.

Memory consolidation during sleep has also been explored in humans. Although place-cell replay has not been observed, other techniques have allowed memory researchers to explore the processes that occur during sleep to promote human memory. For example, Peigneux et al. (2004) explored hippocampal activation in humans and the effect of this activation on memory. Here, participants learned to navigate a virtual town. Cerebral blood flow was measured during learning and during subsequent sleep. Areas of the hippocampus that were active during learning were also more likely to be active during slow-wave sleep (SWS). Interestingly, more hippocampal activity during SWS tended to be linked with a greater memory improvement. The authors suggest that this activity during sleep is an indicator of offline memory processing (Peigneux et al., 2004). A later study combined EEG with fMRI while participants completed a visual perceptual learning task (texture discrimination task) to record sleep and brain activation (Yotsumoto et al., 2009). Areas of V1 that were active during the texture discrimination task

were examined during subsequent sleep and compared to areas that were inactive. It was determined that the region of V1 that was active during training was also active during non-REM (NREM) sleep. Importantly, improved performance after sleep was correlated with the amount of activation, suggesting that memory processing continues during sleep and affects later performance (Yotsumoto et al., 2009).

As the literature has grown, sleep came to be considered a period important for the stability of memory (Stickgold, 2005). Memory consolidation during sleep has primarily been studied in deep sleep and is thought to be a result of endogenous memory reactivation (Ken A Paller, Creery, & Schechtman, 2021; K.A. Paller, Mayes, Antony, & Norman, 2020). For successful consolidation, hippocampal-neocortical interactions are believed to be strengthened during non-REM sleep resulting in better memory for reactivated episodes (Diekelmann & Born, 2010). In order to determine the underlying neural mechanisms that support memory consolidation, researchers have sought to determine whether memory improvements correlate with neural oscillations during sleep, as there are specific, neural signatures characteristic of each sleep stage (Berry et al., 2017). Non-REM (NREM) sleep is comprised of stage 1 (N1), stage 2 (N2), and stage 3 (N3). N3 is also known as slow-wave sleep (SWS). Compared to drowsy wakefulness, N1 is marked by a decrease in alpha waves (8-12 Hz) and an increase in theta (4-8 Hz) waves. N2 is marked by the appearance of 11-16 Hz thalamocortical spindles and an occasional K-complex (a high-amplitude waveform with a positive, sharp wave following a negative, sharp wave). SWS is comprised of neocortical slow oscillations (SOs), 0.5 to 2.0 Hz activity, with a peak-to-peak amplitude of >75μV for at least 6 seconds visible by scalp electroencephalography (EEG). A final signature that has been explored are hippocampal sharp-wave ripples (150-250 Hz). Though not visible from the scalp, these SWRs are considered

important for memory consolidation during sleep, in combination with spindles and slow-oscillations (Born, Rasch, & Gais, 2006; Staresina et al., 2015).

Meier-Koll, Bussmann, Schmidt, and Neuschwander (1999) sought to determine whether sleep architecture changed as a function of learning. Participants in their experiment were faced with either a simple maze, a complex maze, or no maze. Interestingly, participants who navigated a maze before sleep had significantly more spindles than those who did not. Memory for a visuospatial task was also found to correlate with the number of spindles that occurred during subsequent sleep (Z Clemens, Fabo, & Halasz, 2005; Zsófia Clemens, Fabó, & Halász, 2006). Finally, the number of spindles detected during overnight sleep correlated with performance on a verbal free recall task (Z Clemens et al., 2005) as increased spindle activity measured overnight correlated with the number of words recalled during a word-pair association task (Schabus et al., 2004). Antony, Schönauer, Staresina, and Cairney (2019) proposed a framework to explain how sleep spindles may support memory. They hypothesized that sleep spindles are important for the reinstatement of memories while the refractory period allows for further processing.

Another neural signal often implicated in memory consolidation are slow oscillations (SO's) which occur primarily during N3 sleep. Huber, Felice Ghilardi, Massimini, and Tononi (2004) hypothesized that if slow-oscillations support memory, learning should increase slow-wave activity (SWA). In their experiment, participants completed a version of a motor learning task that either elicited activation in right parietal areas or did not. High-density EEG was recorded for two hours following subsequent overnight sleep. Overall, local slow-wave power was indeed found to be greater for those who completed the task hypothesized to elicit activation. Importantly, performance after sleep was correlated with slow-wave activity (Huber

et al., 2004). In a follow-up study, other researchers set out to find a causal role for SWA through slow-wave deprivation (Crupi et al., 2009). Participants completed a motor learning task and during sleep, were exposed to tones timed to decrease SWA or to tones played during other stages of NREM sleep so as to not influence SWA. In accordance with the hypothesis, those in the control condition exhibited improved performance the following day while those in the slow-wave deprivation condition did not (Crupi et al., 2009).

Buzsaki (1998) suggested that information from the neocortex reaches the hippocampus during aroused states but that during sleep, memory is reestablished in the neocortex from the hippocampus via sharp-wave ripples (SWR's) which occur mainly during N3 sleep. In one animal study, researchers were curious about whether SWR-events would occur more frequently if rodents learned an odor-reward association task (Eschenko, Ramadan, Mölle, Born, & Sara, 2008). An examination of neural activity during subsequent sleep showed that learning produced an increase in the number of SWRs. This increase was only found, however, if the rodent learned the odor-reward association before sleep (Eschenko et al., 2008). Finally, as mentioned above, increased activity of the hippocampus during sleep, which corresponded to areas active during learning, correlated with improvement on a route-learning task (Peigneux et al., 2004).

Overall, there is a rich literature that highlights the importance of slow-oscillations, sleep spindles, and sharp-wave ripples during sleep for memory consolidation and how these oscillations work in tandem to support memory consolidation. For example, Molle, Yeshenko, Marshall, Sara, and Born (2006) set out to determine how together, SWRs, and SOs promoted memory consolidation. From the rodent sleeping data, the researchers explored possible interactions between these two mechanisms and found that the up-state of neocortical SOs promoted hippocampal SWRs while the down-state hindered this SWR activity (Molle et al.,

2006). Because spindles and SOs can be measured from the scalp, many experiments have found spindle-SO coupling events during sleep to be beneficial for memory (e.g., Hahn, Heib, Schabus, Hoedlmoser, & Helfrich, 2020; Mikutta et al., 2019; Muehlroth et al., 2019).

Exploring these SWR-spindle-SO events in humans has since become a research area of interest. In one experiment, researchers tested whether this hierarchical nesting of oscillations could be observed during sleep using intracranial EEG recordings in patients with epilepsy (Staresina et al., 2015). They found that thalamocortical spindles were indeed influenced by SO phase; spindles tended to occur in the SO up-state. Further, hippocampal ripples were found to reside in the troughs of the spindles. This nesting was determined to be temporally organized in a top-down fashion as predicted, giving further credence to the idea that coupled SO, spindle, and SWR events function as a mechanism for hippocampal-neocortical interaction (Staresina et al., 2015). Staresina and colleagues theorized that neocortical slow oscillations act to synchronize hippocampal reactivation, which coincides with hippocampal SWRs that are crucial for memory reactivation and consolidation. These SWRs nest within thalamocortical spindles, which are further nested within slow oscillations, providing a means for the hippocampal-neocortical dialogue that supports memory consolidation (Zsófia Clemens et al., 2007; B. Rasch & Born, 2013). Overall, this literature supports the idea that non-REM sleep, and especially slow-wave sleep, is composed of three main neural oscillations which are temporally coupled and essential for successful memory consolidation.

1.2 Targeted Memory Reactivation

Given that memories are reactivated during slow-wave sleep, we may wonder whether we can reactivate specific memories in an attempt to further explore the neural underpinnings of memory consolidation and/or use this information to improve memory. Rasch, Büchel, Gais, and

Born (2007) developed a new technique, now known as targeted memory reactivation (TMR), to explore whether the consolidation of newly acquired information can indeed be prioritized over other learned information. In this experiment, participants studied object locations on a grid and learned their spatial locations while simultaneously being exposed to an olfactory cue. Participants were then allowed to nap in the lab and were exposed to these olfactory cues during slow-wave sleep (SWS) as well as during other stages of sleep and during wake. They found that memory was better, meaning that less spatial information was forgotten compared to in control groups, if the relevant odor cue was presented during SWS. People in the control groups received the odor cue in rapid eye movement (REM) sleep instead of SWS, or they received the odor in SWS but not during initial learning. These findings highlight the importance of SWS for memory consolidation and demonstrate that spatial memory consolidation can be biased using this novel technique (Björn Rasch et al., 2007).

Rudoy, Voss, Westerberg, and Paller (2009) explored whether TMR would be successful with sound cues, which offer increased specificity relative to olfactory cues, for reactivated episodes as one sound can be associated with one learned item. Participants learned object locations and objects could appear anywhere on the screen. Each of 50 objects was paired with a distinct and ecologically relevant sound cues (e.g., teapot-whistle object-sound pairing). During a nap, half of the cues were presented during SWS. It was determined that participants were overall more accurate at placing objects that were cued during sleep compared to the objects which were not cued. This research further implicates SWS in memory consolidation and demonstrates that memory reactivation can be achieved by playing sound cues that are associated with a prior learning episode (Rudoy et al., 2009). The performance benefit as a consequence of targeted memory reactivation for a variety of tasks, including memory for spatial memories,

associative learning, and vocabulary learning, has been substantiated by the results of a recent meta-analysis (Hu, Cheng, Chiu, & Paller, 2020). This literature further implicates sleep for memory consolidation and introduces a technique that allows us to bias memory reactivation to specific items in a non-invasive manner.

1.3 Memory Reactivation for Declarative and Nondeclarative Tasks

It is widely accepted in memory research that long-term memory depends on two memory systems (Squire & Zola, 1996). The type of memory required for the recall and recognition for facts and events is known as declarative memory (or explicit memory); other types of memory are known as nondeclarative memory (or implicit memory). These memory systems are considered to function simultaneously and distinctly, as declarative memory is dependent on the hippocampus whereas nondeclarative memory is not (Squire, 2004; Squire & Zola, 1996). As discussed above, memory reactivation is thought to facilitate hippocampal-neocortical communication through SWR-spindle-SO events (Zsófia Clemens et al., 2007; B. Rasch & Born, 2013; Staresina et al., 2015). The hippocampus, however, is considered essential for declarative, but not nondeclarative memory (Squire & Zola, 1996). The differential way in which TMR may support memory consolidation for declarative and non-declarative tasks remains under investigation.

As stated above, targeted memory reactivation has been explored in a variety of tasks. Many of the tasks that have been studied in TMR paradigms are considered to be hippocampally dependent. One example is spatial memory for object-locations in which participants learn the placement of objects on a grid (Creery, Oudiette, Antony, & Paller, 2015; Eitan et al., 2021; Björn Rasch et al., 2007; Rudoy et al., 2009). Another common declarative memory task used to explore the effects of TMR include learning word-sound associations of foreign and non-foreign

words (Farthouat, Gilson, & Peigneux, 2017; Schreiner & Rasch, 2015, 2017; Tamminen, Ralph, & Lewis, 2017). Combined, this literature underscores TMR as a successful technique to enhance memory in a number of hippocampally dependent, declarative tasks (Hu et al., 2020).

It is also important to explore the impact of TMR for nondeclarative memory. This is of interest not only because nondeclarative memory may function independently of the hippocampus, but because memories may change as they are stored to allow information to be used in new contexts. One promising line of work is the exploration of sleep-based consolidation for generalization learning (Batterink & Paller, 2017; Witkowski et al., 2021; Witkowski, Schechtman, & Paller, 2020). Generalization for our purposes is be defined as the ability to employ learned information in the face of new stimuli or contexts for resulting in performance that is better than expected without this prior learning experience. Witkowski et al. (2021) explored the effect of TMR on a generalization task. Here, participants learned painting styles of six artists. Paintings by each artist were paired with a sound. After viewing many paintings, participants were given 90-minutes to nap. When slow-wave sleep was detected, three of the sounds presented during learning were played to cue memory for the associated artists. To test generalization, participants were shown novel paintings by the six artists after sleep and asked to determine which artist had created the painting. It was determined that neither sleep nor TMR led to improved generalization (Witkowski et al., 2021). A meta-analysis, however, found that sleep studies testing generalization can have differing results based on the test. In particular, extracting hidden regularities tended to improve with sleep especially when there was an important temporal component (Lerner & Gluck, 2019). Following this, Witkowski et al. (2021) hypothesized that TMR may not have been found successful for this type of generalization as the task did not depend on successful rule abstraction. One example which follows from this

hypothesis is that of learning grammar of an artificial language (Batterink & Paller, 2017). Reactivation of phrases was found to improve generalization as a result of successful abstraction of grammatical rules (Batterink & Paller, 2017). TMR has also been found successful in a variety of other implicit learning tasks including but not limited to grammar learning, procedural memory, and complex motor movements (Cheng, Che, Tomic, Slutzky, & Paller, 2021; Johnson, Scharf, & Westlake, 2018; Schönauer, Geisler, & Gais, 2014).

1.4 Influence of Sleep and Sleep Manipulation on the Perception of Auditory Speech

TMR has been studied in several experiments focused on hippocampal-dependent, declarative tasks. One avenue that has yet to be studied extensively is how sleep and sleep manipulation may affect the auditory perception of speech. A first step in this line of research is to understand how this information is learned and stored for later use. Xie and Myers (2017) explored how one might generalize information learned using a second-language (L2) English transcription paradigm. They found that participants have difficulty categorizing L2 talkers, even of the same language background. The authors argued this demonstrates that participants relied on bottom-up input, rather than top down categorization, to better understand novel talkers (Xie & Myers, 2017). Based on these findings, it is likely that implicit memory systems are crucial for this type of generalization process. Though implicit memory systems are critical for learning the nuances of speech that are learned over time through exposure, the nature of sentence transcription also involves top-down processing. As each word in a sentence is revealed, the number of possibilities decreases. This knowledge interacts with the bottom-up signal, affecting transcription.

Generalization of auditory, linguistic information is an interesting line of research as it is incredibly common that people learn to speak new languages in addition to their native language

(L1). It is also the case that if not learned before a critical period, among other factors, one's first language (Patkowski, 1990) can affect the production of their second language resulting in what many refer to as "accented speech." Large variations in the specifics of speech exist even for talkers operating in their first language. For example, people speak at different pitches and speeds. Further, idiosyncratic speech patterns or regional dialects require that a listener have some flexibility in identifying speech sounds. Despite these variations in spoken language, comprehending L1 and L2 speech is often successful, even in difficult listening environments. We also infer that this ability relies on memory as there is evidence that listeners can improve at this task over time (Baese-Berk, Bradlow, & Wright, 2013; Bradlow & Bent, 2008; Cooper & Bradlow, 2016; Xie, Earle, & Myers, 2018; Xie & Myers, 2017).

As stated above, previous research has demonstrated the ability for humans to adapt to L2 speech perception over time and that this knowledge is transferrable in different contexts. For example, Bradlow and Bent (2008) found that if asked to transcribe L2 speech from an L1 Mandarin talker, participants improved, eventually achieving talker-dependent adaptation. Furthermore, when presented with many L1 Mandarin talkers, talker independent adaptation was achieved, meaning that listeners are better able to decipher speech from a new, L1 Mandarin talker (Bradlow & Bent, 2008). These studies demonstrate the ability for this process to be fixed or flexible depending on training which in turn affects following behavior. In a follow-up study, Baese-Berk et al. (2013) asked participants to transcribe L2 speech from five talkers, each with a different L1. Interestingly, training on these talkers allowed participants to recognize words not only from trained talkers, but from a novel talker of a novel language background. These findings suggest exposure to a variety of talkers allows for talker-independent adaptation. This generalization may rely on the extraction of an underlying structure present in L2 speech which

dictates the way L1 talkers produce words in English (Baese-Berk et al., 2013). Because adaptation to L2 speech can improve over time, we infer that there must be a specific learning process whereby humans adjust perceptual representations used for word recognition. Further, research demonstrating that exposure to talkers during training tasks can produce benefits when listeners are faced with novel talkers and language backgrounds further supports the hypothesis that implicit memory plays an important role in this generalization process due to explicit knowledge being extremely limited as participants were faced with new talkers of different language backgrounds.

The ability to generalize learned acoustic information is thought to occur as existing linguistic representations are updated based on new information (Bent & Baese-Berk, 2021). The literature on perceptual categorization, especially speech-categorization, provides useful insights into the mechanisms underlying the generalization of speech (Bradlow, Pisoni, Akahane-Yamada, & Tohkura, 1997; Eisner & McQueen, 2006; Iverson, Hazan, & Bannister, 2005; T. Kraljic & Samuel, 2006; Tanya Kraljic & Samuel, 2007; Norris, McQueen, & Cutler, 2003). In tasks examining speech categorization, participants are typically asked to identify to which of two categories a stimulus belongs. For example, Norris et al. (2003) exposed participants to an ambiguous [f][s] fricative [?]. One group was trained to label this fricative as [f] while the other learned to labeled it as [s]. Later, when presented with ambiguous words, participants indeed labeled the sound in accordance with prior training. This demonstrates that this type of categorization can lead to the same ambiguous sound being perceived as different phonemes if it is mapped onto preexisting knowledge. Wright, Baese-Berk, Marrone, and Bradlow (2015) also explored this type of speech categorization, exposing participants to a non-native category boundary (“mba”-“ba”). Interestingly, passive exposure of auditory stimuli after training

produced the greatest improvement in the ability to distinguish the non-native sound from the native sound. Improvement from exposure was also found to improve the ability to correctly identify L2 Mandarin speech (Wright et al., 2015).

A general principle in this sort of learning is that speech perception depends on knowing the category boundaries that define syllables and words. Because speech sounds vary on so many dimensions, especially across diverse talkers, the listener's challenge is to analyze units of speech correctly based on category-relevant dimensions in the face of ample variability across all dimensions. Another aspect regarding second-language English is that of listener bias. Boduch-Grabka and Lev-Ari (2021) found that exposure to second-language speech improved the ease of processing which in turn reduced bias. Overall, this research allows us to better understand perceptual adaptation which extends beyond basic science, providing information regarding the way humans interact.

In summary, successful word recognition depends on the listener's ability to differentiate between sounds that are critical and sounds that may reflect idiosyncratic variability in speech that is not necessary to the formation of a word. Together, this research demonstrates that learning non-native speech sounds is facilitated both by exposure to the sounds and by practice. Understanding this process is important because improved speech recognition relates to improved speech comprehension as a sentence can be better understood once all words of a sentence are recognized. As stated above, this type of learning likely depends on both bottom-up and top-down processing. We hypothesize that the bottom-up aspect of adaptation to L2 speech is supported by implicit memory systems. Listeners cannot explicitly state all of the information that allows them to tune into the important aspects of speech, though they learn to do so over time. Top-down expectations also play a role as a listener knows they are hearing a second-

language speaker and may increase their attention and put more effort into processing speech. Further, a listener can begin to expect words as a sentence progresses meaning that all words are not equally likely depending on what was previously heard.

Because it is thought that L2 English may have some systematic similarities, regardless of talker or language background (Baese-Berk et al., 2013), we hypothesized that most of the benefit that comes from exposure to L2 speech is that of implicit processing of speech sounds and the tuning of one's perceptual space. It is also thought that TMR may aid in memory for implicit learning that relies on successful rule abstraction (Witkowski et al., 2020). By using TMR to probe the question of whether memory reactivation is helpful for this type of learning, we also can better understand the memory systems that support perceptual adaptation as well as what type of information is learned.

CHAPTER 2

Experiment 1: Generalization of L2 Speech Following Single- or Multiple-Talker Training

2.1 INTRODUCTION

In this experiment we explored a specific type of generalization learning: adaptation to second-language speech. Bradlow and Bent (2008) explored the ability for first-language talkers of English to improve at word recognition of L2 spoken English. It was first determined that participants were able to achieve talker-dependent adaptation after exposure to a Mandarin-accented talker. In a second experiment, participants were exposed to many different Mandarin-accented talkers. After exposure, participants were asked to transcribe sentences from one novel Mandarin-accented talker. Results showed that this type of high-variability training allowed listeners to achieve talker-independent adaptation for a talker of the same language background (Bradlow & Bent, 2008). Building on this research, Baese-Berk et al. (2013) explored high-variability training in the context of many language backgrounds. To do so, participants were exposed to many talkers, each of a different language background. After training, participants were asked to transcribe sentences from a single test talker of a novel language background. It was determined that perceptual adaptation extended beyond one language background as participants were able to generalize what was learned during training and perform well when faced with the novel talker (Baese-Berk et al., 2013). The authors proposed that together, this research demonstrated that L2 English talkers produce variability in speech that is in some ways systematic. Efficiently processing speech in the face of this variability in speech can be learned, which can thereby improve recognition in novel situations.

On the other hand, an alternative view is that multiple talkers are not necessarily helpful in perceptual adaptation to L2 speech (Xie & Myers, 2017). In an experiment designed to test

this idea, L1 English talkers were exposed to Mandarin-accented words. Importantly, training either consisted of exposure to a single talker or to multiple talkers. After exposure, participants were tested on their ability to transcribe novel words spoken by a novel Mandarin-accented talker. Interestingly, acoustic similarity between the exposure and test talkers, and not the number of exposure talkers, predicted greater generalization. These findings demonstrate that the systematic variation described by Baese-Berk et al. (2013) may not always be apparent. High-variability training does, however, provide a large number of acoustic sounds for participants to sample from increasing the probability that phonetic overlap between exposure and test talkers will exist resulting in successful generalization (Xie & Myers, 2017).

Experiment 1 combined what is known about generalization in language to understand how training on either single or multiple talkers affects performance before and after a wake delay. To do so, participants completed two sessions: one in the morning and one in the evening. Informed by the results collected in Experiment 1, in Experiment 2 we explored perceptual adaptation and generalization after learning. Experiment 2 also explored the possible effect of sleep and sleep reactivation in this learning and generalization process.

2.2 METHODS

2.2.1 Participants

Participants were recruited through the online platform Prolific (except for 4 recruited on campus). A total of 365 potential participants attempted to complete the first session. Participants were excluded if they quit the training session early or did not complete the second session. Of these potential participants, 195 people completed both training and test sessions within the desired time frame and are included in the final analyses. Of these, there were 130 women, 64

men, and 1 participant who did not report gender. All participants registered with Prolific as first-language English talkers between the ages of 18-35 years [mean age = 26.87; standard deviation = 5.12]. Participants resided in the United States and reported no literacy challenges or hearing disabilities at the time of the experiment. Participants who completed the training session, which consisted of a 60-sentence auditory transcription task, were individually informed that the second session of the experiment would be available for completion in the evening. Participants were promptly paid for their time after each session.

2.2.2 Stimuli

Sentences were obtained from the ALLSSTAR corpus (Bradlow, n.d.), which contains recordings from many talkers of many languages. For the purposes of this study, sentences from four L2 English talkers were selected. Talkers were selected by the discretion of the research team in an attempt to include a range of language backgrounds and difficulty. All sentences were mono-clausal, with canonical declarative syntax (e.g., “They are running past the house”). The sentences were presented in speech-shaped noise with a 0 dB signal-to-noise ratio (SNR), which corresponds to listening in a moderately noisy environment.

Sentences were spoken by Turkish, Spanish, Farsi, and Brazilian Portuguese native talkers. One-hundred and twenty sentences were divided into two sets of 60 to be transcribed. One set was used in a training phase and one in a test phase. The talkers for the training phase varied across the experimental conditions described below, but the same list of 60 sentences was used for all participants. The talkers for the test phase were the four listed above (15 sentences each), and for all participants the same 60 sentences were used in a random order. The two sets were arranged so as to minimize highly similar words from appearing at both the training and

test. In other words, we attempted to minimize the number of words that were heard during the training session that would be presented again at test.

2.2.3 Behavioral Task

Participants used their own computers to access a web page that hosted the entirety of each session. Data were stored on Google Firebase. After consenting to participate in the experiment, participants answered demographic questions. Participants were then directed to begin the experiment. Participants were instructed to wear headphones throughout the experiment and set the volume to a comfortable setting. First, two practice words, laugh and gas, were presented and participants were asked to transcribe them to ensure the audio was working properly before starting the experiment. The training portion of the experiment began following the audio test. Here, 60 sentences were presented one by one auditorily. Participants used a mouse to press a button labeled “Play” in order for each sentence to be presented. The participant was then prompted to type what they heard into a text box before advancing to the next sentence. Each sentence could only be heard once. The procedure was identical for the training and test session; each session lasted approximately 30 minutes.

2.2.4 Experimental Design

Training

Participants were assigned to either high- or low-variability training conditions, which can also be considered multiple-talker (n = 84) and single-talker (n = 84) training, respectively. We also ran 27 participants who only completed the test phase to serve as a control group. Because we were interested in how performance changed from training to test after a wake delay, we arranged for training to occur in the morning. Because participants could be in any USA time

zone, the training session was made available for a long period (9 hours, from 4AM to 1PM CST) to cover morning periods across all USA time zones.

The multiple-talker training group was trained on 20 English sentences spoken by three of four possible talkers with different language backgrounds. Each participant in the single-talker group trained on only one of the four mentioned talkers (Figure 2.1). No feedback was provided.

The language backgrounds for L2 talkers were Brazilian Portuguese (PBR), Turkish (TUR), Farsi (FAR), and Spanish (SPA). We will refer to the single-talker groups as PBR, TUR, FAR, and SPA to denote which talker was heard. For the multiple-talker condition, we will refer to these multiple talker groups as noPBR, noTUR, noFAR, and noSPA to denote which talker was excluded at training. For example, the PBR talker was excluded from the multiple-talker group noPBR while the TUR, FAR, and SPA talkers were included. Participants were randomly assigned to one of eight training groups. All participants were exposed to the same sentences but with different talkers in each group. Sentence order was randomized for each participant.

Test

The test session was open for participants to complete for 9 hours in the evening. (4PM-1AM CST). The test took place approximately 11 hours after training [M= 10H 53M; SE=8M]. All participants completed a sound test, described above, and were then tested on 60 novel sentences spoken by each of the four talkers outlined above. It is important to note that the single-talker groups were exposed to three novel talkers at test while the multiple-talker groups were exposed to one novel talker at test.

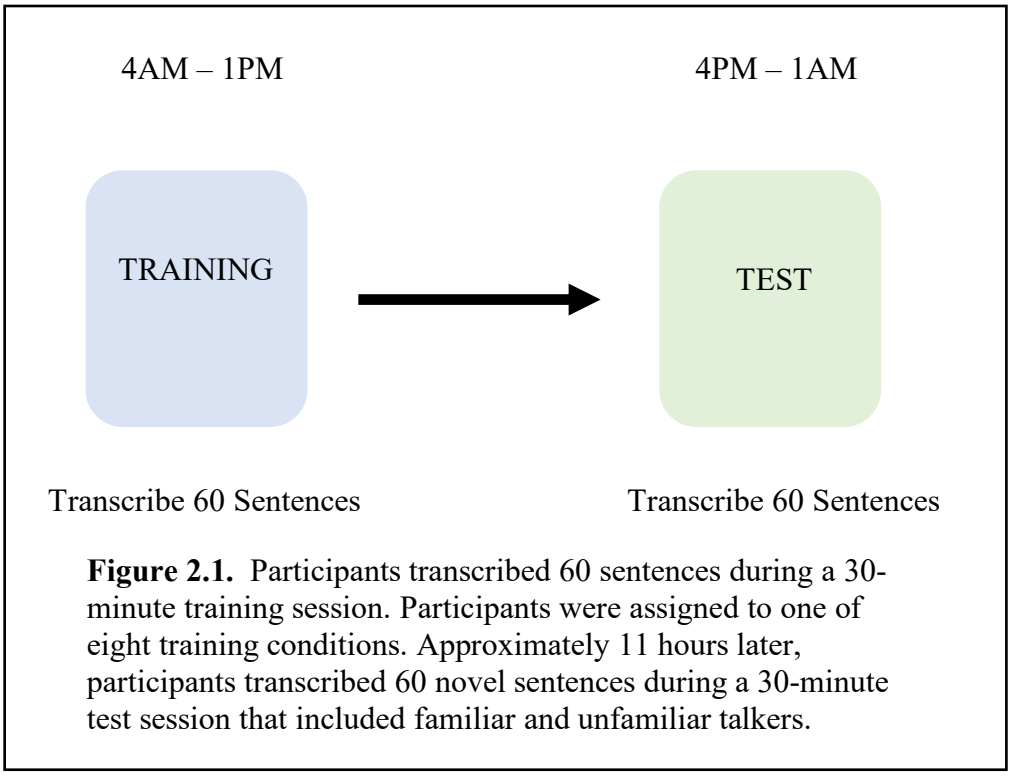

Figure 2.1. Participants transcribed 60 sentences during a 30-minute training session. Participants were assigned to one of eight training conditions. Approximately 11 hours later, participants transcribed 60 novel sentences during a 30-minute test session that included familiar and unfamiliar talkers.

2.2.5 Behavioral Analysis

Scoring

Participant responses were scored using an online, automatic scoring tool (Borrie, Barrett, & Yoho, 2019). The scoring tool calculated total number of words correctly transcribed on each trial for each participant. We allowed for all accepted exceptions in the automatic scoring tool. Response words were scored as correct if the entered word was: (1) a homophone or common misspelling, (2) included as a rootword, (3) omitting a double letter, (4) included "the" in place of "a" and vice versa, (5) was in the incorrect tense, or (6) was entered in either its plural or singular form. These exceptions were made so that scores reflected differences in word recognition rather than spelling or grammar.

Behavioral Analysis

Each participant contributed two behavioral scores. The training score was computed for the second half of training (30 sentences), as the mean number of words correct divided by the total number of words possible. The test score was computed for all test blocks (60 sentences) as the mean number of words correct divided by the total number of words possible. Behavioral analysis was conducted in R version 4.2.3.

2.3 RESULTS

2.3.1 Training

Table 2.1 displays overall performance for all participants during training and test. We first examined transcription accuracy during the training phase to verify evidence of learning. The training results were split into 6 blocks with 10 sentences each. We first plotted performance on each of the blocks separately for the single-talker and multiple-talker training groups (Figure 2.2 AB, respectively). We found that participants tended to improve over time with performance reaching a plateau toward the end of training.

To test for evidence of learning, we compared performance on the first three blocks versus the last three blocks of training. Both groups performed significantly better during the second half compared to the first half [single-talker, $t(83) = 6.60$, $p < .05$; multiple-talker, $t(83) = 4.98$, $p < .05$]. The average improvement across all eight groups was from 68% of words correct in the first half to 72% of words correct in the second half.

Next, we visualized performance for each group separately. For single-talker training groups, as shown in Figure 2.2 C, the number of words correctly identified was largest for participants listening to the PBR talker [M=86%; SD=22%] and smallest for those listening to the TUR talker [M=45%; SD=35%]. That is, the TUR talker was the least-intelligible talker. We

also found that performance was superior for Multiple-Talker training group noTUR [M=76; SD=26], which excluded the least intelligible talker (Figure 2D).

To further explore these differences in talkers, an ANOVA for both the single-talker and multiple-talker training groups was run with each of the four talkers predicting performance on the second half of training. We found a significant effect of talker for the single- and multiple-talker groups [$F(3,80)=37.43$, $p < .05$, $F(3,248)=49.11$, $p<.05$], respectively. After a Bonferroni correction, we found significant differences between the TUR talker and all other talkers for both the single- and multiple-talker groups [p's$<.05$] with performance being the lowest for the TUR talker. Interestingly, for the multiple-talker group we also found a significant difference between the FAR and PBR talkers [$p<.05$] with participants performing worse on the FAR talker.

Finally, we tested whether the same level of accuracy was achieved in single- and multiple-talker groups during the last three blocks of training. We found a marginal difference between these two groups [75% of words correct for single-talker training; 70% words correct for multiple-talker training; $t(146.64) = 1.95$, $p = .053$]. Because performance appears to plateau at block 3 for all groups and was relatively matched, the average performance of blocks 4-6 was considered the appropriate representation of competency achieved after training and served as the comparison value for test performance values obtained after the 11-hour delay.

Table 2.1. Performance at training and/or test for all groups

	Training Performance				Test Performance	
	Blocks 1-6		Blocks 4-6		All Sentences	
Single-Talker						
Training Group	Average Correct (%)	SD (%)	Correct (%)	SD (%)	Correct (%)	SD (%)
FAR	77	28	79	26	76	31
PBR	86	22	89	19	72	32
SPA	80	26	82	25	72	32
TUR	45	35	48	35	67	34
Multiple-Talker						
noFAR	66	34	66	34	72	33
noPBR	62	34	66	34	74	32
noSPA	66	35	67	35	71	34
noTUR	76	28	80	26	77	30
Control Group						
Control	NA	NA	NA	NA	69	34

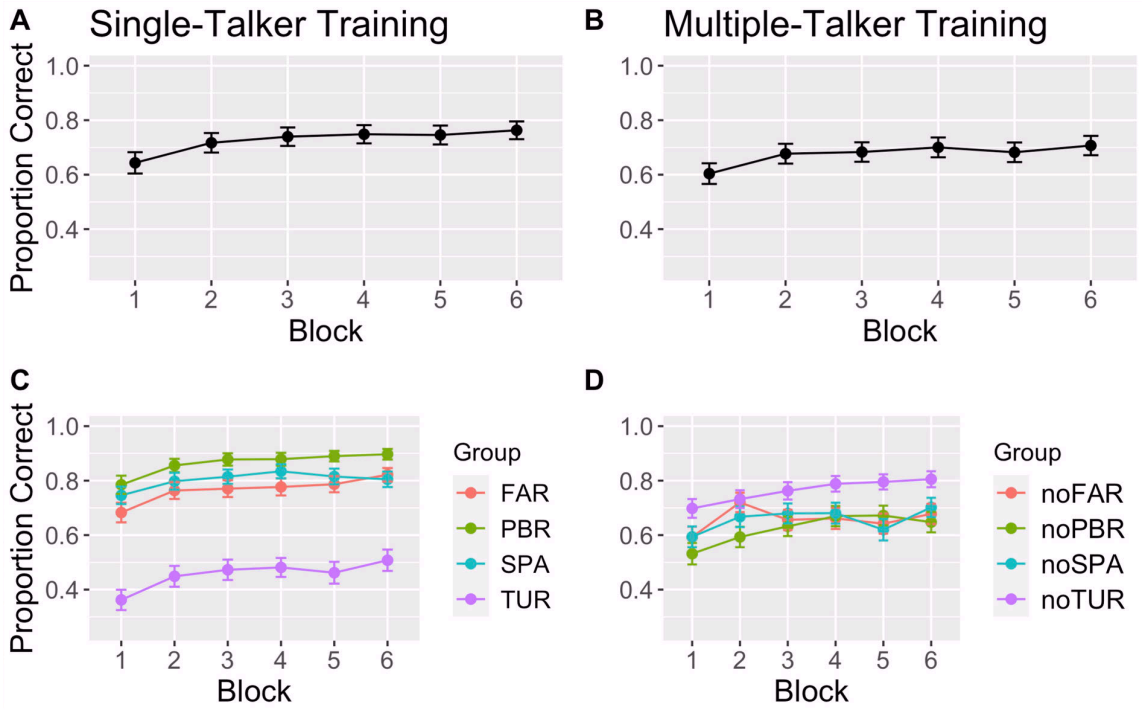

Figure 2.2. Training-phase performance collapsed across single-talker training groups (A) and multiple-talker training groups (B). Performance for the four single-talker training groups (C) and the four multiple-talker training groups (D). Error bars show standard errors of the mean.

2.3.2 Training vs Test vs Control

We next compared performance between training and test based on training group, with results collapsed across the four different talkers in each test (Figure 2.3; Table 2.2). There was a nonsignificant trend for poorer test performance after single-talker training [$t(83) = 1.70$, $p = .09$], with a significant improvement at test after multiple-talker training [$t(83) = 3.37$, $p = <.05$]. For the multiple-talker and single-talker groups, there was a 4% improvement and a 3% decline in performance, respectively. This difference as a function of type of training was substantiated

by a significant group x test interaction [$F(1, 166) = 9.64$, $p = .002$] in a 2x2 ANOVA (Group: Multiple-talker, Single-Talker; Test: Training, Test).

One group of participants completed the Test phase without training and were considered the control group. At test, the single-talker, multiple-talker, and control group correctly identified 72%, 74%, and 69% of words, respectively. We found no difference in scores for these three conditions [$F(2,192) = 1.99$, $p = .14$]; one-way ANOVA with three levels (Single-talker, Multiple-talker, Control). We also ran a t-test between those who received training and those who did not and found no difference [$t(36.05)$, $p = .10$].

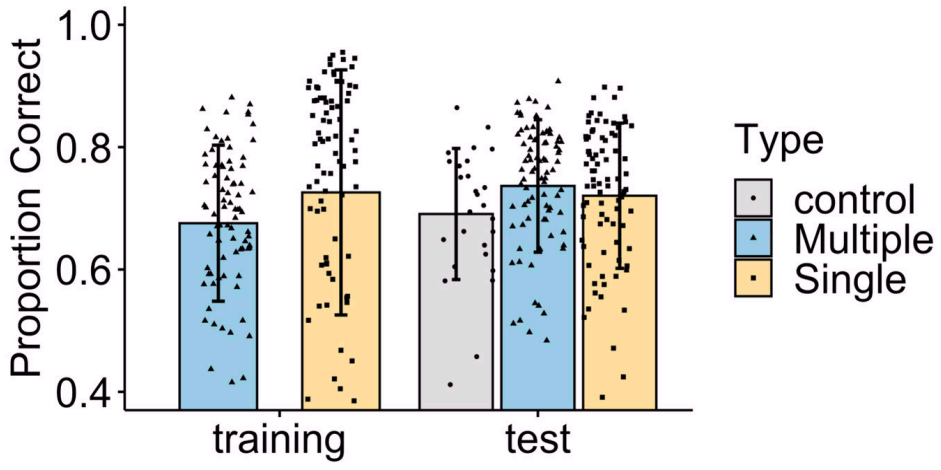

Figure 2.3. Average proportion of words correctly identified at Training (final three blocks) and Test for both types of training group. Participants in the control group (grey) completed the test without a prior training session.

2.3.3 Generalization

Performance for trained and novel talkers

Even though all 60 sentences in the test session were novel, the four talkers were not equally familiar to participants; the design systematically manipulated whether sentences were read by the same person in the two sessions. One test-phase talker was familiar after single-talker training, whereas three test-phase talkers were familiar after multiple-talker training. Although training may lead to improvements for unfamiliar talkers (i.e., generalization), we expected transcription accuracy to be higher for familiar talkers.

Figure 2.4 shows results for these conditions First, test-phase results were subjected to a 2 (Group: Multiple-talker, Single-talker) x 2 (Talker: Familiar, Unfamiliar) mixed-effects ANOVA. The analysis revealed there was no main effect of Group [$F(1, 166) = .51$, $p = .47$], Talker [$F(1, 166) = .23$, $p = .63$], nor a Group vs Talker interaction [$F(1, 166) = .43$, $p = .51$]. It is possible, however, that exposure to a particular talker or set of talkers differentially impacts generalization, given the differences in intelligibility evident in the training phase.

We next explored how each training group performed on familiar and unfamiliar talkers (Figure 2.5). We conducted a 2 (Talker type: Familiar vs Unfamiliar) x 8 (Training Group: FAR, PBR, SPA, TUR, noFAR, noPBR, noSPA, noTUR) mixed-effects ANOVA. Though there was no main effect of Talker [$F(1, 160) = .69$, $p = .41$], there was a main effect of Training Group [$F(7, 160) = 2.58$, $p = .01$], and a Talker x Training Group interaction [$F(7, 160) = 58.08$, $p < .05$]. To further understand these findings, we explored which groups contributed to the main effect of group by running pairwise comparisons of each training group at test. After a Bonferroni correction, taking speaker familiarity into account, we found significant differences between noPBR and TUR, FAR and TUR, and PBR and TUR [p's $< .05$], all performing better

than TUR. Overall, all training groups tend to perform better than those who only trained on the TUR talker while noPBR, FAR, and PBR resulted in the most reliable differences (Table 2.3) This further supports the idea that this particular L1 Turkish talker is less intelligible and that training on this talker hindered generalization to novel talkers. Further, we found that those who trained on the PBR talker performed significantly better on this talker at test [M=84%; SD=25%) compared to the three novel talkers [M=68%; SD=33%) (FAR, TUR, and SPA). As PBR was identified as the most intelligible talker, it seems that exposure to a very intelligible talker may not provide the best learning conditions for generalization to novel talkers of different language backgrounds.

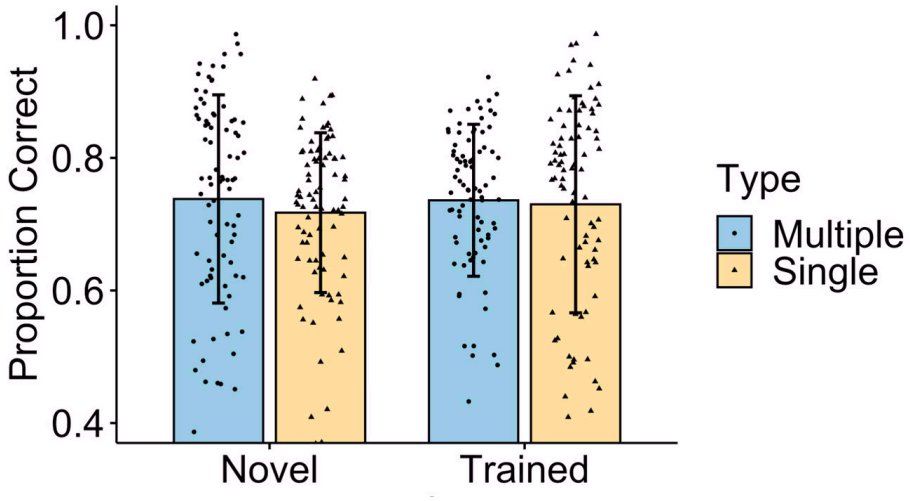

Figure 2.4. Proportion of correctly transcribed words at the test for novel and previously trained talkers. Note that the single-talker group is tested on one trained talker and three novel talkers while the multiple-talker group is tested on three trained talkers and one novel talker.

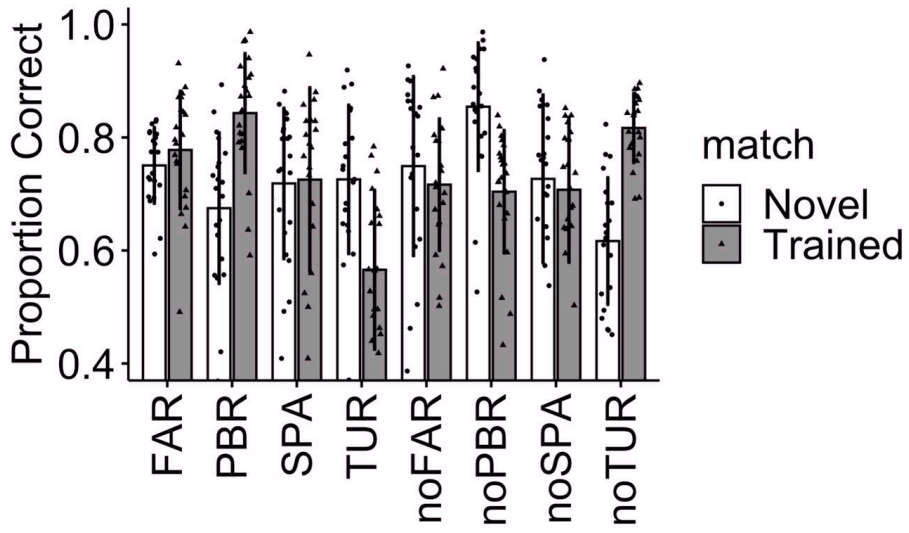

Figure 2.5. Participant performance for all training groups for trained and novel talkers. Participants tended to perform similarly though there was a difference for the single-talker PBR group as they performed significantly worse for the novel speakers on average (FAR, TUR, SPA).

Talker specific performance differences

Here, we took a closer look to explore how training affected performance when faced with each unique talker at test and whether training groups differed in their performance for each talker at test by running a one-way ANOVA's (Figure 2.6). We found that groups were similar in their performance for all talkers [p's > .05] indicating that intelligibility effects were persistent regardless of training group. Next, we ran a 2 (Group: Single-talker, Multiple-talker) x 4 (Talker: SPA, TUR, FAR, PBR) mixed effects ANOVA. We found no effect of Group [$F(1, 166) = .84$, $p = .36$] and a marginal Group x Talker interaction [$F(3, 498) = 2.43$, $p = .06$]. There was, however, a main effect of talker [$F(3, 498) = 377.88$, $p < .05$]. By running pairwise comparisons, we found differences between all talkers except for SPA and FAR following a Bonferroni correction.

Finally, we explored whether training group affected performance on particular talkers at test by running a 4 (Talker: FAR, PBR, SPA, TUR) x 8 (Training Group: FAR, PBR, SPA, TUR, noFAR, noPBR, noSPA, noTUR) mixed effects ANOVA. We found no effect of group [$F(1, 160) = 1.10$, $p = .37$) or Group x Talker interaction [$F(21, 480) = 1.09$, $p = .35$]. There was, however, a main effect of talker [$F(3, 480) = 374.75$, $p < .05$]. Pairwise comparisons again revealed similar results as we found differences between all talkers except for SPA and FAR (Table 2.3).

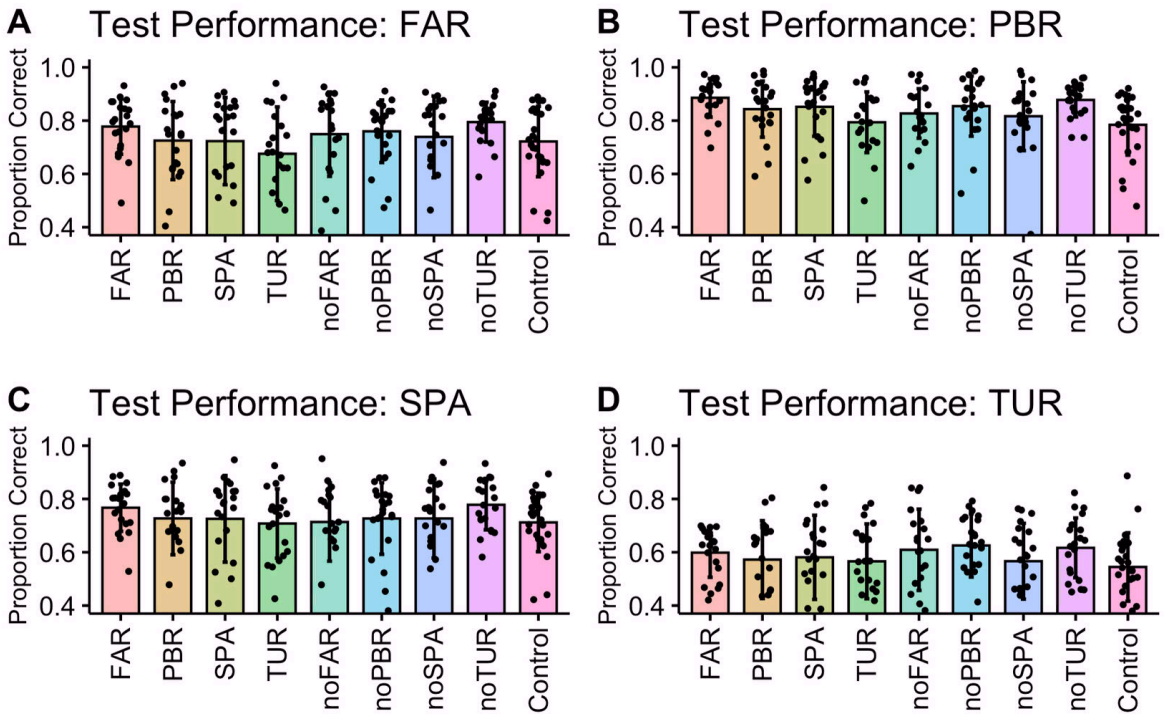

Figure 2.6. Performance for all training groups on each of the four talkers. There were no significant differences in performance based on group for any of the four talkers.

Table 2.3 Significant pairwise comparisons at test		
Training Group 1 (M%; SD%)	**Training Group 2 (M%; SD%)**	***p*-adjusted (Bonferroni)**
FAR (76;31)	TUR (67;34)	.003
PBR (72;32)	TUR (67;34)	.006
noPBR (74;32)	TUR (67;34)	.0003

2.4 DISCUSSION

In this experiment, we tested how exposure to many talkers (high-variability training) or one talker (low-variability training) impacts word recognition when faced with both trained and novel talkers after a delay. Over time, both conditions showed evidence of learning performing significantly better during the second half than the first half. Increased task familiarity likely contributed to this improvement as participants adjusted to recognizing speech in noise from an L2 speaker. Once participants adjusted to the task, it likely that they learned to adapt to the talkers presented allowing them to recognize more words over time (Baese-Berk et al., 2013; Bradlow & Bent, 2008). At training, participants were able to correctly identify more words spoken by the PBR talker and fewer words spoken by the TUR talker indicating the most and least intelligible talkers, respectively. It is important, however, to note that individual talkers speak with idiosyncrasies that affect intelligibility regardless of their L1.

Though performance did not differ overall between training and test for either group, there was evidence that exposure to a specific talker, or set of talkers, affected test performance. As transcription scores were worse overall for the TUR talker, training on only this talker hindered performance at test while PBR proved to be the most intelligible talker. Further, those who trained on all talkers except TUR performed better than all other multiple-talker groups at

test. We did find an interaction between training type and test such that those in the multiple talker group improved from training to test while those in the single talker group showed a decline in performance. When looking at performance, it does not appear that the PBR and TUR talkers are responsible for driving this effect (Figure S1). Overall, performance was affected by the variability of exposure at training as well the intelligibility of the speakers. Xie and Myers (2017) determined that mere exposure to a number of talkers does not promote generalization to new talkers but that the similarity between talkers does. Our findings support this idea as different combinations of training group and talker affected test performance for novel and familiar talkers.

On the other hand, based on Figure 2.6, it seems to be the case that participants perform similarly on each talker regardless of training group. It might be the case that all training groups provided participants with generalized knowledge that was used when transcribing sentences at test. This may explain why performance tended to be similar for all talkers at test. To explore this hypothesis, further testing on the linguistic properties of all speakers should be performed. In addition, follow-up studies on the way exposure to these different talkers affects later generalization should be explored.

One purpose of this experiment was to inform the design of a follow-up nap study which will be discussed in Chapter 3. Because the effects of sleep can be subtle, choosing a training paradigm in which participants have room for improvement on a sensitive measure allows for experimenters to pick up these effects. Here, we found that the TUR talker was the least intelligible indicating room for possible improvement. Based on these findings, we chose to pursue this single-talker training group for the follow-up nap study. Further, this group

demonstrated the most room for improvement from training to test making it possible for the benefits of memory consolidation during sleep to have a greater impact on performance.

Chapter 3

Experiment 2

3.1 INTRODUCTION

People are able to adapt to a variety of second-language (L2) speakers of similar and even novel language backgrounds (Baese-Berk et al., 2013; Bradlow & Bent, 2008). This ability is thought to develop through repeated exposure, allowing us to update our existing representation of meaningful speech sounds (Bent & Baese-Berk, 2021). As this process takes time, it likely depends on memory consolidation, including that which occurs during sleep, such that novel acoustic information becomes integrated into existing knowledge.

The role of sleep in the adaptation to speech sounds has only recently been explored (Xie et al., 2018). Participants were exposed to two Mandarin-accented talkers who were not phonetically similar. This lack of phonetic similarity was shown, as exposure to one of the talkers showed little transfer to the untrained talker. In a second experiment, participants were trained on one of these talkers in either the morning or the evening. After a 12-hour delay, participants were tested on their ability to transcribe sentences spoken by both talkers. Interestingly, only the group that had intervening sleep showed increased ability to generalize to the untrained talker (Xie et al., 2018). The authors proposed that processes occurring during sleep facilitated generalization to the novel talker by allowing for listeners to abstract relevant, acoustic information learned during exposure.

In Chapter 2, we explored how participants performed when faced with the task of transcribing L2 speakers and how this training affected their ability to generalize to novel talkers of different language backgrounds as well as novel sentences from the trained talker. We found that the TUR talker was the most difficult to transcribe but that participants also performed rather

well when faced with new talkers after an 11 hour wake delay (see Sections 2.3.1 and 2.3.2) In this experiment, we manipulated sleep so that we could better understand how this generalization process might be supported by memory consolidation. Because the effects of sleep can be subtle, choosing a training paradigm in which participants have room for improvement on a sensitive measure allows for experimenters to pick up these effects. As stated above, the TUR talker was the least intelligible. Participants could transcribe words from the TUR talker, but there was also room for improvement. Based on these findings, we chose to pursue this single-talker training group for the follow-up nap study.

3.2 METHODS

3.2.1 Participants

We recruited participants via email. We included people who were interested in completing a nap study in our laboratory on Northwestern University's campus. All participants were also first-language English speakers between the ages of 18-35 years [mean age (M) = 21.24; standard deviation (SD) =2.96]. All participants resided near the Evanston or Chicago Illinois area and reported no language or hearing disabilities at the time of the experiment. Fifty-five participants completed the experiment in its entirety. Of these, thirty-eight participants received a sufficient number of auditory cues during sleep for their data to be considered in the final analyses. Eighteen participants were included in the experimental group while twenty participants were included in the control group. Participants were paid for their time spent inside and outside of the lab.

3.2.2 Stimuli

Sentences were obtained from the ALLSSTAR corpus (Bradlow, n.d.), which contains recordings from many talkers of many languages. For the purposes of this study, sentences from four L2 English talkers were selected. All sentences were mono-clausal, with canonical declarative syntax (e.g., “They are running past the house”). The sentences were presented in speech-shaped noise with a 0 dB signal-to-noise ratio (SNR), which corresponds to listening in a moderately noisy environment.

Sentences were spoken by Turkish, Spanish, Farsi, and Brazilian Portuguese native talkers. One-hundred and twenty sentences were divided into two sets of 60 to be transcribed. One set was used in a training phase and one in a test phase. The talkers for the training phase varied across the experimental conditions described below, but the same list of 60 sentences was used for all participants. The talkers for the test phase were the four listed above (15 sentences each), and for all participants the same 60 sentences were used in a random order. The two sets were arranged so as to minimize highly similar words from appearing at both the training and test. In other words, we attempted to minimize the number of words that were heard during training session to be presented again at test.

3.2.3 Behavioral Tasks

We were interested in the way sleep and memory consolidation might affect performance for previously trained talkers as well as the ability to generalize to novel talkers. We employed a technique known as targeted memory reactivation (TMR) during which auditory cues are paired with a learning episode and later presented unobtrusively during sleep. This procedure allowed for us to track when memory reactivation might occur and explore brain physiology during this

period. In this experiment, participants completed an L2-English speech transcription task with sentences presented over speech-shaped noise and these same sentences (without noise) were presented during an afternoon nap.

Often in TMR studies, auditory cues are linked with a distinct learning episode and the experiment can be completed using a within-subjects study design. We hypothesized, however, that accented speech cues may reactivate an entire learning episode rather than the memory for one specific sentence. If this were the case, it would be difficult to determine whether memory reactivation impacted participant performance. To avoid this possibility, participants also completed an implicit motor sequence control task in a between-subjects study design in which TMR cues were either from the L2-English transcription task or the implicit sequence control task. We reasoned that this design would control for generic effects of auditory cues during sleep while avoiding issues regarding cue specificity.

Experimental task

Training: All participants were asked to transcribe 60 sentences spoken by a native Turkish talker. The first 20 sentences were presented one by one through headphones, in the absence of noise, to allow participants to get used to the task and hear the talker clearly. The remaining 40 sentences were presented in noise with a 0dB SNR. For all sentences, participants were asked to type what they heard using a keyboard. After, participants heard the sentence once more with orthographic feedback presented on the screen in the absence of noise. Participants were allowed to move at their own pace and could take a break before listening to the next sentence.

Test: At test, participants transcribed 60 novel sentences presented in noise with an SNR of 0dB. The 60 sentences were split between four talkers such that each talker spoke 15 total sentences. Sentences were spoken by the familiar, native Turkish talker as well as three novel talkers from

three novel language backgrounds: Farsi, Spanish, and Portuguese. No feedback was given at test. As before, participants were allowed to move at their own pace and could take a break before listening to the next sentence presented through headphones. Participants could only listen to one sentence at a time and no sentence could be played more than once.

The in-lab test was completed on a local desktop and headphones were worn throughout. The experiment was coded in Neurobehavioral Systems Presentation version 23.1. The remote follow-up test was administered either through Qualtrics or by accessing a webpage written in Java with data hosted by Google Firebase.

Control task

Participants completed an implicit learning task known as the Serial Interception Sequence Learning task. Here, participants completed a series of button presses which corresponded to a tone creating a melody. There was a repeating sequence embedded within the task which participants have been shown to learn implicitly through more accurate and quicker motor responses for learned compared to random sequences (Antony, Gobel, O'hare, Reber, & Paller, 2012; Sanchez, Gobel, & Reber, 2010).The implicit sequence to be learned was presented 80% of the time. During this task, participants learned to map motor responses to four keyboard keys (D, F, J, or K) for an auditory stimulus. The task required timed keypresses in response to a simultaneous visual and auditory cue. Each 12-item sequence contained 36 tones (3 sequential repetitions of each tone per trial with 12 total trials). Three tones were played rather than one so that the participant could prepare a response. Participants were to press the correct button on the third repetition of the tone. The trial was set at an initial speed of three 100ms tones of the same frequency. Trials were considered correct if the proper key was pressed within 300ms of the third tone. The speed of this task was adaptive in order to keep participants at a performance level of

~80% accuracy. The number of correct responses after six button presses was calculated. If all six responses were correct, speed would decrease (become faster) by ~5%. If fewer than six responses were correct, the speed would increase by 5%. Duration could not exceed ~3.5s/trial so that the in-lab session could be completed within four hours. Participants were able to take a break after every 360 trials. Headphones were worn throughout the task. Participants completed a JavaScript version of this task by accessing a webpage. Data were gathered online via HTML5/JavaScript code through web-based browsers.

Training: Before coming to the lab, participants completed an online training session at home. This session consisted of an eight-trial demo with a fixed sequence (D, F, J, K, J, F, K, D). Participants then completed 720 trials in which the 12-item sequence was covertly presented. In the lab, participants completed another 720-trial training. Participants then completed three 360-trial tests which included the trained and novel sound-key mappings. Trained and novel sequences were chosen randomly for each participant. Participants wore headphones during both training and test.

Test: First, participants completed a 120-trial speed adjustment block which included ten novel 12-item foil sequences so that participants could achieve an accuracy of 80% in the untrained sequence. Participants then completed two 360-trial tests in the trained and untrained sound-key mapping. Speed was not adaptive during this portion of the test so that we could accurately assess learning of the trained sequence. Novel foils were used for the untrained sequences. Participants completed the same test with novel foils one week later.

Scoring

Participant responses for the transcription were scored using an online, automatic scoring tool (Borrie et al., 2019). The scoring tool calculated total number of words correctly transcribed on

each trial for each participant. We allowed for all accepted exceptions in the automatic scoring tool. Response words were scored as correct if the entered word was: (1) a homophone or common misspelling, (2) included as a rootword, (3) omitting a double letter, (4) included "the" in place of "a" and vice versa, (5) was in the incorrect tense, or (6) was entered in either it's plural or singular form. These exceptions were made to allow the scores to reflect meaningful differences in language perception.

3.2.4 Afternoon nap

Participants were assigned to be in either the control condition or the experimental condition according to a counterbalancing order set ahead of time. Once participants reached SWS, indicated by at least 6 seconds of slow oscillations in a 30-s period, sound cues were presented. Slow oscillations are defined as having a frequency range of 0.5-4.5 Hz with a peak-to-peak amplitude of at least 75 μV, according to the American Academy of Sleep Medicine (Berry et al., 2017). Cues were stopped promptly upon signs of arousal. Pink noise was played during the entirety of the nap and volume was adjusted based on participant preference such that they could hear the pink noise but also fall asleep. Sound cues were presented at a sound intensity similar to that of the white noise with a sound pressure level of approximately 38 dB. The average SNR of sound cues to pink noise was .13dB.

EEG recording

After training, participants were fitted with an EEG cap. A Biosemi system was used to record electrical activity from 27 tin electrodes at standard scalp locations. In addition, we also recorded from five other electrodes on the head, including two electrooculogram (EOG) channels, one chin electromyogram (EMG) channel, and both the left and right mastoids. EEG was collected

with a sampling rate of 256 Hz. Once EEG recording began, participants were allowed 90 minutes to nap.

Experimental Group

Eighteen participants were exposed to the same 60 spoken sentences, or cues, presented during training. Sentences were ~1.5s long. Sentences were presented during slow-wave sleep as indicated by polysomnography (PSG). An inclusion criterion was that a minimum of one round of cueing (60 total sentences) must have been presented in the nap. Participants received no more than three rounds of cueing. On average, participants were exposed to 124.05 cues. Nine participants were excluded because all 60 spoken sentences were unable to be presented within the 90-minute timeframe. These participants were exposed to an average of approximately 7 words. A new sentence was presented every 5 seconds and sentences were played in the same order for each participant. There was an average of 5.31 words per sentence.

Control Group

Twenty participants were exposed during sleep to the same 12-tone sequence repeated throughout training. Following another similar study (Antony et al., 2012), the inclusion criterion was that at least 20 rounds of cues were presented during slow-wave sleep for a total of 240 cues. Sounds were presented once SWS began and stopped when signs of arousal became evident. Cues were presented with an ISI of 1.1 times the task speed. The average ISI was 1.02 seconds. On average, 254.74 cues were presented. Seven participants were excluded as they did not receive any cues and two participants were excluded hearing an average of 150 cues.

3.2.5 Experimental design

Participants completed wake behavioral tasks (Figure 3.1A) and were given 90 minutes to nap in the lab (Figure 3.1 B). The order of all behavioral tasks was counterbalanced across participants.

Here, participants were asked to complete a 30-minute initial test in their own home. This test was included to ensure that all participants could complete the control task in the lab given the allotted time slot (See section 3.2.3). Participants were then scheduled for a 4-hour in-lab session.

In the lab, participants consented to participate in the study. They next completed both the experimental and control tasks on a desktop computer (See section 3.2.3). Participants were then fitted with an EEG cap and allowed 90 minutes to nap on a futon in one of the laboratory's sleep chambers. Participants were randomly assigned to either received control cues or experimental cues during SWS (See section 3.2.4). Upon waking, participants were given time to use the restroom and wash their hair to mitigate the effects of sleep inertia. Finally, participants completed a post-nap test for the experimental and control tasks. One week later, participants were asked to complete the same tests remotely.

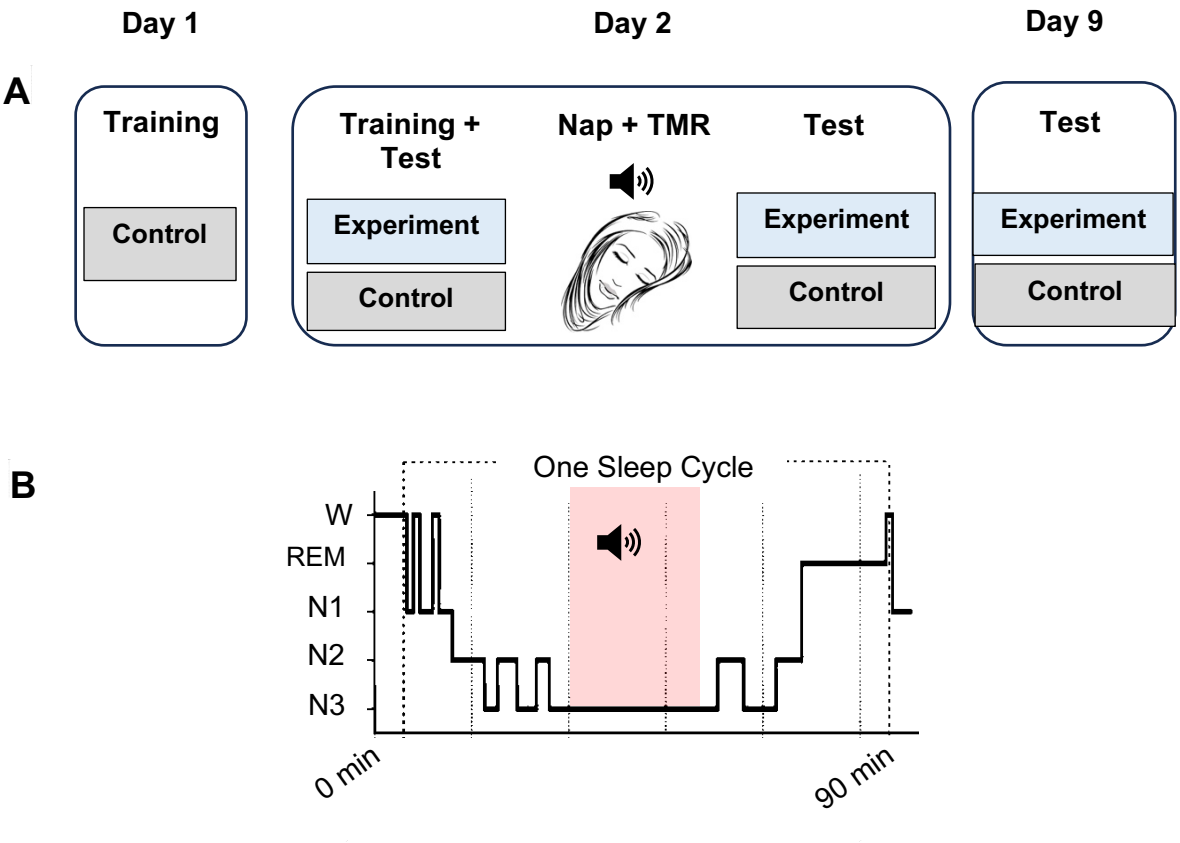

Figure 3.1. Schematic of experimental design. Participants trained on the control task before coming into the lab. In the lab, participants trained on both tasks before being given 90 minutes to nap. Following the nap, participants completed a test on both tasks. One week after the in-lab session, participants completed a follow-up test online. Participants received auditory cues either related to the experimental task or control task during slow-wave sleep.

3.2.6 Behavioral analysis

Scoring

Participant responses were scored using an online, automatic scoring tool (Borrie et al., 2019). The scoring tool calculated total number of words correctly transcribed on each trial for each participant. We allowed for all accepted exceptions in the automatic scoring tool. Response words were scored as correct if the entered word was: (1) a homophone or common misspelling, (2) included as a rootword, (3) omitting a double letter, (4) included "the" in place of "a" and

vice versa, (5) was in the incorrect tense, or (6) was entered in either it's plural or singular form. These exceptions were made to allow the scores to reflect meaningful differences in language perception.

Behavioral Analysis

Each participant contributed three behavioral scores. The training score was computed as the mean number of words correct divided by the total number of words possible for training blocks 4-6. Both follow-up test scores were computed by the mean number of words divided by the total number of words presented. Behavioral analysis was conducted in R version 4.2.3.

3.2.7 EEG Recordings and analysis

EEG analysis

EEG data were filtered in ERPLAB v9.00, a plug in of EEGLAB v 2022.1. We first used a butterworth bandpass filter with a low pass of 30Hz and a high pass of .01Hz followed by a 60-Hz notch filter. Noisy electrodes were interpolated using the spherical method in EEGLAB (less than 5% of all electrodes recorded required interpolation). EEG data were scored using the MATLAB (MATLAB, 2022) package sleepSMG (http://sleepsmg.sourceforge.net). Oscillatory patterns were identified using the MATLAB package CountingSheepPSG (https://sleepsmg.sourceforge.net) and analyzed in MATLAB version 2022b.

3.3 RESULTS

3.3.1 Training

Experiment 2

We first explored how participants performed during training. Because of a computer malfunction, two participants were not able to complete the entirety of block 1 or block 2; their data were not included in the first look at training. We saw that participants tended to improve

over the training period (Figure 3.2). As in Chapter 2 Experiment 1, average performance on the final three blocks served as a pre-nap score. A t-test revealed that for blocks 4-6, performance was similar for the experimental [M=56%; 33%] and control [M=62%; SD = 31%] conditions [$t(25.221) = 1.77$, $p = .09$]. Results were very similar when the two participants mentioned above were included in the data set [$t(26.96) = 1.69$, $p = .10$]. Because the addition of these two participants does not affect what we consider pre-nap performance, these participants were included in the remainder of the analyses.

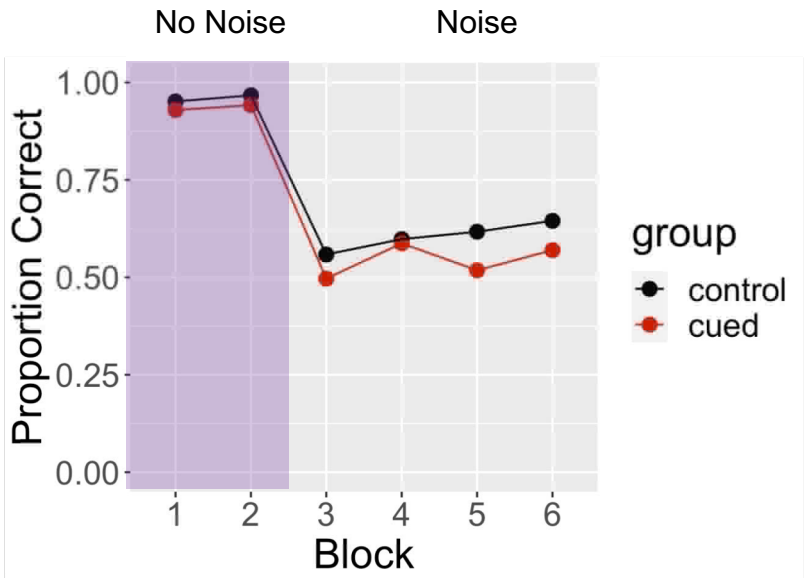

Figure 3.2. Training performance for n = 36 participants who completed the entirety of training. Speech was presented in noise for Blocks 4-6. Performance for the control and cued conditions were nearly matched for blocks 4-6 and will be considered the pre-nap score for each participant.

Experiment 1 and Experiment 2

In Experiment 1, all 60 training blocks consisted of transcription of speech in noise. In Experiment 2, transcription of the 20 initial sentences did not include noise. We also included orthographic and auditory feedback in an attempt to improve training scores. To see how these changes affected performance, we compared training and test data from the TUR single talker group from both experiments. We found that for blocks 4-6, participants correctly identified fewer words for Experiment 1 than Experiment 2, 48% correct and 59% correct, respectively [$t(24.97 = 2.22, p = .04$]. At test, participants correctly recognized 69% of words for Experiment 1 and 71% % correct for Experiment 2. Test performance for Experiment 1 and 2 was not significantly different [$p>.05$]. Interestingly, the adjustments made to training at Experiment 2 improved training performance such that training was better than Experiment 1, but this improvement did not persist into the test. It is possible that there were differences between test sentences that may have affected recognition. It is also possible that the design changes led to improvements that were not long lasting and therefore could not be detected at test.

3.3.2 Training vs Test

To get an overall picture of the way TMR may have affected the ability to recognize words, we first conducted a 2 (Test: Training, Test) x 2 (Nap condition: Cued, Uncued) mixed effects ANOVA. Performance at training and test are represented below (Figure 3.3 A; Table 3.1).

Word recognition was lower in the cued group [M=62%; SD=13%] than in the uncued group [M=68;SD=10%], as confirmed by the significant main effect of group [$F(1,36)=4.52, p = .04$]. Also, word recognition was generally better after the nap [test M=71%] compared to before the nap [training M=59%], as confirmed by the significant main effect of test [$F(1,36)=120.79, p < .01$], but the group by test interaction was nonsignificant [$F(1,36)=.63, p = .43$].

The absence of an interaction effect suggests that the lower performance for the cued group compared to the uncued group was evident to an equivalent extent both before and after the nap. That is, the group differences cannot be attributed to TMR or any other sleep-related factors. Accordingly, we ran another analysis on the pre-nap versus post-nap differences scores (test phase minus training phase). Difference scores showed that the cued group improved by 12% and the uncued group improved by 13%. This group contrast was nonsignificant [$t(35.14) = .81$, $p=.43$ (Figure 3.3 B).

Table 3.1 Performance for both groups with SD in parentheses

Group	**Training (SD) %**	**Test (SD)%**	**Difference scores (SD %)**	***t* (df)**	***p***
Cued	56.2 (2.9)	67.6 (34)	11.4 (1.4)	8.23(17)	<.01*
Control	61.8 (1.6)	75.1 (30)	13.3 (1.7)	7.68(19)	<.05*

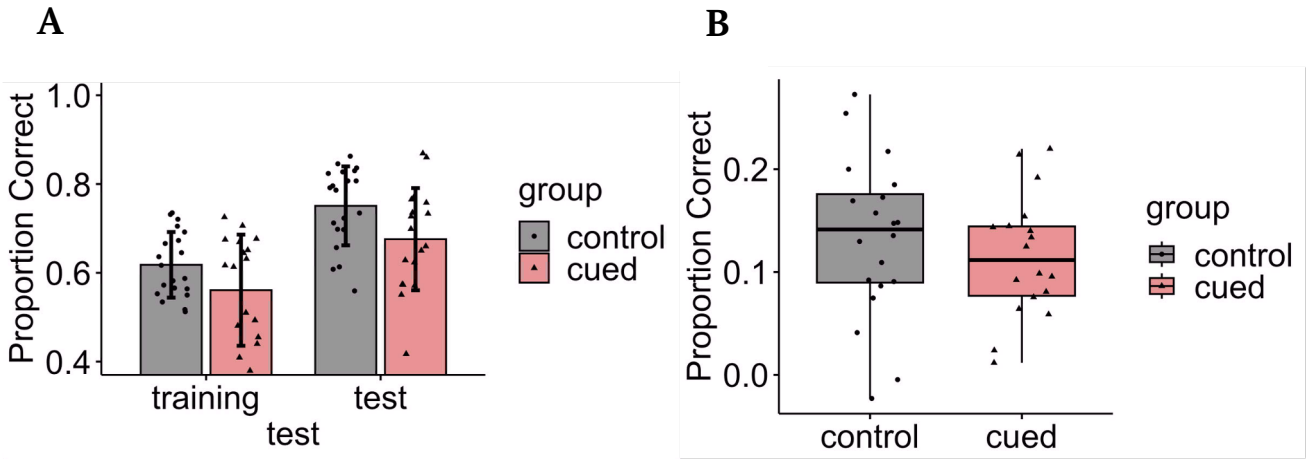

Figure 3.3. Performance for both the cued and uncued groups at training and test (A). Difference scores (test – training) are shown on the right (B). Individual subjects are represented by black circles (control) and black triangles (cued).

3.3.3 Generalization

We next explored how participants performed when faced with novel talkers (Table 3.2). First, we ran a 2 (Group: Cued, Uncued) x 2 (Match: Trained, Novel) mixed ANOVA and found a marginal effect of group [$F(1, 36)=4.49$, $p = .04$] and a significant effect of match [$F(1, 36) = 136.92$), $p < .01$]. Interestingly, we found that participants tended to perform better on novel talkers [M=75%; SD=31%] than trained talkers [M=59%; SD=34%]. Though we must keep in mind that the trained talker (TUR) was shown to be the least intelligible. There was no main effect of group x match [$F(1,36)=.03$, $p = .86$]. When examining group further, pairwise t-tests revealed that the control group performed better than the cued group after a Bonferroni correction [$p = .02$].

To explore the way groups performed based on talker, we ran a 2(Group: Cued, Control) x 4 (Talker: Far, PBR, SPA, TUR) mixed group ANOVA. There was a main effect of both group [$F(1,36) = 5.09$, $p = .03$] and talker [$F(3, 108) = 92.19$, $p < .01$] but no group x talker interaction [$F(3,108)=.01$, $p = 1.0$]. As in Experiment 1 Chapter 2, pairwise t-tests revealed significant differences in performance between all talkers [p's < .01] except for SPA and FAR after a Bonferroni correction. The control group again performed better than the cued group [$p = .001$]. Supplemental Figure 1 displays test performance for all training groups separated by speaker.

Table 3.2 Test Performance and SD for each talker at test

Talker	**Total** **Mean Correct (%) ± SD (%)**	**Cued Group** **M (%) ± SD (%)**	**Control Group** **M (%) ± SD (%)**
FAR	70 ± 31	66 ± 34	74 ± 29
PBR	84 ± 26	80 ± 30	88 ± 22
SPA	71 ± 33	67 ± 35	75 ± 30
TUR	60 ± 34	56 ± 34	63 ± 34

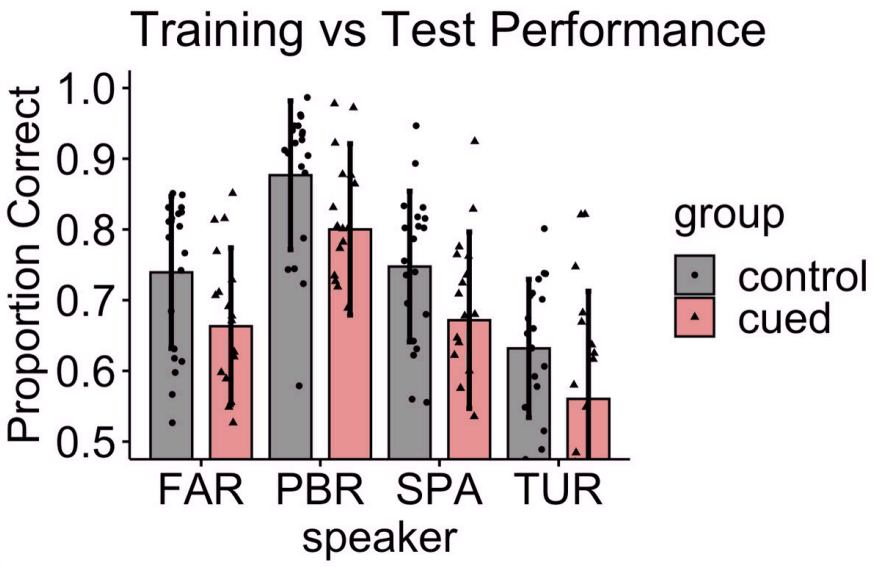

Figure 3.4. Mean performance for participants depending on group and talker. The control and cued group differed for the FAR and PBR talkers. Pairwise t-tests reveal significant difference in performance based on talker for all groups except PBR and SPA.

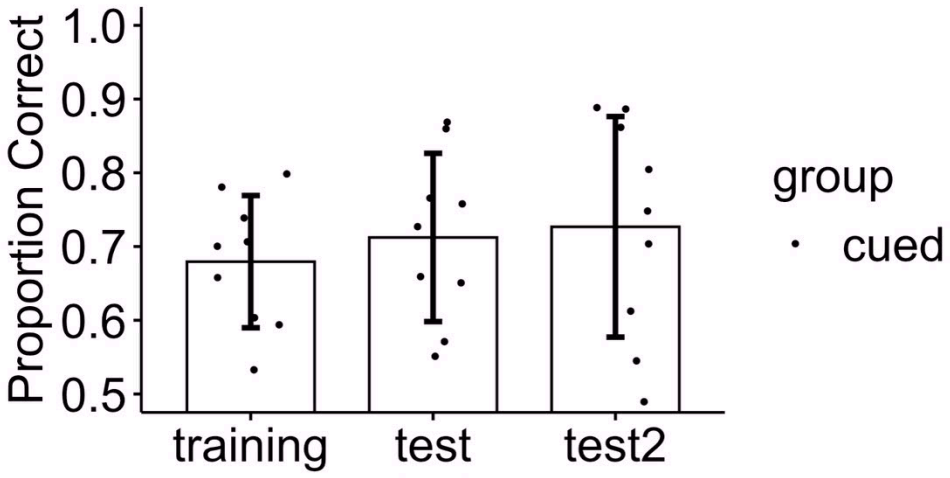

Figure 3.5. Average performance for participants who completed all three tests.

3.3.4 Delayed test

Finally, we explored how participants performed on the follow-up test that took place one week after the in-lab nap session (Figure 3.5). Due to platform issues data was compromised as participants were able to hear sentences multiple times before transcription. Because of this problem, usable data from 9 participants was analyzed. Because all 9 of the participants were in the cued group, combined with the limited amount of data, 2-sample pairwise t-tests were conducted. There was not a significant difference between training [M=68%; SD=34] and the follow-up test [M=73%; SD=32%), $t(8)=1.18$, $p = 0.27$]. A paired t-test also showed no change in performance from the immediate test to the delayed test [$t(8)=.47$), $p = .65$].

3.3.5 Describing sleep

All participants were allowed 90 minutes to nap and received auditory cueing related to either the speech transcription task or the control task. Total time asleep was similar in the uncued and the cued groups [M=74.7 (19.0) and M=73.7 (18.4), respectively; $t(35.79) = .16$, $p = .87$]. All other stages of sleep and two measures of sleep physiology were not significantly different between groups (Table 3.3).

We next explored differences in auditory cues for the two groups (Table 3.4). The control group was exposed to a series of 3 tones for each note of the 12-item sequence which served as one round of cueing. Though 36 tones were actually played per round, we considered one cue to be 3 tones of the same frequency rather than 3 separate cues, making the total number of cues per round 12 tones. The experimental group heard 60 sentences per round. Because of this, there was a significant difference between the number of cues presented [$t(29.70) = 26.69$, $p <.01$] for the control group [M=732.0;SD=89.1] and experimental group [M=120.0;SD=47.8]. Further, cues for the control group were presented at the speed of training each participant achieved. The

experimental group heard each sentence with an ISI of 5s. Due to this, there were also differences in the duration of cueing with the control group hearing cues for an average of 3.98 min [SD=1.17] and the experimental group hearing cues for an average of 10.1 min [SD=3.99), $t(19.64) = 6.24$, $p < .05$]. Due to the differences in cueing, we also compared the percentage of cues presented during each stage as this seemed to be a fairer assessment. We found no difference between groups (p's >.05) (Table 3.4).

Table 3.3 Sleep Measures and Statistics Between Groups

	Cued (M ± SD)	**Control (M ± SD)**	***p*-value**
Wake	20.6 ± 13.1	22.4 ± 21.4	.76
N1	24.8 ± 12.7	23.1 ± 13.8	.69
N2	28.0 ± 15.3	26.8 ± 8.97	.76
N3	19.8 ± 15.1	20.2 ± 13.1	.92
Delta power (Fz)	$427.0 \pm 1.55 \times 10^3$	$3.15 \times 10^6 \pm 1.40 \times 10^7$	.33
Sigma power (Cz)	1.58 ± 3.24	47.70 ± 211	.34
Slow oscillations (count)	465 ± 544	788 ± 828	.16
Spindles (count)	454 ± 152	402 ± 164	.32

Table 3.4 Between group statistics			
Raw number	**Cued (M ± SD)**	**Control (M ± SD)**	***p***
Cues in Wake (#)	.67 ± 1.71	1.30 ± 3.19	.15
Cues in N1 (#)	.61 ± 1.91	.87 ± 2.7	.31
Cues in N2 (#)	15.8 ± 15.7	55.8 ± 68.5	.004*
Cues in N3 (#)	103.0 ± 56.90	186.0 ± 70.9	<.001*
Total Cues (#)	120.0 ± 47.80	244.0 ± 29.70	<.001*
TMR duration (min)	10.10 ± 3.99	3.98 ± 1.17	<.001*
Percentage	**Cued: % total cues (M ± SD)**	**Control: % total cues (M ± SD)**	***p***
Cues in wake	.86 ± 2.24	.52 ± 1.26	.58
Cues in N1	.78 ± 2.54	.33 ± 1.06	.49
Cues in N2	18.40 ± 7.20	22.50 ± 26.60	.61
Cues in N3	79.90 ± 23.4	76.70 ± 27.4	.69
Total cues	NA	NA	NA
TMR duration (min)	NA	NA	NA

3.3.6 Effect of sleep on performance

Between group comparisons

To explore the effects of sleep on performance, participants were split into two groups: Cued and control. A difference score was created by subtracting the average training score from the average test score such that a positive number indicates improvement from pre-sleep to post-sleep. After calculating this improvement score, a series of Pearson correlations were run to

explore the effects of different sleep measures considered to play a role in memory consolidation (Table 3.5). When looking at all participants, neither time spent in N3, total time asleep, delta power measured at electrode Fz, nor sigma power measured at electrode Cz correlated with improvement on the speech transcription task [*p*'s > .05]. We did find a significant effect of N2 [$r(36)=.37$, $p = .02$]. It appeared that this finding was driven by the control group, [($p = .18$) for the cued group and ($p=.02$) for the control group]. We also explored the relationship between TMR duration and improvement. We did not find a correlation between the duration of TMR and improvement scores [for all participants ($p=.44$), the control group ($p=.91$), or the cued group ($p=.69$)].

Table 3.5 Correlation of sleep measures on improvement for all participants

Measure	***r***	***p*-value**
N2 min	.37	.02*
N3 min	.23	.17
Time asleep	.21	.20
Delta power (Fz)	.01	.94
Sigma power (Cz)	.04	.83

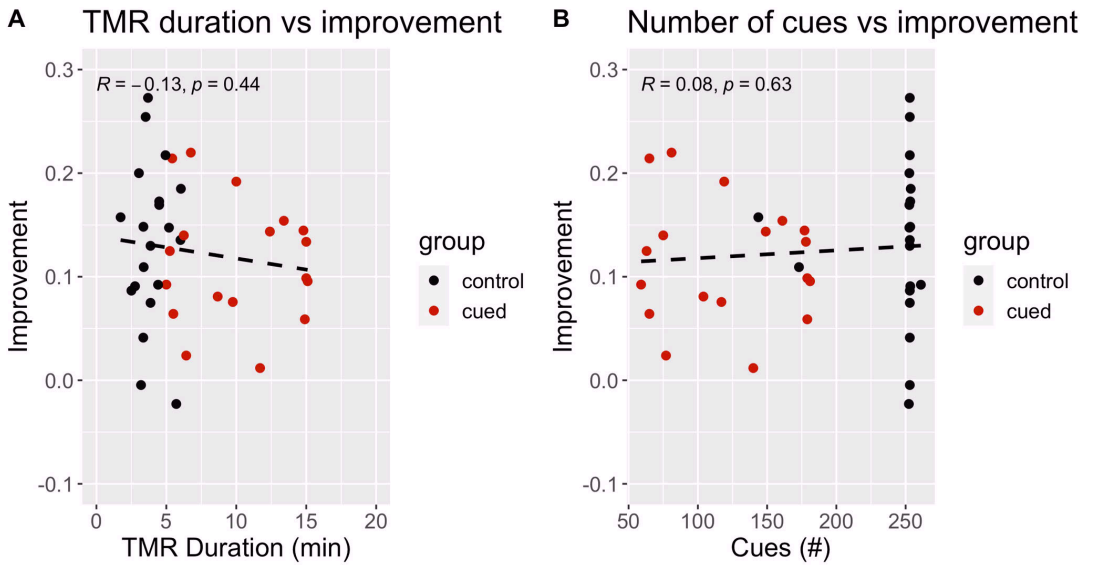

Figure 3.6 TMR duration by improvement (A). Number of cues presented during the nap by Improvement (B). The control cued participants are shown in black, and the task cued participants are shown in red.

Cued group

Because of the differences in cueing between the cued and control group, outlined above, we next ran a series of Pearson correlations to explore how EEG measures and memory reactivation affected performance for only the cued group (Table 3.6). Neither time in N2, time in N3, delta power measured at electrode Fz, sigma measured at electrode Cz, nor the total number of cues or cueing duration were correlated with the memory improvement score. There was a significant effect of total time asleep such that more time asleep correlated with a greater improvement from training to test [$p = .047$] (Figure 3.6A). Time asleep was not correlated with improvement for those who received the control cues [$p = .86$].

Recent research has also shown that auditory cues may disrupt sleep and negatively affect the memory consolidation process, correlating with a decrease in performance (Whitmore,

Bassard, & Paller, 2021; Whitmore & Paller, 2023). For the cued group, each auditory cue was presented 5s apart. The sleep disruption score was measured as the absolute change in the power spectrum 3s before cue onset compared to 5s post cue. There was a marginal effect of sleep disruption for the participants cued with spoken sentences such that a greater change in the EEG power spectrum after cue presentation compared to before cue presentation was marginally correlated with a decline in performance from training to test. [p = .06] (Table 3.6; Figure 3.6B), though there was one participant who experienced much more sleep disruption than the others. We cannot be sure whether sleep disruption played a role in poorer improvement or if this participant would have scored low on L2-English transcription more generally.

Time asleep and sleep disruption were not correlated [p=.10]. Because the control and cued group spent a similar time in each stage of sleep (Table 3.3), it may be that task relevant cues in addition to longer and less disrupted sleep play a role in successful generalization.

Table 3.6 Correlation of sleep measures on improvement for cued participants

Measure	***r***	***p*-value**
N2	.33	.18
N3	.25	.32
Time asleep	.47	.05*
Total number cues	.10	.69
Cueing Duration	.10	.69
Sleep disruption	.45	.06
Delta (Fz)	.08	.74
Sigma (Cz)	.05	.83

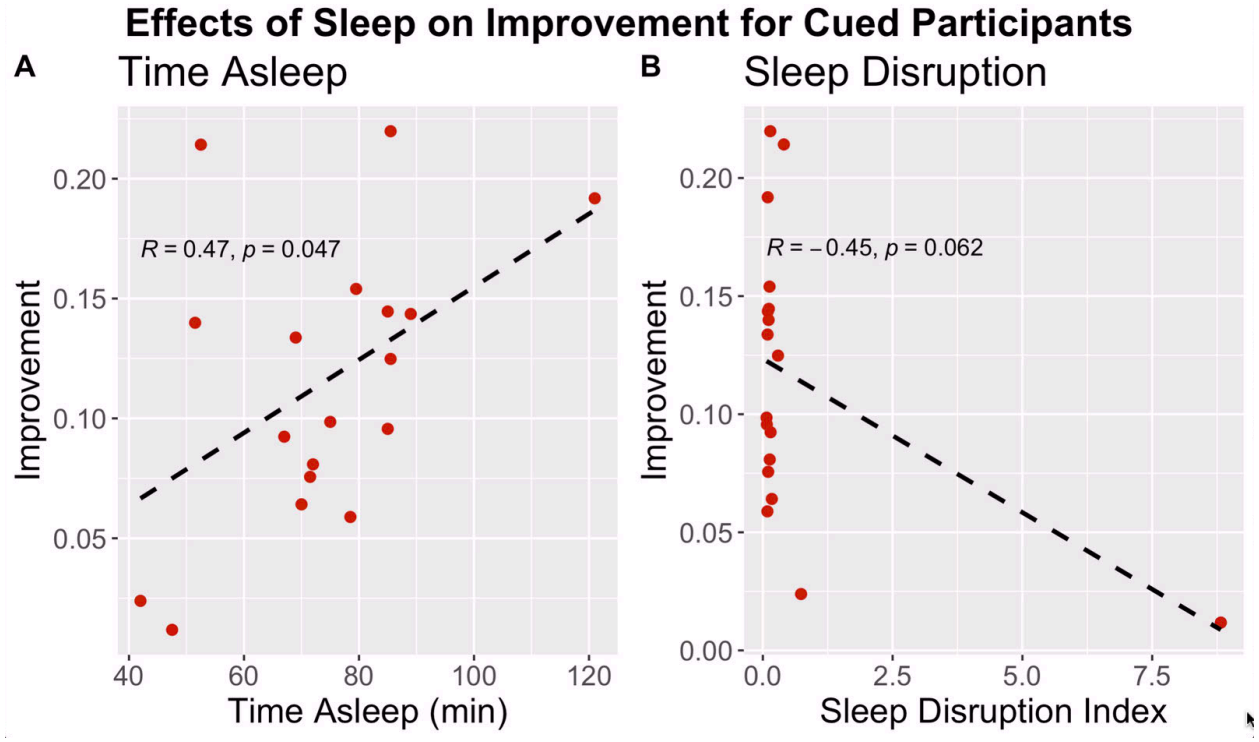

Figure 3.7. Pearson correlations predicting improvement by total time asleep (A) and sleep disruption (B). Individual participant data are shown as red dots.

3.4 DISCUSSION

It is very common for us to converse with people who may have a different first-language (L1). This may affect one's production of a second language (L2). Though it may be difficult to understand a new L2 speaker, we adapt relatively quickly though factors such as a noisy environment and the idiosyncratic ways in which a person speaks can prolong this process. What we are learning and how this information is transformed during sleep is not well understood. One way to study this type of learning is by asking L1-English participants to transcribe L2-English in noise and see how they perform over time (Strori, Bradlow, & Souza, 2020).

In Chapter 2, we found that participants who were exposed to a native Turkish talker improved at transcription for this talker as well as talkers of other language backgrounds. In this

experiment, we explored the effect of sleep and memory reactivation for an L2-English transcription task in which English dominant speakers were only exposed to the same Turkish talker as in Experiment 1 (Chapter 2). Participants also completed a finger tapping sequence control task. During a 90-minute nap, participants were exposed to the auditory cues during learning of either the transcription task or the control task. The between-subjects design was chosen based on the nature of the transcription task. Because we felt it was likely that one cued sentence, or even a single word from that sentence may have reactivated the entire task, therefore activating the perceptual process of L2 word recognition. The between-subjects design helped to mitigate these possible "bleed-over" effects. It is a possibility, however, that the mere presentation of auditory cues could affect post-nap behavior. Including the control task in the design made it possible for all participants to receive a task-related cue. Although there were differences in cues — for example one group heard sentences while the other heard tones — this design allowed for auditory cues during sleep to be distinct but also present.

We determined that the cued and control conditions were statistically matched for performance at training. Though this was the case, we must acknowledge that the cued group did start at a slight disadvantage, transcribing 6% fewer words than the control group (See section 3.3.1). Group was counter-balanced across participants and would likely have appeared more equal with a greater number of participants in each group.

Though both groups were able to correctly identify more words at the post-nap test compared to the pre-nap test, TMR did not improve performance for the cued group compared to the control group. Though it is not easy to identify one clear reason, there are hypotheses as to why TMR was unsuccessful in improving performance on this task. First, it is possible that the cues were not presented at a volume loud enough to be processed by the brain during sleep.

Because speech sentences are more complex than tone cues, our aim was for the sentences to be played loud enough such that individual words could be processed but that they were not loud enough to wake the participant. Though we increased the volume after each round was presented, it is still possible that our cues were not loud enough to be processed. Similarly, it may be the case that spoken sentences are not as clear as auditory cues, such as tones or characteristic sounds which are often used, and failed to elicit memory reactivation. Aside from TMR, it is possible that this type of learning may improve with wake practice and not during sleep. For example, one could imagine that learning a new language is strengthened while you are sleeping but that the largest benefits come from conversing with others in the language. Though this may be the case, we might consider changes to this protocol so that researchers can use this technique to study how sleep physiology may help us understand how neural processes may play a role in auditory perception.

Another reason may be the nap design. It has been hypothesized that REM sleep is helpful for integration and generalization (Sterpenich et al., 2014; Tamminen et al., 2017; Witkowski et al., 2020). Because our participants did not have sustained REM, as it is more prevalent in the second half of the night, it is possible the afternoon nap design hindered this generalization process. When looking at the sleep data, we found that the control and cued group spent a similar amount of time in all sleep stages of sleep and had a similar total time asleep. Many of these measures did not correlate with change in performance from pre-nap to post-nap, though we did find a positive correlation between time spent in N2 and improvement from pre-nap to post-nap for all participants., During N2 sleep, there is an abundance of thalamic spindles. It is possible that sleep spindles played a role in this time of memory consolidation as improvement did not correlate with other stages of sleep. While the effect may have been driven

by the control group a positive correlation seemed to emerge (r = .33). A more sensitive measure of N2, such as the number of sleep spindles , and their relation to improvement after sleep should be explored. Because we found no TMR effect, we explored sleep measures within the cued group to better understand why this may have been the case. First, we found that the total time asleep was correlated with participants' improvement score such that those participants who spent more total time asleep correctly identified more words on average at post-test compared to pre-test.

Next, we explored sleep disruption as it has been hypothesized that disruptions during sleep following cue presentation may negatively affect memory (Whitmore et al., 2021; Whitmore & Paller, 2023). Here, sleep disruption was considered a change in the power spectrum before and after cue presentation. We found that sleep disruption was marginally, negatively correlated with improvement where participants with more sleep disruption improved the least from pre-test to post-test. When presenting auditory cues during a nap, one goal is to do so without waking the participant which explains why most participants scored relatively low on this measure of sleep disruption. Though we cannot describe the change in performance on this task for those with moderate sleep disruption, this result follows the literature on sleep disruption as it was negatively associated with behavioral performance. When looking at the plot, however, one participant had much greater sleep disruption than the others and also performed the worst which influenced the results (Figure 3.6). It is possible that this participant had an overall lower performance due to other reasons that cannot be attributed to sleep disruption which could have led to a spurious finding.

In summary, we found that participants could improve on the TUR talker despite it being the least intelligible. We also found that the PBR and TUR talkers were the most similar as there

were no significant differences in performance. We did not, however, find a performance effect of task-relevant cues. The possible explanations above may explain this finding. We did find an effect of the total time asleep such that more time asleep correlated with an improvement in the L2-English transcription task and not the control task. A more fine-grained measure of sleep, such as the number of sleep spindles and slow oscillations, should be explored. Finally, we did find an effect of sleep disruption as disruption in relation to the cues correlated with worse performance suggesting that more quality sleep may be important for generalization.

Overall, interesting interactions between performance on this task for different speakers and TMR groups seemed to emerge. This experiment adds to the literature by exploring generalization of L2 speech and exploring sleep physiology to better understand this process. Participants in Experiment 1 and Experiment 2 completed very similar tasks with the exception of the type of training and the inclusion of a nap between the training and test (see Sections 2.2.3 and 3.2.3 for task descriptions of Experiment 1 and 2, respectively).

Chapter 4

In-Depth Statistical Analysis

4.1 INTRODUCTION

Chapter 2 and Chapter 3 of this dissertation detailed a second-language (L2) English transcription task by either first-language English and/or English dominant listeners. The task included one native Farsi (FAR), Brazilian Portuguese (PBR), Spanish (SPA), and Turkish (TUR) talker for a total of four talkers. In Experiment 1 (Chapter 2), We were interested in testing whether exposure to multiple talkers or a single talker affected performance on familiar and novel talkers after a delay. Here, 168 participants transcribed 60 sentences spoken by either 3 or 1 of the mentioned talkers. During this initial training session, participants completed one of two training types: multiple-talker and single-talker where groups were exposed to three or one talker, respectively. Approximately 11 hours later, participants completed a test which involved the transcription of sixty novel sentences spoken by all four talkers.

In Experiment 2 (Chapter 3), we tested whether sleep as well as targeted memory reactivation (TMR) a technique in which auditory stimuli paired with a leanring episode are presented unobtrusively during slow-wave sleep (SWS). Thirty-eight participants completed the same general task; however they were only exposed to the TUR talker at training. At test, all talkers were included. This time, training and test were separated by a 90-minute nap during which auditory cues were presented in SWS. During sleep, participants were either exposed to the L2 English sentences heard during the transcription task or auditory cues paired with a control task which tested implicit learning of a motor sequence task. Each participants transcribed the same sentences, presented in a random order, at training and test. Data from Experiment 1 was collected from the online platform Prolific while data from Experiment 2 was

collected in the laboratory. Although there are design differences between the two studies, here, we pooled and analyzed the data collected from both in order to further assess previous findings and explore relationships between variables. We did so by employing a mathematical modeling approach known as multi-level modelling (MLM).

Regardless of design differences stated above, data collected from both experiments can provide valuable insights related to L2-English transcription. In Chapters 2 and 3, we thoroughly analyzed the data using standard techniques which are popular in the psychology field, particularly with ANOVA and *t*-tests. These analyses allowed us to directly compare differences in groups based on training type as well as individual training group, speaker transcribed, and whether or not participants received task-relevant cues. However, it becomes difficult to consider these many things in one model. Often, one may run an ANCOVA or ANOVA. Follow-up t-tests may then be necessary to explore where the significant difference in groups lie. Further one must decide which correction will be used to account for the multiple comparisons.

Multi-level models, also referred to as hierarchical linear models (HLM), have often been used to assess data that have a natural, hierarchical structure (J. Hox, 1998; Stephen & Anthony, 2002). MLM has also been used to combine data for the purpose of meta-analyses. Though a hierarchical structure may not be as apparent as the previous example, participant data collected can be considered to be nested within conditions which are further nested within studies (Fernández-Castilla et al., 2020). We chose MLM to assess our data as it accounts for variance by incorporating how the higher levels of hierarchy affect the dependent variable (Steenbergen & Jones, 2002). MLM also allows for results that can be more easily generalized. Because of the grouping factors present in both experiments, with MLM we can incorporate these differences into the model by incorporating them as a level when possible. Finally, MLM is able to do so

within one concise model (Steenbergen & Jones, 2002). Due to the potential advantages of MLM, in this chapter, we explored the use of this statistical tool in psychology.

There are decision points which must be considered when building a multi-level model. One is the way you will categorize the predictors. There are two options: fixed or random effects. A fixed effect is considered to have a constant relationship with the response variable. A fixed effects model is a linear regression of y on x. MLM computes an estimate, often based on maximum likelihood estimation, for each fixed effect in relation to y, controlling for all other included predictors, is computed (Clark & Linzer, 2015; Hayes, 2006). In other words, a fixed-effect coefficient estimates how a change in the predictor will affect the dependent variable. Random effects are considered those which vary in their relationship to y. Here, rather than directly calculating coefficients, such as in fixed effects, they are assumed to follow a normal distribution where the mean and standard deviation of the normal distribution are used as estimators (Clark & Linzer, 2015). Generally, random effects provide information regarding the variance of the dependent variable at different levels.

Here, we used the mixed-models approach, which includes both fixed and random variables, to explore data collected from Experiment 1 and Experiment 2. We did so to explore the use of MLM to analyze data which is typically analyzed with ANOVA and t-tests. Doing so also allowed us to compare findings from this approach with the more typical approaches. We also pooled the data from Experiment 1 and Experiment 2 together to build one final model. Doing so allowed us to compare data from both experiments in one model. It also accounted for the differences in each experiment that may affect the findings, as they could be included as levels in the hierarchy.

4.2 METHOD

4.2.1 Data

Data: Scoring

Participant responses were scored using an online, automatic scoring tool (Borrie et al., 2019). The scoring tool calculated total number of words correctly transcribed on each trial for each participant. We allowed for all accepted exceptions in the automatic scoring tool. Response words were scored as correct if the entered word was: (1) a homophone or common misspelling, (2) included as a rootword, (3) omitting a double letter, (4) included "the" in place of "a" and vice versa, (5) was in the incorrect tense, or (6) was entered in either it's plural or singular form. These exceptions were made to allow the scores to reflect meaningful differences in language perception.

Behavioral Analysis

Each participant contributed three behavioral scores. The training score was computed as the mean number of words correct divided by the total number of words possible for training blocks 4-6 (30 total sentences). Both follow-up test scores were computed by the mean number of words divided by the total number of words presented. Behavioral analysis was conducted in R version 4.2.1.

4.2.2 Model building

Variables

The variables for possible inclusion consisted of participant ID, proportion correct (proportion of words correctly identified variable), format in which the test was completed (online, in-lab), session (training, test), speaker transcribed (FAR, TUR, PBR, SPA), type of training (single-talker, multiple-talker), and TMR group (cued, control). The variable, proportion correct, served

as the response variable (Table 4.1). The intercept and participant ID were included as random variables. The rest of the variables included in all models were considered fixed effects which allowed us to meaningfully estimate the effect of each predictor variable and interaction term on performance.

Running the models

The model was created using R version 4.2.3 using the "nlme" package. Because of the nature of the data, a mixed effects model, consisting of both random and fixed variables, were used. A mixed model follows the equation:

$$y\text{ij} = \beta_0 + \beta_1 x_{ij} + u_i + \cdots + e_{ij}$$

Where *y* is the response variable, β_0 is the intercept, β_1 is the coefficient of the first fixed effect term, and *i* is an individual observation for each level, *j*. *u* is a random-effect, and *e* is an error term.

To compare the models, the Akaike information criterion (AIC) and Bayesian information criterion (BIC) were used. While there is no set number that is considered a good indicator of fit, a lower AIC and BIC indicate a better model fit. ANOVA was used to compare models one by one. Due to our relatively small sample size, restricted maximum likelihood estimates (REML) were used (McNeish, 2017).

Table 4.1. All possible variables	
Variable Name	**Variable Type**
ID	Random
Proportion Correct	Response
Format	Fixed
Test	Fixed
Type	Fixed
Speaker	Fixed
Cueing	Fixed

Hierarchy

Chapter 2: Experiment 1 Online Data

The data collected online from Experiment 1 were separated into three levels, represented in Figure 4.1. Participant ID was included as a random effect as we expected performance to vary across participants in no particular pattern. From lowest to highest, the hierarchy for Experiment 1 data included, the four transcribed talkers, the session, training group, and the type of training one received (single-talker or multiple-talker.

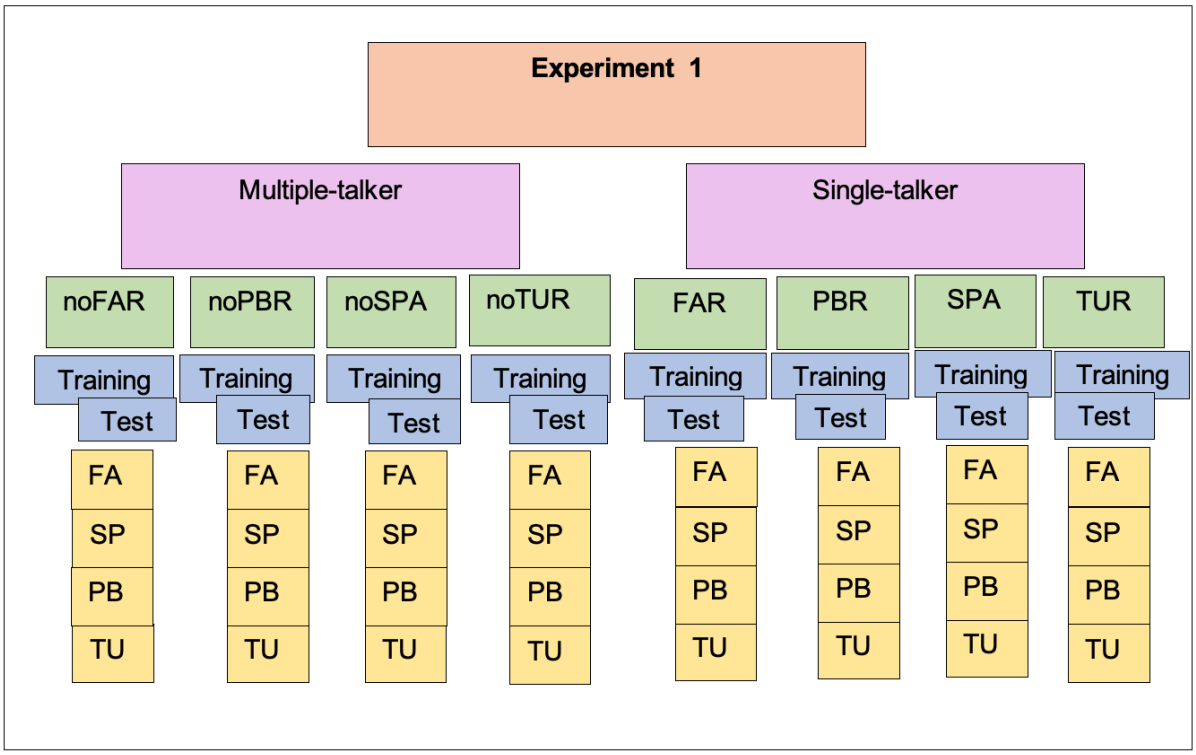

Figure 4.1 Hierarchy of data from experiment 2. Level 1 includes speakers. Level 2 includes the session, which is nested within level 3, training group, and finally level 4 training format.

Chapter 3: Experiment 2 In-lab Data

Data from Experiment 3 was separated into two levels represented in Figure 4.2. Level 2 was separated by those who received cues related to the second-language English transcription task and those who received control cues. Data was then further grouped by session: training and test (level 1). Participant ID was included as a random variable.

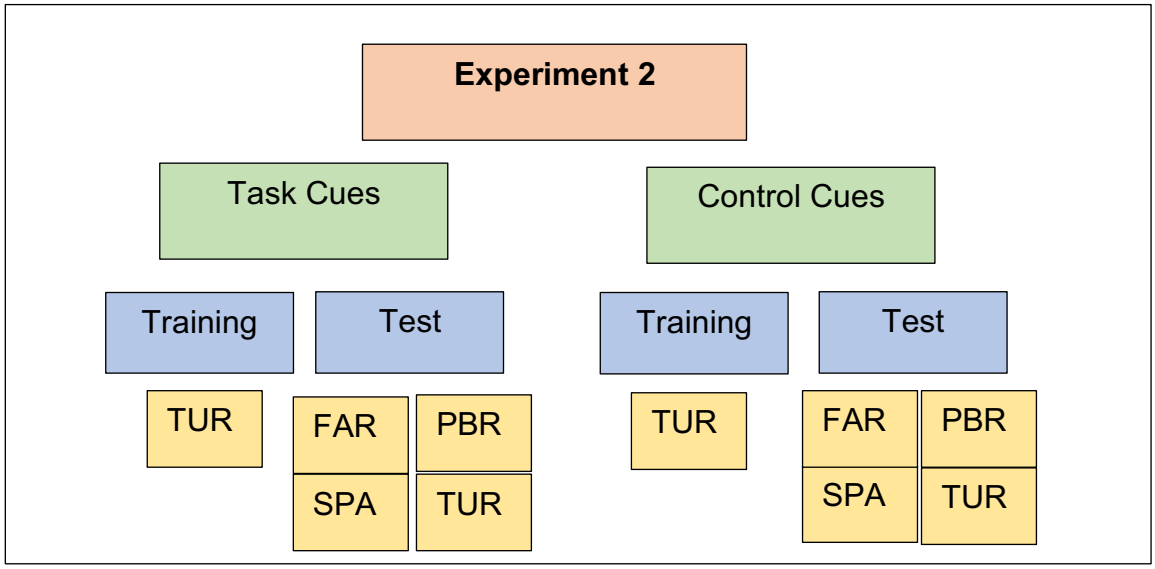

Figure 4.2. Hierarchy of Experiment 2 variables. Level 1 is the session, which is nested within level 2, the cueing format.

Experiment 1 and 2 Combined Data

For the combined data, the broadest level was separated by format: In-lab and online data (level 3). Data was then further separated into training and test data (level 2). Talker was also included as a level one fixed variable. All levels are shown in Figure 4.3. Due to the single-talker in-lab study design, talkers PBR, SPA, and FAR were only included at test. Due to this, we only interpreted coefficients of talker at test to try and better understand the effect of talker for the in-lab and online studies. It is also important to keep in mind that any effects related to the three mentioned talkers are only relevant for the test session.

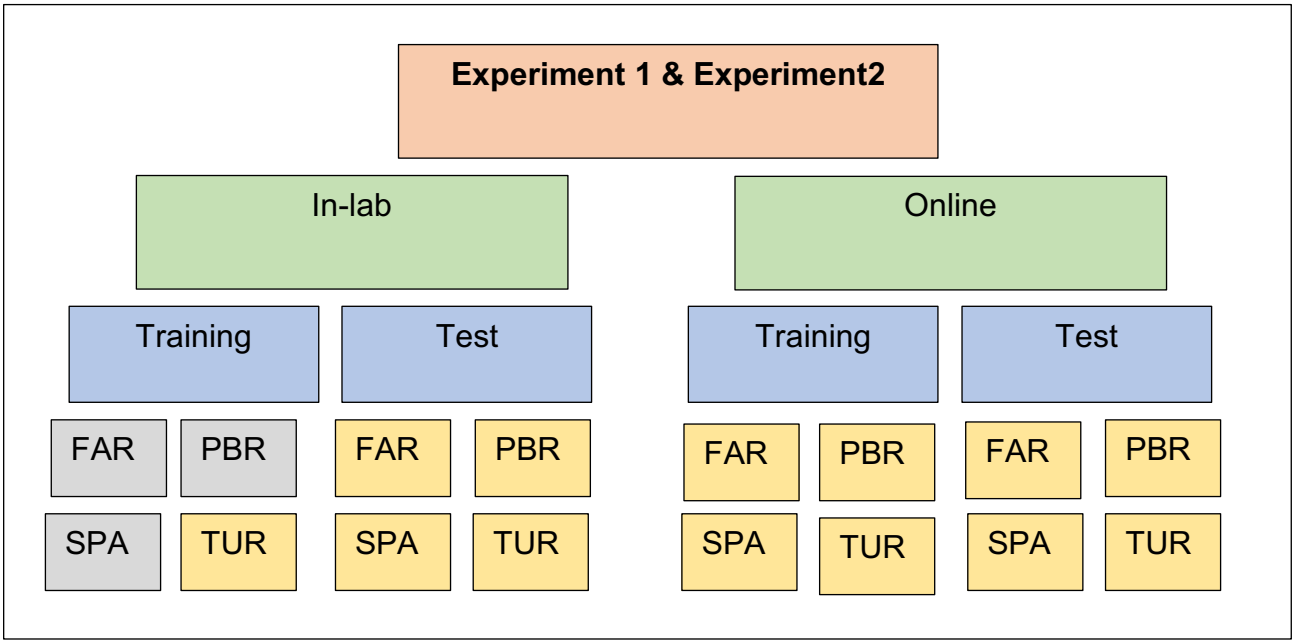

Figure 4.3 Hierarchy of experiment 1 and 2 combined data. Level 1 includes speaker. Level 2 includes the session which is nested within the format in which one completed the task. FAR, PBR, and SPA are represented in grey as these speakers were not included in the in-lab training session.

4.3 RESULTS

4.3.1 Chapter 2: Experiment 1 online model

To fit these data, we first created a simple regression model including participant ID as a random effect at the intercept level to predict the response variable proportion correct. We first ran an intercept only model including ID to predict performance. This was compared with a model which also included the type of training (multiple-talker or single-talker) as a level one fixed effect with no changes to the random effect. There was no difference between these two models [p=.70]. Next, we included test type (training and test) as a level two fixed effect (model 3) though including test at level two did not improve the model [p = .74]. However, including test, type, and a test by type interaction term only as fixed effects did improve the model [p < .05]. We then compared this model with one that included talker as a level three fixed effect which

significantly accounted for more error [$p<.05$]. Finally, we determined whether including two- and three-way interactions between type, test, and talker improved the model. This was included following the a priori hypothesis that the type of training one receives, the session, and the speaker heard would interact in a way that affected performance. This proved to be the best performing model as it accounted for the most overall error (Table 4.2, Model 6).

For those who completed the experiment online, performance was greatest for the PBR and SPA talkers compared to other talkers [participants performed an average of 12% better ($t(168) = 13.66$, $p < .05$) for the PBR talkers and 5% better for the SPA talker ($t(168) = 3.05$, $p < .05$) controlling for other variables and interaction terms]. At test, however, participants tended to perform worse (-4%) on the SPA talker [$t[168] = 2.68$, $p = .007$] compared to other talkers. Data for these models are displayed in Table 4.3 The intraclass correlation (ICC) represents how much of the variance can be attributed to between subjects rather than within-subjects. The ICC for the best performing model was 12% which indicated that most of the variance (88%) came from within participants.

Though findings were significant, it is important to take into consideration the degree to which a predictor affects the outcome variable, measured by the estimate, and the degree to which this finding is reliable based on the confidence interval. There were 627 total words which contributed to the score with 310 at training and 317 at test. Because participants were split into eight groups with the number of speakers changing, we can estimate that each speaker spoke roughly a quarter (157) of all total words. Considering this, we expect an improvement of 8 words on average for the SPA talker and an increase of 19 words on average for the PBR talker. Further, we would expect a decrease of 3 words at test for the SPA talker. Overall, the largest beta estimate is that of the PBR talker as nearly 19 more words (12%) were estimated to be

transcribed with a rather narrow 95% confidence interval [10%, 14%]. This finding indicated that this speaker may indeed be the most intelligible out of the four included.

Table 4.2. Models tested to fit Experiment 1 data

Model	Variable(s) Included	AIC	BIC	Comparison	*p*-value
1	ID	7277.11	7300.00	NA	NA
2	ID, Speaker	5687.7	5733	1 vs 2	<.05
3	ID, Speaker + Test	5689.40	5742.80	2 vs 3	.59
4	ID, Speaker + Test + Group	5691.10	5797.80	4 vs 2	.12
5	ID, Speaker + Test + Group + Type	5691.10	5797.80	5 vs 2	.12
6	ID, Speaker * Test	5622.60	5698.00	6 vs 2	<.05
7	ID, Speaker * Test * Group	5607.4	5988.60	7 vs 6	<.01*
8	ID, Speaker * Test * Group * Type	5607.4	5988.60	8 vs 7	NA

Table 4.3 Significant statistics from best fitting model for Experiment 1 Data				
Predictor	**Estimate (β) (%)**	**95% CI (%)**	***t***	***p***
Intercept	72	70-75	54.45	<.001
PBR	12	10-14	13.66	<.001
SPA	5	2-8	3.05	.002
SPA * Test	-4	-1 - -7	2.68	.007
$\sigma^2 = .08$ ICC=.11 *n* = 168				

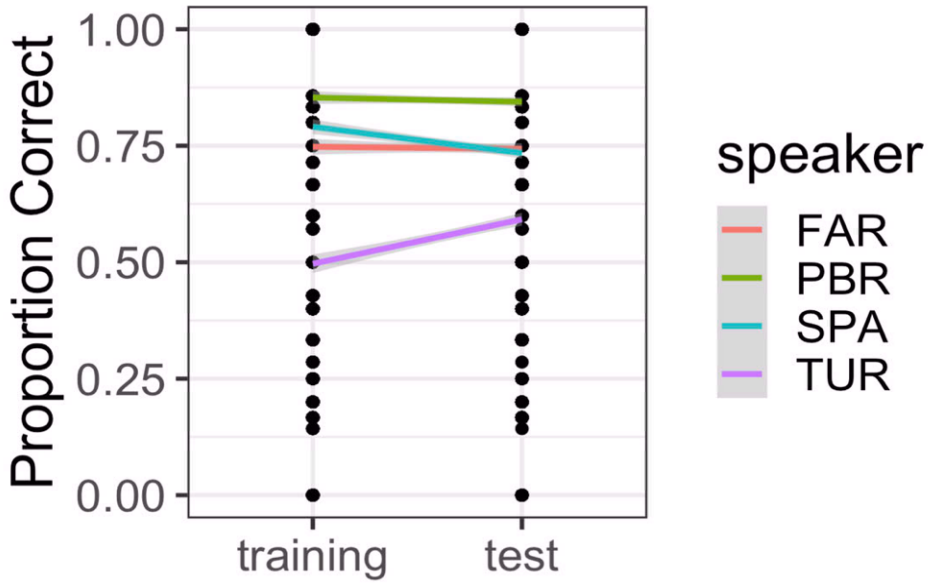

Figure 4.3 Data representing performance at test and training for each of the four speakers.

4.3.2 Chapter 3: Experiment 2 in-lab model

For the data collected from Experiment 2, we first compared a model including a fixed effect of test with an intercept only model which performed significantly better (p<.01). The addition of Speaker, however, improved the model even further (p <.01). Interestingly, the inclusion of the test or the interaction between TMR group and Test was not significant (p's > .05) indicating that the increased complexity did not account for any additional variance (Table 4.4). Based on the best performing model, we found that participants speaker accounted for the majority of the error with participants performing the best on the PBR group (14%) and worse on the TUR group by 11% (p's<.01) (Table 4.5) controlling for all talkers. Grouped data is represented in Figure 4.4. The ICC indicated that 90% of the variance is attributable to within-person performance.

There was a total of 607 words transcribed with 290 at training and 317 at test. Given this information, one would expect an additional 44 (14%) words to be transcribed at the test session on average controlling for all other talkers. We would also expect 44 fewer words to be correctly transcribed on average for the TUR talker between both the training and test sessions.

Table 4.4 Models tested to fit Experiment 2 data

Model	**Variable(s) Included**	**AIC**	**BIC**	**Comparison**	***p*-value**
1	ID	1852.80	1871.10	NA	NA
2	ID, Speaker	1565.60	1602.3	1 vs 2	<.01
3	ID, Speaker + Test	1567.40	1610.20	2 vs 3	.66
4	ID, Speaker + Test + TMR group	1564.60	1613.50	2 vs 3	.08
5	ID, Speaker * Test	1567.40	1610.20	2 vs 5	.66
6	ID, Speaker * Test * TMR group	1571.90	1645.30	6 vs 2	.46

Table 4.5 Significant statistics from best fitting model for Experiment 2 data				
Predictor	**Estimate (β)(%)**	**95% CI**	***t***	***p*-value**
Intercept	70	66-74	35.04	<.001
PBR	14	10-17	7.69	<.001
TUR	-11	-14 - -8	7.51	<.001
$\sigma^2 = .09$ ICC=.09 *n*= 38				

Figure 4.4 Performance for the training and test sessions for each speaker.

4.3.3 Experiment 1 and 2 combined data model

In this analysis datasets from both experiments were included. We explored a simple model, including ID as a random effect, with one that included speaker as a fixed effect. The latter

model was significantly better (p<.05). The addition of both test and format also improved the model. Finally, the inclusion of all interaction terms significantly accounted for the greatest amount of error (p<.05) (Table 4.6 Model 5).

Performance was calculated as the number of words correctly identified divided by the total number of words presented. There were 627 words in Experiment 1 and 607 words for Experiment 2. For this analysis we will consider there to be a total of 617 words (average number of words in Experiment 1 and Experiment 2) with each talker speaking roughly 154 words. Overall, we found that participants performed better at test and for the PBR and SPA talkers [PBR performance was greater by 13% and the SPA talker by 3%]. Performance at test for the in-lab participants was significantly worse in general [-1%; p<.01] and when faced with the SPA talker [-3%; p <.05]. Finally, participants who completed the test in-lab were worse for the FAR talker [-1%; p<.05] and at test [-1%; p<.05]. Statistics for all predictors are displayed in Table 4.7 and represented visually in Figure 4.5. The majority of the variation was attributable to within person differences (88%).

We found that the greatest predictor affecting performance is that of the PBR talker with a 13% increase, which translates to roughly 20 words, in performance. The narrow confidence interval [12%, 14%] suggests this was an accurate estimate. The remaining estimates fell between 1% and 3%, or roughly 3 to 9 words, respectively. Though these estimates had a less substantial impact on improvement of performance, they are still reliable in their prediction of performance outcomes based on confidence intervals (Table 4.7)

Table 4.6 Models tested to fit Experiment 1 and 2 data

Model	**Variable(s) Included**	**AIC**	**BIC**	**Comparison**	***p*-value**

1	ID	9139.70	9163.20	NA	NA
2	ID, Speaker	7274.70	7321.60	2 vs 1	<.05
3	ID, Speaker + Test	7276.4	7331.2	3 vs 2	.60
4	ID, Speaker + Test + Format	7278.30	7340.90	4 vs 2	.82
5	ID, Speaker * Test	7229.9	7308.10	5 vs 2	<.05
6	ID, Speaker * Test * Format	7199.2	7316.5	6 vs 5	<.05*

Table 4.7 Significant statistics from best fitting model for Experiment 1 and 2 data

Predictor	**Estimate (β) (%)**	**95% CI (%)**	***t***	***p*-value**
Intercept	73	71-75	70.33	<.001
PBR	13	12-14	19.86	<.001
SPA	3	2-4	4.86	<.001
Test	-1	-2-0	2.91	.004
SPA * Test	-3	-4 - -2	6.47	<.001
FAR * In-lab	-1	-2 – 0	2.14	.033
Test * In-lab	-2	-2 - -1	3.28	.001
$\sigma^2 = .08$ ICC=.12 *n* = 206				

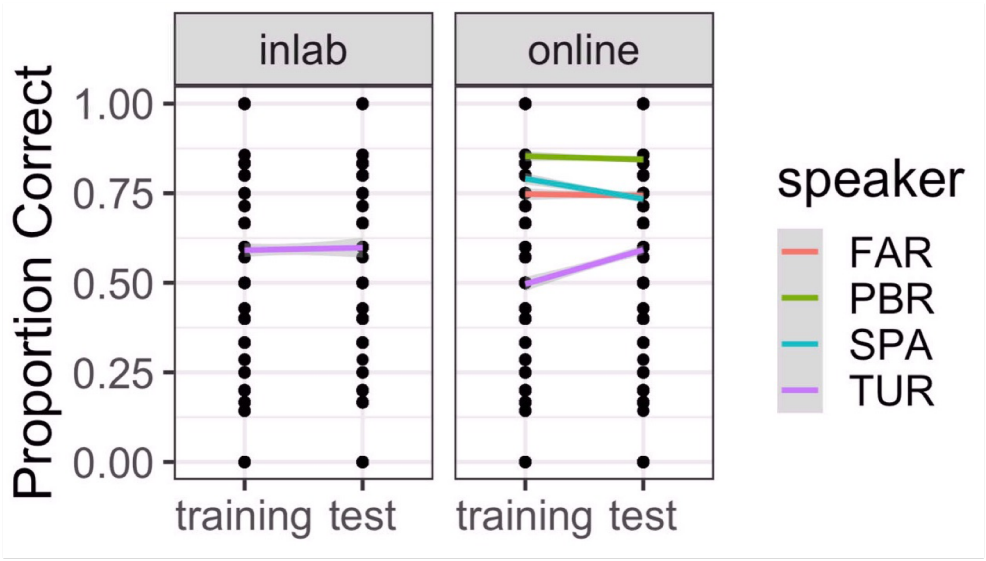

Figure 4.5 Performance for data collected both in-lab and online at each session. Change in performance from speaker to test is also represented where applicable.

4.3.4 Testing model assumptions

There are assumptions that MLM makes which should be checked in order to have confidence in estimates and interpretation. First, we explored the assumption of linearity. To check this assumption, we plotted the residuals of the best fitting model for each dataset, looking for no clear pattern between the residuals and outcome variable. Fox (2008) argues that visual inspection is sufficient to test for linearity. We next test for the assumption of homoscedasticity, or an equal variance in the residuals. To do so, we ran an ANOVA on the square of the residuals. A nonsignificant result indicates that the assumption of homoscedasticity is met. If this assumption is not met, a transformation may be required. We also tested that the residuals of the model were normally distributed by visually assessing a plot of standardized quantiles by standardized residuals, looking for a near perfect correlation. Finally, we checked for multicollinearity amongst predictors to see whether there was a difference in performance that

could be attributed to a correlation amongst predictors. Because all of the predictors were categorical variables, we ran a Chi-squared test on the relationship between each of the predictors with the dependent variable. A nonsignificant difference indicates that multicollinearity is not an issue in any given model.

Experiment 1: Online data

Linearity of variance

To check for linearity of variance, we plotted the residuals of the best fitting model against the true value, the proportion of words correctly identified. This indicates that unexplained variance is correlated with the proportion correct which indicates that this model violates the assumption. Violations of the assumptions tend to happen in collected data. It has been argued that mixed-effects models tend to be robust, though we should still be aware that model estimates may be less accurate (Schielzeth et al., 2020).

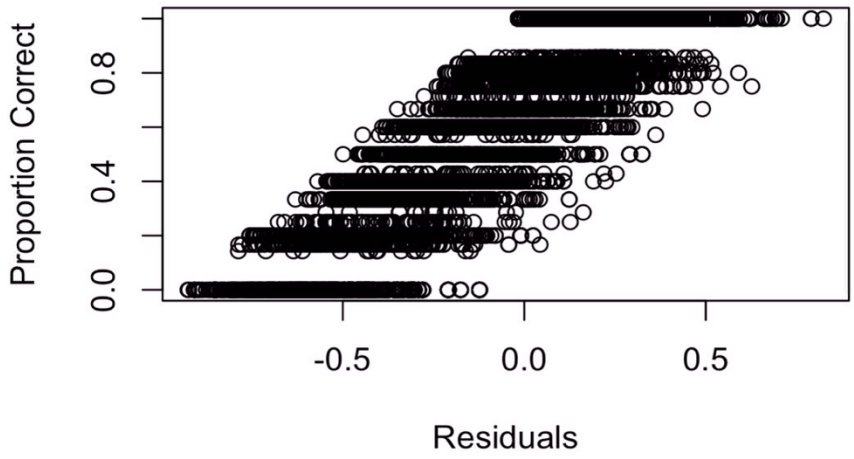

Figure 4.6 Residuals of the best fitting model plotted against the true values. Individual data points are represented by black circles.

Homogeneity of variance

One assumption of multilevel models is that of homoscedasticity. An ANOVA was run to test whether the variance of the residuals was equal. We found that the assumption was met as there was not a significant difference in variance at the individual level [$F(1,15118) = .83$, $p=.36$]. Figure 4.7 is a visual representation. Our data was centered around the zero line, however, indicating an even spread in the residuals.

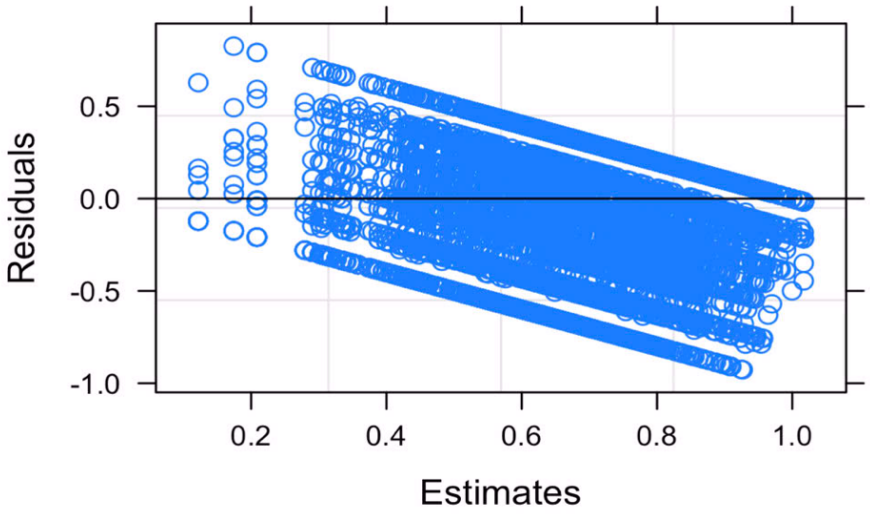

Figure 4.7 Plot of estimates by residuals. Individual data points are represented by black circles.

Normally distributed residuals

We explored normality based on a visual inspection of the relationship between standardized quantiles and residuals (Figure 4.8A). We noticed slight deviation near (-2, -2) and tried a logarithmic transformation on the data to explore whether or not this would improve our findings (Figure 4.8B). We found that the log transformation caused the residuals to appear less Normal.

Because the standard plot (Figure 4.8A) appears to be generally Normal, we concluded that the assumption of normality was met.

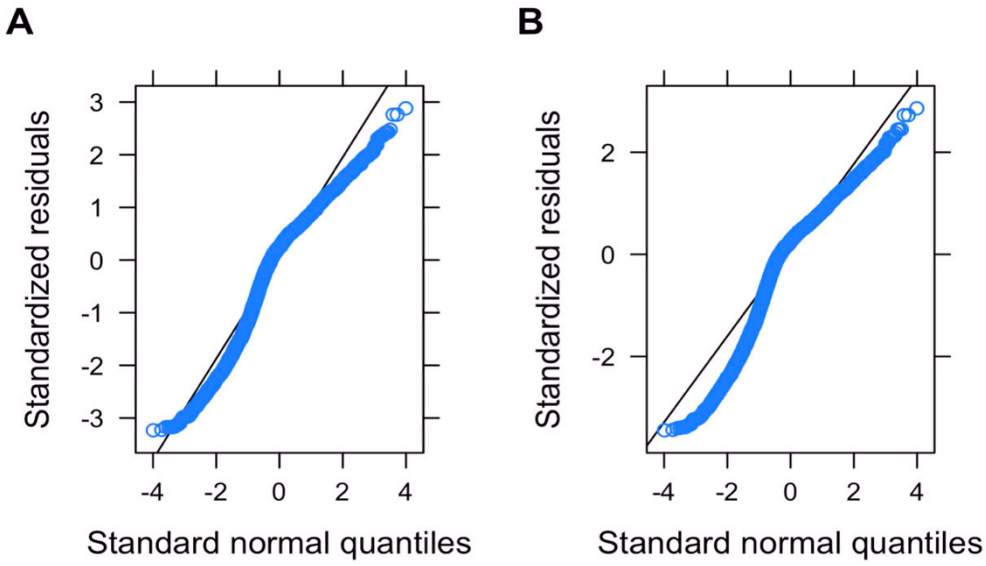

Figure 4.8 Plot of standardized quantiles by standardized residuals for data collected from Experiment 1. Individual observations are represented by blue circles.

Multicollinearity

Finally, we tested for existing collinearity amongst predictors on the dependent variable. After running a Chi-squared test, we found no evidence of multicollinearity [$\chi^2(3013) = 2980.10$, $p = .66$], indicating that the variables chosen were appropriate predictors for our multi-level modeling analysis.

Experiment 2: In-lab data

Linearity of variance

For Experiment 2, we also found that the assumption of linearity was violated. As explained above, mixed-effects models may be robust, but consideration should be taken into account when interpreting results of the model.

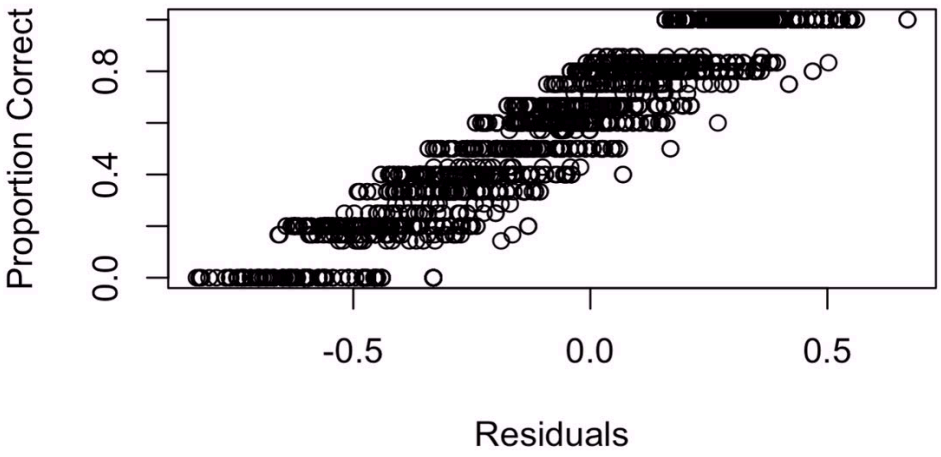

Figure 4.9 Residuals of the best fitting model plotted against the true values. Individual data points are represented by black circles.

Homogeneity of variance

We ran an ANOVA to test whether homoscedasticity was met. We found that the assumption was not met as there was not a significant difference in variance at the individual level [$F(1,3353) = 10.34$, p=.001]. This is likely, in part, due to our linearity violation. We did, however, find that the data were mostly centered around the zero line.

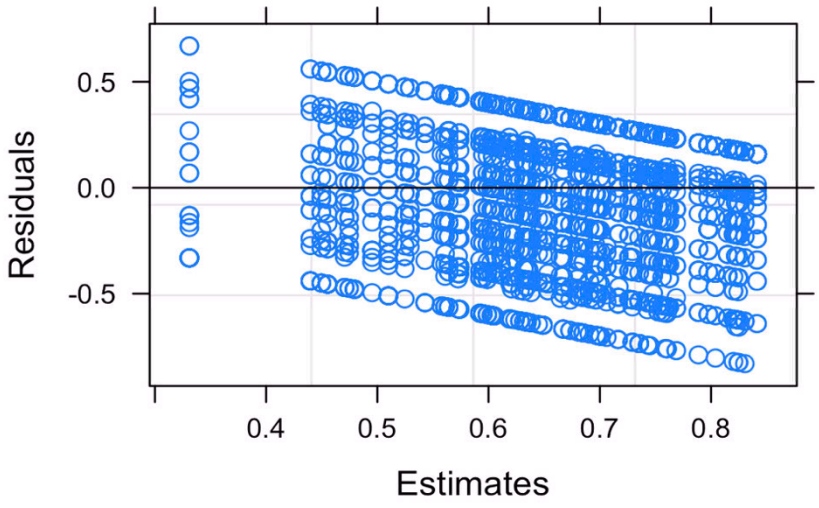

Figure 4.10 Plot of estimates and residuals. Individual observations are represented by blue circles.

Normally distributed residuals

We first examined the relationship between standardized quantiles and standardized residuals. Based on visual inspection, the residuals appeared to be relatively, normally distributed though there was slight variation (Figure 4.11A). A logarithmic base 10 transformation was performed on the data to see whether this would improve our findings (Figure 4.11B). We found that this did, though transformations make interpreting the outcomes more difficult. It has been argued, however, that violations of normality may have little impact (Schielzeth et al., 2020). Because the data seemed generally Normal, the logarithmic transformation may not have provided additional value given the challenges in interpretation.

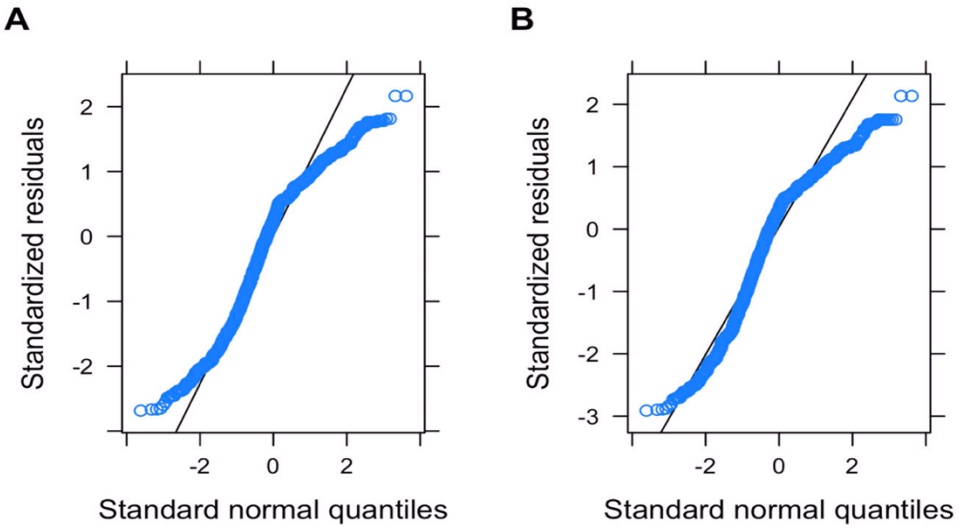

Figure 4.11 Plot of standardized quantiles and residuals before (A) and after (B) a logarithmic transformation. Individual observations are represented by blue circles.

Multicollinearity: Finally, we tested for any existing collinearity amongst predictors and the dependent variable. Using a Chi-squared test, we found no evidence of multicollinearity [$\mathcal{X}^2(52) = 56$, $p = .33$], indicating that the variables chosen were appropriate predictors for our multi-level modeling analysis.

Experiment 2 and 3: Combined dataset

Linearity of variance

As this data is a combination of data collected from the prior two experiments, it was expected that the linearity assumption was violated (4.12). As stated, we should remain cautious when interpreting the output.

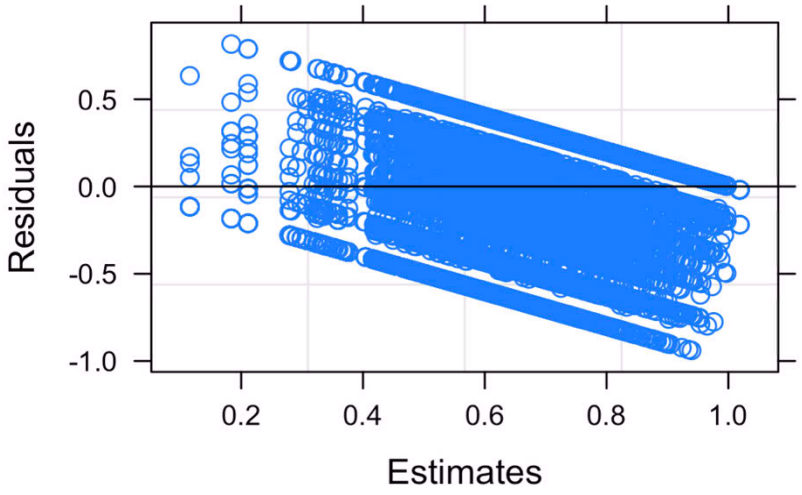

Figure 4.13 Plot of estimates by residuals. Individual data points are represented by blue

Homogeneity of variance

We ran an ANOVA to test whether homoscedasticity was met. We found that the assumption was not met as there was not a significant difference in variance at the individual level [$F(1,18473) = 4.49$ p=.03]. Visual inspection showed that, similar to the in-lab data assessment, the violation in linearity of residuals likely affected the results (Figure 4.13). Similarly to the plots above, the data were centered around zero indicating equal spread in the variance.

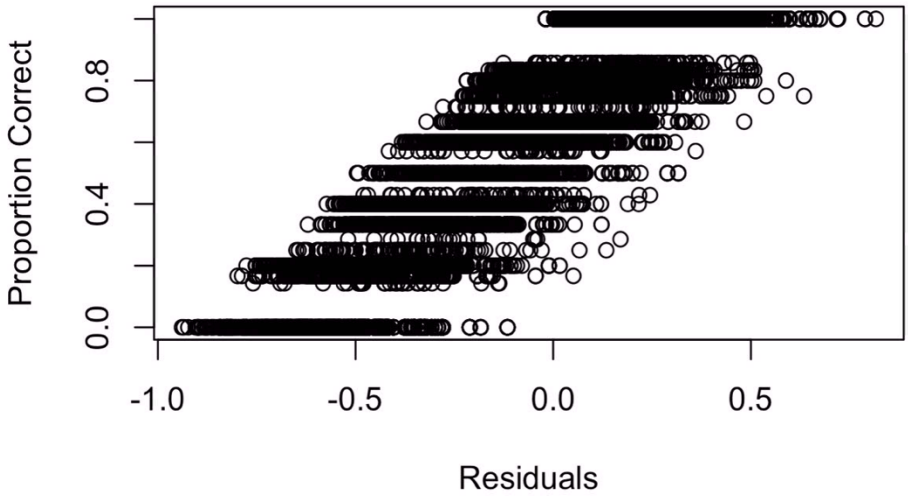

Figure 4.12 Residuals of the best fitting model plotted against the true values. Individual data points are represented by black circles.

Normally distributed residuals

Visual inspection of the relationship between standardized quantiles and residuals revealed a generally Normal distribution (Figure 4.14A). Though variation at the tails is common, we explored whether a logarithmic transformation would improve the data (Figure 4.14B). We found, however, that the relationship between standardized quantiles and residuals appeared less Normal than before. Therefore, we concluded that the data did not require a transformation and that the normality assumption was met.

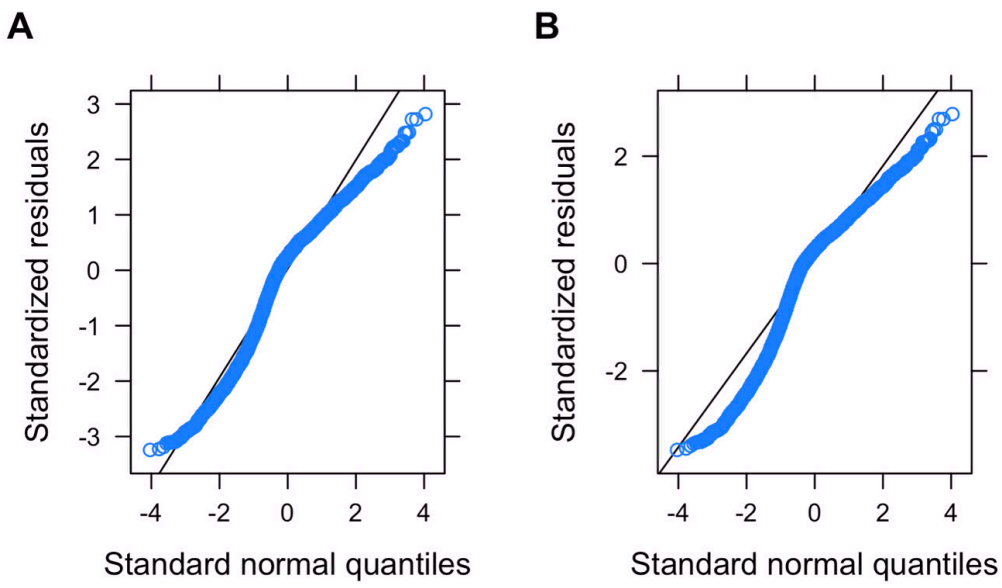

Figure 4.14 Plot of standardized quantiles by standardized residuals for data collected from Experiment 2. Individual observations are represented by blue circles.

Multicollinearity

Finally, we tested for any existing collinearity amongst predictors and the dependent variable. Using a Chi-squared test, we found no evidence of multicollinearity [$\mathcal{X}^2(3013) = 2980.10$, $p = .66$].

4.4 DISCUSSION

In contemporary studies exploring sleep and memory reactivation, ANOVA is most often used. Replication allows psychologists to have confidence in findings, although in some cases we may want to pool data from similar studies to explore broad effects. This approach was used recently in a meta-analysis of targeted memory reactivation studies (Hu et al., 2020).

In this dissertation, data were collected from participants in two separate experiments. Though the two experiments differed in format and study design, there were also similarities as both explored how participants performed on a second-language English transcription task

before and after a delay. In Chapters 2 and 3, data were analyzed for the online Prolific experiment and the in-lab nap experiment using mixed effects ANOVA and Pearson correlations. Here, we analyzed data from each experiment separately using MLM. We also pooled data and explored broader effects while also accounting for differences in study design. This provided information which would not be possible with a more standard ANOVA analysis.

We found results that supported our previous findings such as the PBR talker being the most intelligible of the four included speakers. In Chapter 2 (Experiment 1), we found differing results from training to test based on training group. In Chapter 3 (Experiment 2), we found that speaker alone was sufficient in accounting for the majority of the error in the model, allowing us to predict performance in this specific design from speaker alone. From our combined model, however, we saw that test and training sentences were more similarly matched than we may have thought as we would expect participants to score 1% worse than the training condition.

It is important to note, however, that the assumption of homogeneity and linearity of variance were violated in many of the models. We speculate that this is due to the nature of the intelligibility of each speaker. Each predictor added to the multi-level models was categorical. Furthermore, speaker type was shown to greatly affect performance regardless of training type (Figure 2.6). Because there was a clear order of intelligibility of speakers which affected performance, this likely affected the variance such that it was much more structured than anticipated. Violating these assumptions affects our ability to interpret the data accurately though it may be the case that the fixed effects were more resilient to these effects (Schielzeth et al., 2020). Nevertheless, we should be cautious when interpreting these findings.

Overall, the MLM model was conducive to exploring performance differences based on format. For example, we found participants who completed the in-lab study, exposure to only the

native Turkish speaker may have hindered performance for the FAR speaker (Table 4.7), though by what may be considered a small amount. Regardless, this finding may, indicate that speakers might be learning to generalize information that only applies to certain speakers. Further linguistic analyses on these two speakers compared to the others are needed to determine whether an explanation might help explain this effect. We also found that participants tended to perform a bit worse (1%) in the lab for the test compared to online session controlling for speaker and session. Our in-lab participants, however, also slept in between the training session and test session which may have affected their performance compared to the online participants.

Though these analyses led to interesting findings, there were also limitations in our approach. First, because of the differences in Experiment 1 and Experiment 2, our abilities to nest data and develop a hierarchy were hindered. Incomplete groups due to experimental design, along with a relatively small sample size at the individual and especially group level, and a low intraclass correlation greatly affected the robustness of our findings (J. J. Hox & Maas, 2001). An increase in sample size at each level and more similarity in study design should be explored. It is also apparent that the estimates and ICC's tend to be quite small. It may be possible that this task and/or study design is not sensitive enough for this type of model especially when it comes to slight differences in performance based on speaker or memory reactivation.

Despite these limitations, MLM provides an added benefit as we could combine data and support existing findings while also highlighting novel avenues for exploration. Though we could have attempted to combine data using ANOVA and t-tests, MLM depends on a hierarchical structure of predictors. This structure considers the differences in study design, explained above, and groups data to better estimate effects at distinct, nested levels. In the future, we should explore how increasing sample size or more similarity in experimental design may

affect findings. We should also explore other, novel, approaches to better understand our data. Overall, throughout this chapter we able to analyze our data in a novel way, highlighting points for future exploration.

Chapter 5

General Discussion

5.1 THESIS OVERVIEW AND FINDINGS

Introduction

In today's technical and interconnected world, it is not unlikely for a conversation to occur where one person is functioning in their first language (L1) while the other is functioning in their second language (L2). Understanding any talker may be difficult until the listener learns to adapt to the particular speaker, especially in difficult listening conditions. To better understand how this process occurs, researchers have recreated this situation in a laboratory setting by having participants transcribe L2 speech in noise. Previously, it has been shown that listeners can improve in their ability to recognize words after training on the same talker, or many talkers of the same language background (Bradlow & Bent, 2008). Further, exposure to a variety of different L2 talkers of different language backgrounds has been found support the ability to generalize to talkers from novel language backgrounds which is thought to indicate there may exist systematic properties of L2 English that listeners can learn overtime (Baese-Berk et al., 2013). It has been found that participants have difficulty categorizing the language background of L2 as they rely on bottom-up rather than top-down information for successful transcription (Xie & Myers, 2017). The authors suggest this finding demonstrates that explicit categorization or clustering of speakers based on L1is not a strategy listeners rely on. Based on these findings, it is likely that implicit memory systems are crucial for this type of generalization process. There has also been evidence that sleep may play an important role in the ability to generalize to new talkers (Xie et al., 2018).

To explore the role of sleep in perceptual generalization of speech, we used a technique known as targeted memory reactivation (TMR). TMR relies on the pairing of external stimuli with learning episodes that are played unobtrusively, in an attempt to reactivate and strengthen memory (Rasch et al., 2007; Rudoy et al., 2009). TMR has been shown successful for both implicit and explicit memory tasks (Hu et al., 2020).

Experiment 1

To inform the sleep study, we first wanted to explore how participants performed on trained and novel talkers before and after a delay based on the variability of incoming speech sounds. In Experiment 1, L1-English listeners completed an L2-English transcription task (Figure 2.1). The four talkers included in the experiment had one of four L1 backgrounds: Brazilian Portuguese (PBR), Spanish (SPA, Farsi (FAR), or Turkish (TUR). All sentences were spoken in English and the number of words correctly transcribed per session divided by the total number of words presented served as the participant score (Section 4.2.1 Scoring). During the initial session, they were exposed to either one talker (low variability) or three talkers (high variability), termed single and multiple talker groups, respectively. Roughly 11 hours later, participants transcribed 60 novel sentences split between each of the four mentioned speakers. The single talker group heard the trained talker and three novel talkers while the multiple talker group heard three trained talkers and one familiar talker.

We first found that at training, participants in the single talker group transcribed more words correctly than the multiple-talker group. This follows as the multiple talker group was exposed to many talkers while the single talker group had an opportunity to achieve speaker-dependent adaptation (Bradlow & Bent, 2008). We also found that both groups performed similarly at the test (Table 2.1; Figure 2.2; Figure 2.3). The most glaring difference regarding

speakers was between two particular talkers: PBR and TUR. We found that participants performed significantly better on the PBR talker and significantly worse on the TUR talker. When looking at all four multiple-talker training groups, we also saw that those who were not exposed to TUR (noTUR) tended to perform worse on the novel talker (TUR) than the other multiple-talker training groups did when faced with a novel talker (Figure 2.5). Though the TUR talker was less intelligible, we found that those who were only exposed to this speaker improved the most from training to test (Table 2.1). It is likely, however, that being exposed to more intelligible speakers at the test session contributed to this finding. Interestingly, the multiple-talker groups performed better on the TUR talker than the single-talker groups on average, indicating that exposure to a variety of talkers was aided in generalization for the most difficult talker (Table S1). For Experiment 2, we focused on this single-talker training group to explore whether sleep and memory reactivation could lead to improvements in word recognition.

Experiment 2

For Experiment 2, we ran a similar experiment in which participants transcribed 60 sentences spoken by the TUR talker at training and 60 sentences at test split evenly amongst the PBR, FAR, SPA, and TUR talker (Figure 3.1). Participants also completed a control task during which they pressed keys in response to visual and auditory cues. A repeating sequence was embedded within the task. Previous literature has shown that TMR during sleep has proven to improve participant performance on the repeating sequence measured by increased speed and accuracy (Antony et al., 2012). Between the training and test session, participants were given 90-minutes to nap during which they received auditory cues either of the repeating sequence from the control task or the spoken words transcribed during training. Cues were presented during slow-wave sleep (SWS) and paused during signs of arousal.

Here, we found that at training, participants improved over time, performing better than the TUR group from Experiment 2. After calculating the final training score, we did find a marginal effect that emerged between the cued and control group. This effect has no clear explanation as participants were randomly assigned to groups and tasks were counterbalanced across participants, so it is likely that this is a spurious finding. At test, both groups improved by a similar degree (Table 3.1; Figure 3.3). When looking at the performance at test for the novel and learned talkers, the control group tended to perform better for all speakers (Table 3.2).

When exploring the various sleep measures, we found that both groups of participants were matched (Table 3.3; Table 3.4). Time spent in N2 positively correlated with improvement for both groups. We, however, wanted to explore how sleep may have affected performance, particularly for the cued group. For these participants, total time spent asleep correlated with a greater degree of improvement. We also found that improvement was negatively correlated with a greater amount of sleep disruption. The majority of participants had very little sleep disruption our goal was to present the sounds so as not to wake participants and disrupt their sleep. Considering the sleep results, it may be the case that total time asleep and sleep disruption, combined with the fact that the cued and control groups were not equal at training may have contributed to our finding that TMR did not improve performance in this task.

Finally, we also tested multi-level models (MLM) to explore how this mathematical tool might improve our understanding of the data collected. Though there were a number of significant findings (see Chapter 4), a few stood out as being particularly impactful. For the data from Experiment 1, we found that participants performed well on the SPA talker and exceptionally well on the PBR talker compared to other talkers. For Experiment 2, we found that performance was significantly better at test than at training. Finally, by comparing all of the data,

we found that participants performed much better on the PBR talker compared to all other talkers.

Overall, we found that multiple-talker training tended to support generalization while single-talker training did not. There is complexity in this finding, however, as the PBR and TUR talkers greatly affected performance as they were the most and least intelligible talkers, respectively. However, it appears that performance on these two talkers did not drive this finding as performance is similar across talkers regardless of training group (Table S1). This indicates that there may indeed have been a base of generalized knowledge that allowed participants in the multiple-talker group to improve at training.

We also found that neither sleep nor memory reactivation improved generalization. There are a couple of potential explanations for this finding. First, despite randomization, participants in the control and cued groups were significantly different at training. Further, when looking at sleep measures, a decline in performance was significantly correlated with sleep disruption related to the cues. These two factors likely contributed to the null effect of TMR. It may also be the case that sleep and/or TMR may not be helpful for this type of task though we cannot be confident in this conclusion based on the other factors explained above.

Limitations

In Experiment 1, there were limitations that may have affected our results. First, the experiment was run online. Because of this, it's possible that participants may not have followed instructions as precisely as if they were in the lab. For example, they may not have worn headphones or may have completed the task in a noisy environment. We also did not know what participants may have done during the day that could have affected their performance at either the training or nap session. Further, because we did not have access to a participant's location, we cannot be sure of

the exact time zone during which the experiment was completed. To account for this issue, we opened the experiment for 6 hours to account for a number of time zones across the United States. Finally, we were able to select for L1 English speakers between the ages of 18-35 who experienced no hearing difficulties. Though running participants outside of the lab makes it more difficult to control for external variables, participants from the online platform, Prolific, have been shown to provide useful data for L2 language experiments (Strori et al., 2020).

There were also limitations that affected Experiment 2. First, because it is common that people learn more than one language, finding L1 English speakers was difficult. Due to this, we limited our participants to those who considered themselves to be English dominant. We believe, however, that participants who functioned mainly in English had a perceptual space that was attuned to English sounds to limit the effect this had on performance. Along these lines, though a participant may have considered themselves to be English dominant, it is possible that they often heard L2 English speakers on any given day which may have affected performance.

When looking at the description of sleep, none of our participants achieved REM. This is not surprising as reaching REM during a 90-minute nap in a new space may be difficult. It has been hypothesized, however, that REM may play an important role in the ability to use learned information in a more generalized context (Sterpenich et al., 2014; Tamminen et al., 2017). Because Experiment 2 was a nap design, we cannot know whether REM sleep may have supported generalization for this task. Finally, it is difficult to assess whether the auditory cues were processed during sleep. To address this possibility, speakers were set to a volume that each participant considered to be optimal for sleep based on pink noise. We presented sounds at this volume while increasing the volume by .1-.2dB for each round of auditory cues presented, stopping upon signs of arousal.

In Chapter 4, we tested multi-level models to determine which best fit data from Experiment 1, Experiment 2, and both Experiment 1 and 2 combined. Although we were able to gather useful information, the assumptions of linearity and homogeneity of variance were not met. Further, because these experiments were not originally designed to be analyzed with HLM, there are likely design choices that, if changed, would allow for a better assessment of the data.

5.2 FUTURE DIRECTIONS

Given the interesting interaction of performance based on training group and the speaker transcribed, future research should explore the similarities and differences in these speakers, and in other speakers to explore the necessary requirements for successful generalization. Further, a similar study should be conducted using an overnight study design to determine whether REM plays an important role in the generalization of L2 English speech. Doing so will provide insight into the way the brain might support the integration of information and how this information might be used based on differences and similarities in the speakers heard at training and test.

5.3 CONCLUSION

This dissertation was focused on better understanding human communication. Experiment 1 and 2 allowed us to explore the effect of exposure to variability, how intelligibility may affect our ability to understand novel talkers, and the way sleep may support this type of learning. This this research also focuses on L2-English transcription which contributes to a body of work that rebalances the communicative burden of talkers by exploring communication from the perspective of the listener (Bent & Baese-Berk, 2021). Researchers should continue to explore the perception of L2 speech so that we might (1) better understand how the brain supports this process and (2) learn how we can become better listeners.

Chapter 6

References

Antony, J. W., Gobel, E. W., O'hare, J. K., Reber, P. J., & Paller, K. A. (2012). Cued memory reactivation during sleep influences skill learning. *Nature Neuroscience, 15*(8), 1114.

Antony, J. W., Schönauer, M., Staresina, B. P., & Cairney, S. A. (2019). Sleep spindles and memory reprocessing. *Trends in Neurosciences, 42*(1), 1-3.

Baese-Berk, M. M., Bradlow, A. R., & Wright, B. A. (2013). Accent-independent adaptation to foreign accented speech. *The Journal of the Acoustical Society of America, 133*(3), EL174-EL180.

Batterink, L. J., & Paller, K. A. (2017). Sleep-based memory processing facilitates grammatical generalization: Evidence from targeted memory reactivation. *Brain and Language, 167*, 83-93. doi:10.1016/j.bandl.2015.09.003

Bent, T., & Baese-Berk, M. (2021). Perceptual learning of accented speech. *The Handbook of Speech Perception*, 428-464.

Berry, R. B., Brooks, R., Gamaldo, C., Harding, S. M., Lloyd, R. M., Quan, S. F., . . . Vaughn, B. V. (2017). Aasm scoring manual updates for 2017 (version 2.4). In (Vol. 13, pp. 665-666): American Academy of Sleep Medicine.

Boduch-Grabka, K., & Lev-Ari, S. (2021). Exposing individuals to foreign accent increases their trust in what nonnative speakers say. *Cognitive science, 45*(11), e13064.

Born, J., Rasch, B., & Gais, S. (2006). Sleep to remember. *The Neuroscientist, 12*(5), 410-424.

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), 392-399.

Bradlow, A. R. (n.d.). Allsstar: Archive of l1 and l2 scripted and spontaneous transcripts and recordings.

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition, 106*(2), 707-729.

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura, Y. i. (1997). Training japanese listeners to identify english/r/and/l: Iv. Some effects of perceptual learning on speech production. *The Journal of the Acoustical Society of America, 101*(4), 2299-2310.

Buzsaki, G. (1998). Memory consolidation during sleep: A neurophysiological perspective. *Journal of Sleep Research, 7*(S1), 17-23.

Cheng, L. Y., Che, T., Tomic, G., Slutzky, M. W., & Paller, K. A. (2021). Memory reactivation during sleep improves execution of a challenging motor skill. *Journal of Neuroscience, 41*(46), 9608-9616.

Clark, T. S., & Linzer, D. A. (2015). Should i use fixed or random effects? *Political Science Research and Methods, 3*(2), 399-408.

Clemens, Z., Fabo, D., & Halasz, P. (2005). Overnight verbal memory retention correlates with the number of sleep spindles. *Neuroscience, 132*(2), 529-535.

Clemens, Z., Fabó, D., & Halász, P. (2006). Twenty-four hours retention of visuospatial memory correlates with the number of parietal sleep spindles. *Neuroscience Letters, 403*(1-2), 52-56.

Clemens, Z., Mölle, M., Erőss, L., Barsi, P., Halász, P., & Born, J. (2007). Temporal coupling of parahippocampal ripples, sleep spindles and slow oscillations in humans. *Brain, 130*(11), 2868-2878.

Cooper, A., & Bradlow, A. R. (2016). Linguistically guided adaptation to foreign-accented speech. *The Journal of the Acoustical Society of America, 140*(5), EL378-EL384.

Creery, J. D., Oudiette, D., Antony, J. W., & Paller, K. A. (2015). Targeted memory reactivation during sleep depends on prior learning. *Sleep, 38*(5), 755-763.

Crupi, D., Hulse, B. K., Peterson, M. J., Huber, R., Ansari, H., Coen, M., . . . Tononi, G. (2009). Sleep-dependent improvement in visuomotor learning: A causal role for slow waves. *Sleep, 32*(10), 1273-1284.

Diekelmann, S., & Born, J. (2010). The memory function of sleep. *Nature Reviews Neuroscience, 11*(2), 114-126.

Eichenbaum, H. (2004). Hippocampus: Cognitive processes and neural representations that underlie declarative memory. *Neuron, 44*(1), 109-120.

Eisner, F., & McQueen, J. M. (2006). Perceptual learning in speech: Stability over time. *The Journal of the Acoustical Society of America, 119*(4), 1950-1953.

Eitan, S., Antony, J. W., Lampe, A., Wilson, B. J., Norman, K. A., & Paller, K. A. (2021). Multiple memories can be simultaneously reactivated during sleep as effectively as a single memory. *Communications Biology, 4*(1).

Eschenko, O., Ramadan, W., Mölle, M., Born, J., & Sara, S. J. (2008). Sustained increase in hippocampal sharp-wave ripple activity during slow-wave sleep after learning. *Learning & Memory, 15*(4), 222-228.

Farthouat, J., Gilson, M., & Peigneux, P. (2017). New evidence for the necessity of a silent plastic period during sleep for a memory benefit of targeted memory reactivation. *Sleep Spindles & Cortical Up States, 1*(1), 14-26.

Fernández-Castilla, B., Jamshidi, L., Declercq, L., Beretvas, S. N., Onghena, P., & Van den Noortgate, W. (2020). The application of meta-analytic (multi-level) models with multiple random effects: A systematic review. *Behavior Research Methods, 52*(5), 2031-2052. doi:10.3758/s13428-020-01373-9

Fox, J. (2008). Applied regression analysis and general linear models2nd editionthousand oaks. In: CASage Publications.

Hahn, M. A., Heib, D., Schabus, M., Hoedlmoser, K., & Helfrich, R. F. (2020). Slow oscillation-spindle coupling predicts enhanced memory formation from childhood to adolescence. *eLife, 9*, e53730.

Hayes, A. F. (2006). A primer on multilevel modeling. *Human communication research, 32*(4), 385-410.

Hox, J. (1998). *Multilevel modeling: When and why.* Paper presented at the Classification, data analysis, and data highways: proceedings of the 21st Annual Conference of the Gesellschaft für Klassifikation eV, University of Potsdam, March 12–14, 1997.

Hox, J. J., & Maas, C. J. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Structural equation modeling, 8*(2), 157-174.

Hu, X., Cheng, L. Y., Chiu, M. H., & Paller, K. A. (2020). Promoting memory consolidation during sleep: A meta-analysis of targeted memory reactivation. *Psychological Bulletin, 146*(3), 218.

Huber, R., Felice Ghilardi, M., Massimini, M., & Tononi, G. (2004). Local sleep and learning. *Nature, 430*(6995), 78-81.

Iverson, P., Hazan, V., & Bannister, K. (2005). Phonetic training with acoustic cue manipulations: A comparison of methods for teaching english/r/-/l/to japanese adults. *The Journal of the Acoustical Society of America, 118*(5), 3267-3278.

Johnson, B. P., Scharf, S. M., & Westlake, K. P. (2018). Targeted memory reactivation during sleep, but not wake, enhances sensorimotor skill performance: A pilot study. *Journal of Motor Behavior, 50*(2), 202-209.

Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin & Review, 13*(2), 262-268.

Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language, 56*(1), 1-15.

Lerner, I., & Gluck, M. A. (2019). Sleep and the extraction of hidden regularities: A systematic review and the importance of temporal rules. *Sleep Medicine Reviews, 47*, 39-50.

MATLAB, T. U. s. G. (2022). Natick, massachusetts: The mathworks inc. In.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review, 102*(3), 419.

McNeish, D. (2017). Small sample methods for multilevel modeling: A colloquial elucidation of reml and the kenward-roger correction. *Multivariate behavioral research, 52*(5), 661-670.

Meier-Koll, A., Bussmann, B., Schmidt, C., & Neuschwander, D. (1999). Walking through a maze alters the architecture of sleep. *Perceptual and Motor Skills, 88*(3_suppl), 1141-1159.

Mikutta, C., Feige, B., Maier, J. G., Hertenstein, E., Holz, J., Riemann, D., & Nissen, C. (2019). Phase-amplitude coupling of sleep slow oscillatory and spindle activity correlates with overnight memory consolidation. *Journal of Sleep Research, 28*(6), e12835.

Molle, M., Yeshenko, O., Marshall, L., Sara, S. J., & Born, J. (2006). Hippocampal sharp wave-ripples linked to slow oscillations in rat slow-wave sleep. *Journal of Neurophysiology, 96*(1), 62-70.

Muehlroth, B. E., Sander, M. C., Fandakova, Y., Grandy, T. H., Rasch, B., Shing, Y. L., & Werkle-Bergner, M. (2019). Precise slow oscillation–spindle coupling promotes memory consolidation in younger and older adults. *Scientific Reports, 9*(1), 1940.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology, 47*(2), 204-238.

O'keefe, J., & Nadel, L. (1978). *The hippocampus as a cognitive map*: Oxford university press.

Paller, K. A., Creery, J. D., & Schechtman, E. (2021). Memory and sleep: How sleep cognition can change the waking mind for the better. *Annual Review of Psychology, 72*, 123-150.

Paller, K. A., Mayes, A., Antony, J., & Norman, K. A. (2020). Replay-based consolidation governs enduring memory storage. *The Cognitive Neurosciences*, 265-276.

Patkowski, M. S. (1990). Age and accent in a second language: A reply to james emil flege. *Applied Linguistics, 11*(1), 73-89.

Pavlides, C., & Winson, J. (1989). Influences of hippocampal place cell firing in the awake state on the activity of these cells during subsequent sleep episodes. *Journal of Neuroscience, 9*(8), 2907-2918.

Peigneux, P., Laureys, S., Fuchs, S., Collette, F., Perrin, F., Reggers, J., . . . Aerts, J. (2004). Are spatial memories strengthened in the human hippocampus during slow wave sleep? *Neuron, 44*(3), 535-545.

Rasch, B., & Born, J. (2013). About sleep's role in memory. *Physiological Reviews, 93*(2), 681-766. doi:10.1152/physrev.00032.2012

Rasch, B., Büchel, C., Gais, S., & Born, J. (2007). Odor cues during slow-wave sleep prompt declarative memory consolidation. *Science, 315*(5817), 1426-1429.

Rudoy, J. D., Voss, J. L., Westerberg, C. E., & Paller, K. A. (2009). Strengthening individual memories by reactivating them during sleep. *Science, 326*(5956), 1079-1079.

Sanchez, D. J., Gobel, E. W., & Reber, P. J. (2010). Performing the unexplainable: Implicit task performance reveals individually reliable sequence learning without explicit knowledge. *Psychonomic Bulletin & Review, 17*(6), 790-796.

Schabus, M., Gruber, G., Parapatics, S., Sauter, C., Klösch, G., Anderer, P., . . . Zeitlhofer, J. (2004). Sleep spindles and their significance for declarative memory consolidation. *Sleep, 27*(8), 1479-1485.

Schielzeth, H., Dingemanse, N. J., Nakagawa, S., Westneat, D. F., Allegue, H., Teplitsky, C., . . . Araya-Ajoy, Y. G. (2020). Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods in ecology and evolution, 11*(9), 1141-1152.

Schönauer, M., Geisler, T., & Gais, S. (2014). Strengthening procedural memories by reactivation in sleep. *Journal of Cognitive Neuroscience, 26*(1), 143-153.

Schreiner, T., & Rasch, B. (2015). Boosting vocabulary learning by verbal cueing during sleep. *Cerebral Cortex, 25*(11), 4169-4179.

Schreiner, T., & Rasch, B. (2017). The beneficial role of memory reactivation for language learning during sleep: A review. *Brain and Language, 167*, 94-105.

Squire, L. R. (2004). Memory systems of the brain: A brief history and current perspective. *Neurobiology of Learning and Memory, 82*(3), 171-177.

Squire, L. R., Genzel, L., Wixted, J. T., & Morris, R. G. (2015). Memory consolidation. *Cold Spring Harbor Perspectives in Biology, 7*(8), a021766.

Squire, L. R., & Wixted, J. T. (2011). The cognitive neuroscience of human memory since hm. *Annual Review of Neuroscience, 34*, 259-288.

Squire, L. R., & Zola, S. M. (1996). Structure and function of declarative and nondeclarative memory systems. *Proceedings of the National Academy of Sciences, 93*(24), 13515-13522.

Staresina, B. P., Bergmann, T. O., Bonnefond, M., Van Der Meij, R., Jensen, O., Deuker, L., . . . Fell, J. (2015). Hierarchical nesting of slow oscillations, spindles and ripples in the human hippocampus during sleep. *Nature Neuroscience, 18*(11), 1679-1686.

Steenbergen, M. R., & Jones, B. S. (2002). Modeling multilevel data structures. *American Journal of Political Science, 46*(1), 218-237. doi:10.2307/3088424

Stephen, R., & Anthony, B. (2002). *Hierarchical linear models*: Sage Publications, Thousand Oaks, CA.

Sterpenich, V., Schmidt, C., Albouy, G., Matarazzo, L., Vanhaudenhuyse, A., Boveroux, P., . . . Collette, F. (2014). Memory reactivation during rapid eye movement sleep promotes its generalization and integration in cortical stores. *Sleep, 37*(6), 1061-1075.

Stickgold, R. (2005). Sleep-dependent memory consolidation. *Nature, 437*(7063), 1272.

Strori, D., Bradlow, A. R., & Souza, P. E. (2020). Recognition of foreign-accented speech in noise: The interplay between talker intelligibility and linguistic structure. *The Journal of the Acoustical Society of America, 147*(6), 3765-3782.

Tamminen, J., Ralph, M. A. L., & Lewis, P. A. (2017). Targeted memory reactivation of newly learned words during sleep triggers rem-mediated integration of new memories and existing knowledge. *Neurobiology of Learning and Memory, 137*, 77-82.

Whitmore, N. W., Bassard, A. M., & Paller, K. A. (2021). Targeted memory reactivation of face-name learning depends on ample and undisturbed slow-wave sleep. *bioRxiv*.

Whitmore, N. W., & Paller, K. A. (2023). Sleep disruption by memory cues selectively weakens reactivated memories. *Learning & Memory, 30*(3), 63-69.

Wilson, M. A., & McNaughton, B. L. (1994). Reactivation of hippocampal ensemble memories during sleep. *Science, 265*(5172), 676-679.

Witkowski, S., Noh, S., Lee, V., Grimaldi, D., Preston, A. R., & Paller, K. A. (2021). Does memory reactivation during sleep support generalization at the cost of memory specifics? *Neurobiology of Learning and Memory, 182*, 107442.

Witkowski, S., Schechtman, E., & Paller, K. A. (2020). Examining sleep's role in memory generalization and specificity through the lens of targeted memory reactivation. *Current Opinion in Behavioral Sciences, 33*, 86-91.

Wright, B. A., Baese-Berk, M. M., Marrone, N., & Bradlow, A. R. (2015). Enhancing speech learning by combining task practice with periods of stimulus exposure without practice. *The Journal of the Acoustical Society of America, 138*(2), 928-937.

Xie, X., Earle, F. S., & Myers, E. B. (2018). Sleep facilitates generalisation of accent adaptation to a new talker. *Language, Cognition and Neuroscience, 33*(2), 196-210.

Xie, X., & Myers, E. B. (2017). Learning a talker or learning an accent: Acoustic similarity constrains generalization of foreign accent adaptation to new talkers. *Journal of Memory and Language, 97*, 30-46.

Yotsumoto, Y., Sasaki, Y., Chan, P., Vasios, C. E., Bonmassar, G., Ito, N., . . . Watanabe, T. (2009). Location-specific cortical activation changes during sleep after training for perceptual learning. *Current Biology, 19*(15), 1278-1282.

Chapter 5

Supplemental Materials

Table S1. Experiment 1 test performance for each training group by speaker					
Single Talker			**Multiple Talker**		
Group	**Talker**	**Average % ± SD %**	**Group**	**Talker**	**Average % ± SD %**
FAR	FAR	78 ± 29	noFAR	FAR	75 ± 31
FAR	PBR	89 ± 23	noFAR	PBR	83 ± 29
FAR	SPA	77 ± 30	noFAR	SPA	71 ± 33
FAR	TUR	60 ± 35	noFAR	TUR	61 ± 36
PBR	FAR	72 ± 31	noPBR	FAR	74 ± 30
PBR	PBR	84 ± 35	noPBR	PBR	85 ± 26
PBR	SPA	73 ± 31	noPBR	SPA	73 ± 32
PBR	TUR	57 ± 35	noPBR	TUR	63 ± 34
SPA	FAR	72 ± 33	noSPA	FAR	74 ± 30
SPA	PBR	85 ± 25	noSPA	PBR	82 ± 30
SPA	SPA	72 ± 32	noSPA	SPA	73 ± 31
SPA	TUR	58 ± 37	noSPA	TUR	57 ± 37
TUR	FAR	67 ± 34	noTUR	FAR	79 ± 27
TUR	PBR	79 ± 30	noTUR	PBR	88 ± 23
TUR	SPA	71 ± 32	noTUR	SPA	78 ± 30
TUR	TUR	56 ± 36	noTUR	TUR	62 ± 31